Modeling Differentiated Pricing Scheme for Heterogeneous Cloud Computing Environments

by

Mariam Mayengo

Reg. No: 2013/HD05/496U

Bachelor of Information Technology (MUST)

Department of Networks

School of Computing and Information Sciences, Makerere University

E-mail: nanmarium@gmail.com, Tel: +256775 232335

A Research Proposal Submitted to **School of Graduate Studies** for the Study Leading to a **Research Report** in Partial Fulfillment of the Requirements for the Award of Master of Science in Data Communication and Software Engineering of Makerere University.

OPTION: Communication Networks

Supervisor

Dr. Drake Mirembe (PhD)

Department of Networks

School of Computing and Information Science, Makerere University

Email:dpmirembe@gmail.com, Tel: +256-41-540628 Fax: +256-41-540620

January 2018

LIST OF ACRONYMS

BS	Base Station
FCFS	First Come First Served
IaaS	Infrastructure as a service
IEEE	Institute of Electrical and Electronic Engineers
MAC	Medium Access Control
Paas	Platform as a service
QoS	Quality of Service
SaaS	Software as a Service
TDMA	Time Division Multiple Access

Chapter 1

Introduction

Cloud computing is emerging as a vital practice for the online provisioning of computing resources as services and enables scalable on-demand sharing of resources and costs among a large number of end users. A majority of technology experts expect that by 2020 most people will access software applications online, and share and access information through the use of remote server networks using the cloud, rather than depending primarily on tools and information housed on their individual personal computers [1]. Cloud computing technology allows scalable on-demand sharing of resources and costs among a large number of end users.

Foster et al. [2] defined cloud computing as ”a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet.” Cloud computing comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams [3, 4]. Cloud computing entrusts remote services with a user’s data, software and computation.

Among the many types of cloud computing services delivered internally or by third party service providers, the most common are:

(a) Software as a Service (SaaS) - software runs on computers owned and managed by the SaaS provider. The software is accessed over the public Internet and generally offered on a monthly or yearly subscription.

(b) Infrastructure as a Service (IaaS) - computing, storage, networking, and other elements (security, tools) are provided by the IaaS provider via public Internet, Virtual Private Network, or

dedicated network connection. Users own and manage operating systems, applications, and information running on the infrastructure and pay by usage.

(c) Platform as a Service (PaaS) - All software and hardware required to build and operate cloud-based applications are provided by the PaaS provider via public Internet, VPN, or dedicated network connection. Users pay by use of the platform and control how applications are utilized throughout their lifecycle [5].

Pricing is a critical factor for organizations offering services or products [48]. Pricing is the process of determining what a service provider will receive from an end user in exchange for their services. According to Weinhardt *et al.,* [7] cloud computing success can be obtained only by developing adequate pricing techniques. Therefore, developing an appropriate pricing model will help achieve higher revenues.

Basically, a Service Level Agreement (SLA) represents an agreement between a customer and a provider to receive a particular service provision [8]. SLAs contain Quality of Service (QoS) parameters that must be maintained by a provider (e.g. response time, bandwidth, storage, reliability, deadline, throughput, delay, and cost) [9]. A customer submits a service request to a service provider, receives the desired result from the service provider with certain service level agreement, and pays for the service based on the amount of the service and the quality of the service.

Previous studies on pricing in cloud computing environments assume cloud servers are homogeneous i.e. they have the same characteristics [10, 12, 13]. Homogeneous cloud computing servers consists of same storage capacity, processing power, energy supply and same service rate. However, the deployment scenarios of cloud server systems is not represented by homogeneous servers. Cloud server systems consist of heterogeneous servers with different service rates [14]. Heterogeneous Computing refers to those systems that make use of different types of computational units. The computational unit can be a general-purpose processor, a special-purpose processor, or a co-processor [15].

The performance and utilization of cloud computing systems are heavily constrained by the characteristics of jobs being served (e.g., their sensitivity to latency) [16]. For example, in the current pricing design of Google, users may prefer to immediately process their jobs upon arrival and are not willing to tolerate any latency. This may result in reduced revenue for the cloud service provider since fewer requests are produced as a result of poor ordering of requests.

The need for increased heterogeneity in the computing systems is partially as the need for high-performance, highly reactive systems that interact with other environments too [47]. In practice, both latency-critical and delay-tolerant jobs coexist in the cloud [16]. Theoretical and experimental results show that latency-sensitive requests induce a low resource utilization of cloud resources of between 6% and 12% [18]. Conversely, delay-tolerant requests tend to lead to a much higher utilization [18, 19, 20].

In cloud computing, differentiated service is used to give different services to different classes of customers. An appropriate design of pricing could serve as a tool to improve performance and utilization of cloud computing systems. The need for the different pricing mechanisms to efficiently satisfy expectation of each class of customers in heterogeneous cloud computing environment is the recipe behind this study.

1.1 Statement of the problem

Pricing is a critical factor for organizations offering cloud computing services. How prices are set determines customer's behaviour and loyalty. An appropriate design of pricing could serve as a tool to incentize users to express their application's characteristics thus promoting the utilization of cloud resources which in turn reduces the unit cost of computing resources, enabling cloud providers to provide cheaper computing services without any loss of revenue. In a recent study, Nansamba et al. [21] proposed a pricing scheme for heterogeneous multiserver cloud computing system. The pricing scheme considers heterogeneous servers with different service rates and capacities. Although the model considers practical deployment scenarios of cloud servers, the users are assumed to have the same type of application and therefore served using First In First Out policy. The assumption of same type of applications heavily constraints the performance and utilization of computing systems since requests are processed upon arrival even if they are delay-tolerant. In practice, both latency-critical and delay-tolerant jobs coexist in the cloud [16]. Theoretical and experimental results show that latency-sensitive requests induce a low resource utilization of cloud resources of between 6% and 12% [18]. Conversely, delay-tolerant requests tend to lead to a much higher utilization [18, 19, 20, 22, 23]. Moreover, the current pricing scheme used in [21] restricts the choice of cloud users who may want to pay less but are comfortable with

incurring extra delay. To overcome the above challenges, we propose to introduce queue management techniques that differentiates task execution via queue reordering to service latency-sensitive requests before serving delay-tolerant requests. In addition, we propose to charge the cloud users who are not willing to tolerate any delay in the completion of their jobs using a standard pricing model in the cloud market (e.g., on-demand instances in Amazon EC2, the pricing in Google Cloud Platform). On the other hand, cloud users will be charged lower prices at the expense of delaying job completion time, the more delay a user can tolerate to complete its jobs, the lower the price is.

1.2 Objectives

1.2.1 General Objective

The main objective of the study is to develop a pricing scheme which takes into consideration the heterogeneity of the servers, differences in user applications and charges different prices according to achieved service performance in terms of mean response time.

1.2.2 Specific Objectives

Specific objectives of the project will include:

1. To review literature on pricing schemes in cloud computing environments.
2. To model a pricing scheme which takes into consideration the heterogeneity of the servers, differences in user applications and charges according to achieved service performance in terms of mean response time.
3. To evaluate the performance of the proposed model against the heterogeneous pricing scheme proposed in [21].

1.3 Scope

The scope of the research shall include modeling a pricing scheme for heterogeneous servers with different service rates and capacities. The scheduling policy used is First In First Out policy. For

practical purposes two classes of requests, latency-critical and delay-tolerant requests are considered. The cloud users are charged in such a way that those who are not willing to tolerate any delay in the completion of their requests are charged using a standard pricing model in the cloud market while the cloud users who are willing to delay their requests completion time are charged less. The pricing mechanism will be based on achieved service performance in terms of mean response time. Developing a more realistic pricing mechanism will help service providers estimate more accurately the revenue generated.

1.4 Significance of the research

Cloud computing is emerging as a promising field offering online computing resources as services. Since pricing is a critical factor for organizations offering services or products [48], therefore how the price is set affects customer behavior, loyalty to a provider, and the organization's success. The assumption of same type of users heavily constraints the performance and utilization of computing systems since requests are processed upon arrival even if they are delay-tolerant. In addition, providing customers with payment options can bring customer satisfaction, since each customer goes for what they have chosen. A customer satisfied with a provider's services will continue to use them in the future and recommend them to peers, and this eventually results in higher revenues. Therefore, this study will come up with a more realistic model that will be useful in providing useful insights about pricing in cloud computing. Furthermore, this study will provide a base line for future studies concerning pricing in cloud computing.

Chapter 2

Literature Review

In this chapter, we present an overview of the general literature on cloud computing with particular focus on pricing models. It specifically presents the pricing models used in cloud computing.

2.1 Overview of cloud computing networks

Cloud computing is an on-demand service that has obtained mass appeal in corporate data centers. Cloud computing is quickly becoming an effective and efficient way of computing resources and computing services consolidation [25]. Cloud computing is Internet-based computing, whereby shared resources, software, and information are provided to computers and other devices on demand. The cloud enables the data center to operate like the Internet and computing resources to be accessed and shared as virtual resources in a secure and scalable manner.
Cloud computing represents the infrastructure as a cloud from which businesses and users are able to access the applications from anywhere in the world on demand. The University of California at Berkeley defines cloud computing as follows: ”Both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services” [26].

Cloud computing promises reliable service delivery through the next generation data centers that are built on the basis of virtualization, storage and virtual machine technologies [27]. This implies that users are able to access their applications by a cloud anywhere, anytime across the world easily.

The ability of organizations to tap into computer applications and other software via the cloud and thus free themselves from building and managing their own technology infrastructure seems potentially irresistible. Some companies providing cloud services have been growing at double-digit rates despite the recent economic downturn [27].

2.1.1 Service models of cloud computing

Cloud computing providers offer their services according to three fundamental models [28]: Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). Figure 2.2 shows cloud computing logical diagram.

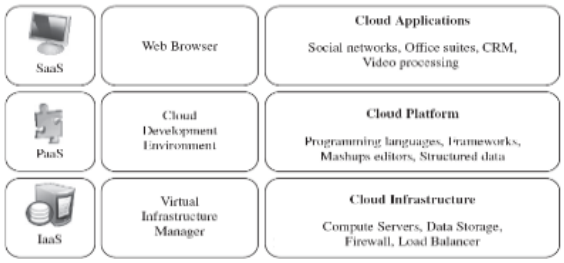

Figure 2.1: Cloud computing service model. Adopted [29]

1. **Platform as a Service:**

   Is a paradigm for delivering operating systems and associated services over the Internet without downloads or installation [30]. Platform as a Service (PaaS) provides computational resources through the platform such as Operating System. PaaS is a proven model for running applications without the hassle of maintaining the hardware and software infrastructure at your company. A PaaS provider hosts the hardware and software on its own infrastructure. PaaS is built upon the principals of Infrastructure as a Service by providing an environment where applications can be built and deployed in a secure, rapid and high quality manner. PaaS eliminates the hardware dependency and capacity concerns. However, the consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

2. **Infrastructure as a Service:**

   The core computing resources are hardware and software components. They lay the founda-

tions of every computing infrastructure [30]. This model involves outsourcing the equipment used to support operations, including storage, hardware, servers and networking components. The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control over some selected networking components (e.g., host firewalls). Amazon Elastic Computing Cloud (EC2) is the leading Infrastructure as a Service (IaaS) provider [31]. Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud.

3. **Software as a Service:**

   SaaS is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, typically the Internet. Software as a service (SaaS) allows users to connect to and use cloud-based applications over the Internet. Common examples are email, calendaring, and office tools (such as Microsoft Office 365). SaaS is the most basic form of cloud computing [5]. It includes implementation of specific business functions, customized business applications, etc. The major benefit of SaaS is that there is no licensing risk involved and no version compatibility issue. SaaS reduces the hardware costs as well. Customers pay for the software and the underlying infrastructure and does not require technical know-how. In this model, the customer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, except limited user-specific application configuration settings.

2.1.2 Deployment Models

Cloud computing consists of four deployment models [28]:

1. *Public cloud:* In this case, the cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. In this case the cloud exists on the premises

of the cloud provider. In public clouds, resources are offered as a service, usually over an Internet connection, for a pay-per-usage fee. Users can scale their use on demand and do not need to purchase hardware to use the service. Public clouds are available to the general public or large organizations, and are owned by a third party organization that offers the cloud service [32]. Examples of public cloud include: Amazon AWS, Google Apps, Salesforce.com, Microsoft BPOS, and Microsoft Office 365.

2. *Hybrid cloud:* Hybrid clouds are more complex than the other deployment models, since they involve a composition of two or more clouds (private, community, or public). Each member remains a unique entity, but is bound to others through standardized or proprietary technology that enables application and data portability among them [33]. A hybrid cloud is a composition of at least one private cloud and at least one public cloud. A hybrid cloud is typically offered in one of two ways: a vendor has a private cloud and forms a partnership with a public cloud provider, or a public cloud provider forms a partnership with a vendor that provides private cloud platforms [34].

3. *Private cloud:* In this case, the cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party , or some combination of them, and it may exist on or off the premises. Traits that characterize private clouds include the ring fencing of a cloud for the sole use of one organization and higher levels of network security. They can be defined in contrast to a public cloud which has multiple clients accessing virtualised services which all draw their resource from the same pool of servers across public networks. Private cloud services draw their resource from distinct pool of physical computers that may be hosted internally or externally and may be accessed across private leased lines or secure encrypted connections via public networks.

4. *Community cloud:* Here the cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). Community cloud is similar to a private cloud, but the infrastructure and computational resources are exclusive to two or more organizations that have common privacy, security, and regulatory considerations, rather than a single organization [32]. Community cloud may be owned, managed, and operated

by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off the premises.

2.2 Pricing Models in Cloud Computing

Pricing is an important factor for the company which provides cloud services because it affects the clients directly and the organization profit. The price also has a major impact in economic aspect, where key concepts such as fairness and competitive pricing in a multi-provider marketplace affect the actual pricing [35]. Each provider has a scheme for calculating the price for the cloud services offered for clients. The provider's goal is to have a greater benefit, while each client's goal is to have the maximum service for low price. Therefore, satisfying both parties requires an optimal pricing scheme. The price charged is one of the most important metrics that a service provider can control to encourage the usage of its services [36]. According to Weinhardt et al., [7] cloud computing success can be obtained only by developing adequate pricing techniques. How the price is set affects customer behavior, loyalty to a provider, and the organization's success. The price determined for a service or product must consider the manufacturing and maintenance costs, market competition, and how the customer values the service or product offered [37]. Therefore, developing an appropriate pricing model will help achieve higher revenues.
The following are the most pertinent factors that influence pricing in cloud computing [38, 36, 39]:

1. QoS. This is the set of technologies and techniques offered by the service provider to enhance the user experience in the cloud, such as data privacy and resource availability. The better QoS offered, the higher the price will be.

2. Age of resources. This is the age of the resources employed by the service provider. The older the resources are, the lower the price charged will be. This is because resources can sustain wear over time, which reduces their financial value.

3. Initial costs. This is the amount of money that the service provider spends annually to buy resources.

4. Lease period. This is the period in which the customer will lease resources from the service provider. Service providers usually offer lower unit prices for longer subscription periods.

5. Cost of maintenance. This is the amount of money that the service provider spends on maintaining and securing the cloud annually.

2.2.1 Cloud computing pricing approaches

Cloud computing pricing describes the process by which the price is determined.

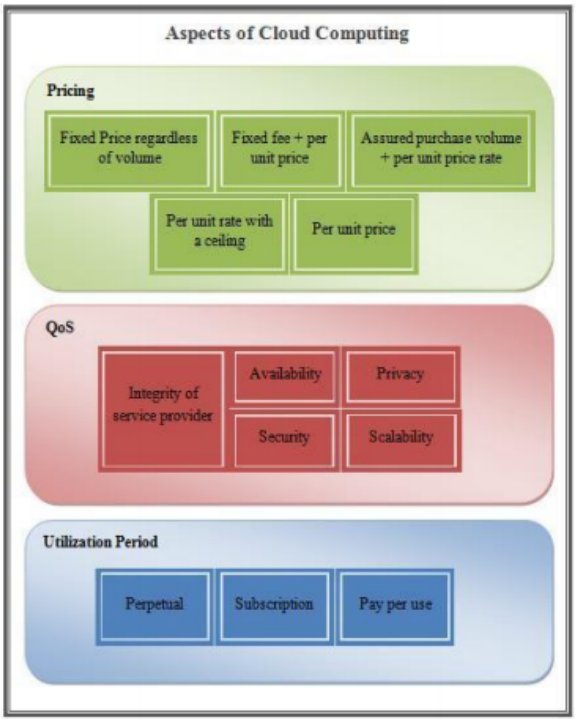

Figure 2.2: Aspects of Cloud Computing. Adopted [37]

The most common pricing approaches include the following [40]:

1. Fixed price regardless of volume: The fixed price regardless of volume charges the customer a fixed price regardless of the volume of the service or product utilized.

2. Fixed price plus per-unit rate: The fixed price plus per-unit charges the customer a fixed price plus a unit rate.

3. Assured purchase volume plus per-unit price rate: In the assured purchase volume plus per-unit price rate, the customer pays a fixed price for a certain quantity. If the customer's utilization exceeds that quantity, the customer has to pay a fixed rate per unit for the extra utilization.

4. Per-unit rate with a ceiling: In the per-unit rate with a ceiling approach, the customer pays the per-unit rate up to a certain limit. The provider will not charge the customer above that limit.

5. Per-unit price: In the price per unit approach, the customer is charged a different price per unit.

6. The quality of service (QoS) describes the requirements for what a service provider should provide to his customers. These requirements include the availability of service, security, privacy, scalability, and integrity of the service provider. If the service provider ensures that these requirements are maintained at a high level, the quality of the service offered will increase.

7. The utilization period can be defined as the period in which the customer has the right to utilize the provider services based on Service Level Agreements between the two parties.

Different service providers employ different schemes and models for pricing. However, the most common model employed in cloud computing include:

1. **Fixed Pricing**
   Fixed pricing includes pricing mechanism as pay-per-use pricing, subscription and list price/menu price[39]. Amazon, considered the market leader in cloud computing, utilizes such a model by charging a fixed price for each hour of virtual machine usage [41]. The ”pay-as-you-go” model is also implemented by other leading enterprises such as Google App Engine[42] and Windows Azure[43].

2. **Pay for resources** Another common scheme employed by leading enterprises is the ”pay for resources” model. A customer pays for the amount of bandwidth or storage utilized and pays in advance for the services he is going to receive for a pre-defined period of time.

3. **Pay-per-use** In this model, users only have to pay for what they use. Customer pays in function of the time or quantity he consumes on a specific service. Pay-per-use makes users aware of the cost of doing business and consuming a resource. Jaatmaa*et al.*[44] emphasized that a ”pay-per-use” pricing mechanism is regarded as the key characteristic of cloud

computing pricing. The study found that pay-per-use pricing significantly changed the risk-sharing model between the service provider and the customer as the customer's commitment decreased.

4. **Dynamic pricing scheme** Mihailescu *et al.*[45] proposed the dynamic pricing scheme for federated clouds, in which resources are shared among the various service providers in cloud computing. Federated clouds are implemented for improving the reliability and scalability for both service provider and end users. End users in the federated environment are assumed to be capable of both buying and selling resources. As compared to fixed prices, the dynamic pricing that reflects the real-time supply demand relationship represents a more promising charge strategy that can better exploit user payment potentials and thus larger profit gains at the cloud provider[46].

2.3 Homogeneous multiserver systems

Most of the traditional computing processors were homogenous meaning that they had the same characteristics[47]. Homogenous processing environments were overwhelmed by the increasing work load which led to the evolution of heterogeneous processing environments. These have different characteristics for the same processor unlike the homogenous processors. The following conditions should be satisfied for a system to be considered as homogeneous [47]:

1. The hardware or processor ensures the same storage representation.

2. The software module (operating system, compiler, compiler options) on each processor also guarantees the same storage representation.

Homogeneous servers have the same server speed which makes the service rate to be constant, and therefore the load depends on the arrival rate only, on the other hand heterogeneous servers have variable service rates that make the load to depend on both the service rate and arrival rate. In addition, the requirements for a homogeneous computing environment are rigorous and are frequently not met in networks of workstations, even when each computer in the network is the same model. On the other hand, heterogeneous multiserver system exhibits reduced execution time for several tasks with overall increase in performance. Although cloud computing infrastructure assumes homogeneous data centers, the investigation of queueing systems with heterogeneous servers is called for because many queueing systems arise in practice and work at different speeds[46].

2.4 Heterogeneous multiserver systems

Heterogeneous computing refers to those systems that make use of different types of computational units. The computational unit can be a general-purpose processor, a special-purpose processor, or a co-processor. Cloud server systems consist of heterogeneous servers (which means service rate of all servers can be different) because failed or misbehaved servers of a multi-server system are replaced by new and more powerful ones that cause systems to be heterogeneous. The need for increased heterogeneity in the computing systems is partially as the need for high-performance, highly reactive systems that interact with other environments too [47].

2.5 Research gap

Determining what a service provider will receive from an end user in exchange for their services is important. Pricing is considered a critical factor for organizations offering services or products[48]. The price determined for a service or product must consider the manufacturing costs, maintenance costs, market competition, and how the customer values the service or product offered[37]. Hence, developing an appropriate pricing model is important in achieving higher revenues.

Yeoa et al.[49] described the difference between fixed and variable prices. Fixed prices were easier to understand and more straightforward for users. However, fixed pricing could not be fair to all users because not all users had the same needs. Their study proposed charging variable prices with advanced reservation. Charging variable pricing with advanced reservation would let users know the exact expenses that are computed at the time of reservation even though they were based on variable prices. The advantage of advanced reservations is that users can not only know the prices of their required resources in the future but are also able to guarantee access to future resources to better plan and manage their operations.

Mihailescu *et al.*[45] introduced a dynamic pricing scheme for federated clouds, in which resources are shared among many cloud service providers. The authors carried out simulations to determine the efficiency of this approach by comparing it to a fixed pricing scheme. They found that dynamic pricing achieved better average performance with increasing buyer welfare and numbers of successful requests. However, fixed pricing achieved better scalability in the case of high demand in

the market.

Wang *et al.* [50] developed two distributed algorithms for the net profit optimization: Net Profit Optimization for Divisible jobs (NPOD), and Net Profit Optimization for Indivisible Jobs (NPOI). An indivisible job is a job that cannot be interrupted, while a divisible job is one that can be interrupted. The authors proved via simulations that the two algorithms can increase revenues and reduce electricity costs by comparing it to the Largest Job First (LJF) algorithm. However, the authors assumed that the servers at all data centers were homogenous, which does not depict the real cloud server deployment scenarios. The disadvantage of homogeneous multiserver system is that it exhibits increased execution time for several tasks with overall reduction in performance.

Cao *et al.*[10] proposed an optimal multiserver configuration for a cloud computing environment. The pricing mechanism proposed is biased towards the service provider and aims to increase the service provider's revenues[10]. In addition all servers are assumed to be homogeneous which does not depict realistic cloud deployment scenarios.

In an effort to maximize revenue, Feng *et al.*[13] scheduled the Cloud resources among different service instances adaptively based on the dynamically collected information. In the study, each service instance, a virtual machine associated with a user, is modeled as a FIFO (First In First Out) M/M/1/FIFO queue system. The authors proposed two customer-oriented pricing mechanisms; Mean Response Time (MRT) and Instant Response Time (IRT), in which the customers are charged according to achieved service performance in terms of mean response time. The optimal number of servers required to maximize profit was obtained. However, the multiserver system is assumed to be homogeneous.

In an effort to accurately model practical deployment scenarios of cloud servers, Nansamba et al.[21] proposed a pricing model for heterogeneous cloud computing servers based on response time and slowdown. The authors observed that heterogeneous multiserver system generated more revenue than homogeneous multiserver system. However, this study assumed that customers have the same type of application and therefore served using First In First Out policy. The assumption of same type of application heavily constraints the performance and utilization of computing systems since requests are processed upon arrival even if they are delay-tolerant.

Table 2.1 shows a summary of the research gap.

Author	Pricing model	Type of Servers	Type of Service
Wang *et al.* [50]	Distributed algorithms	Homogeneous	Same for all traffic
Cao *et al.* [10]	Service provider oriented	Homogeneous	Same for all traffic
Feng *et al.* [13]	Customer-oriented	Homogeneous	Same for all traffic
Nansamba et al. [21]	Service provider-oriented	Heterogeneous	Same for all traffic
Proposed	Service provider-oriented	Heterogeneous	Differentaited service

Table 2.1: Summary of research gap

2.5.1 Pricing model for heterogenenous multiserver systems

The heterogeneous multiserver system consists of servers with different capabilities in terms of general-purpose processor, special-purpose processor, or a co-processor. The heterogeneous multiserver system can be modeled using the $M/M_i/m$ queue system [21]. In this case the first M denotes Markovian and represents Poisson arrivals into systems, m_i represents the service rates for servers $i = 1, 2, ..., m$. The service rates are exponentially distributed and variable, and depends on the state i in which the system is. The allocation policy in the system is FIFO. The service rate can be defined as shown in equation 2.1.

$$\mu_i = \begin{cases} 0 & i = 0 \\ \mu_1 & i = 1 \\ \mu_1 + \mu_2 & i = 2 \\ . & \\ . & \\ . & \\ \mu_1 + \mu_2 + ... + \mu_m & i \geq m \end{cases} \tag{2.1}$$

In this case, $\mu_1 \geq \mu_2 \geq ...\mu_m$. . Equation 2.2 can be formulated in two ways when the system contains less than m jobs, in which μ_i is a variable and when there are m or more jobs in the system, in which case μ_i is a constant.

Define

$$m_i = \begin{cases} \sum_{j=1}^{i} \mu_j & i < m \\ \\ m_m = \sum_{j=1}^{m} \mu_j & i \geq m \end{cases} \tag{2.2}$$

When $i = 1$, the system is in state 1 and there is 1 job present in the system and only one server is processing the work (this server is assumed to be the fastest).
When $i = 2$, the system is in state 2 and there are 2 jobs present in the system and two servers are processing the work (the two servers are assumed to be the fastest). The mean revenue G brought by a service provision is

$$G = a\left(\frac{P_o.(m_m)^m.\rho^{(m+1)}}{(\pi_{j=1}^{m} m_j).\,(1-\rho)^2}.\frac{1}{\lambda}\right) \tag{2.3}$$

where a is the service charge per unit amount of service, and P_o is as given in equation 2.4.

$$P_o^{-1} = \sum_{i=0}^{m-1}\left(\frac{\lambda^i}{\pi_{k=1}^{i}(\sum_{j=1}^{k}\mu_j)}\right) + \left(\frac{\lambda^m}{(1-\rho)\pi_{j=1}^{m} m_j}\right) \tag{2.4}$$

Equation 2.3 represents the revenue generated from a multiserver heterogenous cloud computing system, however in real cloud computing environments, cloud customers have different type of applications (delay tolerant and delay sensitive) and hence cannot be serviced in the order in which requests arrive. Therefore, we propose to introduce queue management techniques that differentiates task execution via queue reordering to service latency-sensitive requests before serving delay-tolerant requests. In addition, we propose to charge the cloud users who are not willing to tolerate any delay in the completion of their jobs using a standard pricing model in the cloud market while cloud users who can tolerate some delay will be charged lower prices at the expense of delaying job completion time, the more delay a user can tolerate to complete its jobs, the lower the price charged.

Chapter 3

Methodology

This study proposes to use analytical methodology. Analytical methodology is a generic process combining the power of the Scientific Method with the use of formal process to solve any type of problem. An analytical model therefore is a set of computational algorithms or formulae used to analyze systems. Analytical models provide a faster and more computationally efficient methods of obtaining perfomance measures. Simulation on the other hand is the imitation of the operation of a real-world process or system over time. The act of simulating something first requires that a model be developed; this model represents the key characteristics or behaviors/functions of the selected physical or abstract system or process. The model represents the system itself, whereas the simulation represents the operation of the system over time. Therefore, we propose to use analytic models to analyse the proposed system.

3.1 Proposed allocation and pricing model

A cloud service provider serves user's requests by using a multiserver system which is constructed and maintained by an infrastructure vendor and rented by the service provider. We consider a heterogeneous multiserver system where the service rate of servers is different. Customers submit service requests to a service provider, and the service provider serves the requests by using the multiserver system.

The servers are grouped into clusters depending on their speeds and each server can only join one cluster. Every service instance is mapped to a server cluster and each cluster is virtualized as a single machine.

In this model, requests arrive randomly into the system. Requests are classified into delay sensitive class and delay tolerant class and queued in their corresponding queues $(Q_1$ and $Q_2)$. Delay sensitive requests are assigned the servers and served before the delay tolerant requests. The response time for each priority class is determined and it is upon which the price is charged. Algorithm 1 shows the pseudocode for the differentiated heterogeneous multiserver cloud computing system.

Algorithm 1 Pseudocode for the differentiated heterogeneous multiserver cloud computing system

Require: Requests arrive randomly into the system

Require: Requests are classified into differentiated classes
  if Request is delay sensitive **then**
    Assign requests to queue Q_1 and give service immediately
  end if
  if Request is delay tolerant **then**
    Assign requests to Q_2 and give service when Q_1 is empty
  end if
  if Request belongs to Q_1 **then**
    Charge requests basing on mean response for Q_1
  end if
  if Request belongs to Q_2 **then**
    Charge requests basing on mean response time for Q_2
  end if

The charging model will be given as follows:

If the response time r to process a service request is less than T then the customer will pay an amount ar, where a is the service charge per unit amount of service, and T is the benchmark response time in the service level agreement. On the other hand, if the response time r to process a service request is longer than T then the customer will pay an amount $ar(1-d)$, where d indicates the degree of penalty incurred by the service provider for delaying the service. This is represented in equation 3.1.

$$Cost = \begin{cases} ar, & r \leq T \\ ar(1-d), & r > T \end{cases} \tag{3.1}$$

A cloud service provider serves user's requests by using a multiserver system which is con-

structed and maintained by an infrastructure vendor and rented by the service provider. We consider a heterogeneous multiserver system where the service rate of servers is different. Customers submit service requests to a service provider, and the service provider serves the requests by using the multiserver system. The servers are grouped into clusters dynamically and each server can only join one cluster. Every service instance is mapped to a server cluster. Each cluster is virtualized as a single machine. A cloud computing service provider serves users' service requests by using a multiserver system, which is constructed and maintained by an infrastructure vendor and rented by the service provider. Examples are blade servers and blade centers where each server blade is a server, clusters of traditional servers where each server is an ordinary processor [52], [49], and multicore server processors where each server is a single core [51]. We refer to all these blades/processors/cores as servers. Users submit service requests to a service provider, and the service provider serves the requests on a multiserver system.

In addition, we consider a multiserver heterogeneous queueing system in which the arrivals follow a Poisson process with mean arrival rate λ and exponentially distributed inter arrival times. Poisson arrival rates are assumed since the requests into the servers are random and memoryless. Memoryless due to the fact that the arrival of the next request doesnot depend on the arrival of the past requests. The multiserver system maintains a queue with an infinite capacity. The multiserver system is treated as an $M/M_i/m$ queueing system. The $M/M_i/m$ queue model is used to derive the mean revenue brought by a service provision. There are m servers (i.e., blades/processors/cores) with different service rates (measured by the number of packets that can be executed per unit time) μ_m, $(i = 1, 2, ..., m)$ for each of the m servers and the service times at each server follows exponential distribution. Each request requires exactly one server and delay sensitive tasks are served before delay tolerant tasks.

3.2 Model Metrics

The model metric will be revenue. Revenue is the income generated. Revenue has been used in literature as a performance metric to evaluate the performance of different pricing schemes [10, 12]. The revenue will be expressed in terms of performance parameters like task mean response time. A service provider should keep the mean response time to a low level by providing enough servers and/or increasing server speed, and be willing to pay back to a customer in case the mean

slowdown exceeds certain limits.

3.3 Modeling process

In modeling mean response time, we use queueing theory. Queueing models are suitable in a variety of environments ranging from common daily life scenarios to complex service and business processes, operations research problems, or computer and communication systems. Queueing theory has been extensively applied to evaluate and improve system behavior [13, 10]. Specifically, we shall use the $M/M_i/m$ queue system, where M represents Poisson arrival with mean arrival rate (λ) per request with exponentially distributed inter arrival times. Poisson distribution best models random arrivals into systems. Poisson probability distribution is given in [11] as:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}; \quad x = 0, 1, 2, ... \tag{3.2}$$

where

x=number of arrivals in a specific period of time,

$\lambda =$ average, or expected number of arrivals for the specific period of time,

$e = 2.71828.$

M_i represents the service time for server i. The amount of time dedicated to each request on each server is exponentially distributed, with a service rate μ_i, where $i = 1, 2, 3, ..., m$. The number of parallel servers is m. On arrival requests are classified into delay sensitive and delay tolerant and placed into different queues. The queue discipline for each queue is First Come First Served. However, the delay sensitive requests are served before delay tolerant requests. The service intensity, ρ is defined as the ratio of arrival rate to the service rate, $\rho = \lambda/\mu$. The last m represents the number of servers.

Additionally we shall form the optimization problem of the profit generated. The optimization problem will be solved using Lagrange Multiplier Method. In doing this, we shall determine the number of servers that give maximum profit. In mathematical optimization, the method of Lagrange multipliers [53] is a strategy for finding the local maxima and minima of a function subject to constraints. The method of Lagrange multipliers is a powerful tool for solving this class of problems. The number of servers to ensure maximum profit for the service Provider-Oriented pricing model will then be determined.

3.4 Performance Evaluation

To evaluate the performance of the proposed models, we shall:

(i) Analyze the variation of Revenue with execution requirement (i.e., the amount of service),

(ii) Analyze the variation of Revenue with arrival rate of requests,

(iii) Analyze the variation of Revenue with number of servers.

(iv) Analyze the variation of Revenue with the system load.

In all cases, the Revenue under the Differentiated Pricing Scheme will be compared with Revenue under Non Differentiated Pricing Scheme for Heterogeneous Cloud Computing Environments.

3.5 Tools to be used

We employ Monte Carlo simulation which is a package of MATLAB to simulate the $M/M_i/m$ queuing system. Monte Carlo simulation involves the use of random sampling techniques and often the use of computer simulation to obtain approximate solutions to mathematical or physical problems especially in terms of a range of values each of which has a calculated probability of being the solution.
Monte Carlo simulation is widely used in a wide range of areas for statistical simulation. In communication engineering, this technique is basically used to simulate the behavior of a given communication system using artificial means but based on input statistics similar to those of the inputs used in practice [54].

During a Monte Carlo simulation, values are sampled at random from the input probability distributions. Each set of samples is called an iteration, and the resulting outcome from that sample is recorded. Monte Carlo simulation does this hundreds or thousands of times, and the result is a probability distribution of possible outcomes. In this way, Monte Carlo simulation provides a much more comprehensive view of what may happen. Monte Carlo simulation indicates not only what could happen, but how likely it is to happen. Monte Carlo simulation provides a number of advantages over deterministic, or single-point estimate analysis, these include:

1. Probabilistic Results: Results show not only what could happen, but how likely each outcome is.

2. Graphical Results: Because of the data a Monte Carlo simulation generates, its easy to create graphs of different outcomes and their chances of occurrence.

3. Sensitivity Analysis: With just a few cases, deterministic analysis makes it difficult to see which variables impact the outcome the most. In Monte Carlo simulation, it's easy to see which inputs had the biggest effect on bottom-line results.

4. Scenario Analysis: In deterministic models, it's very difficult to model different combinations of values for different inputs to see the effects of different scenarios. Using Monte Carlo simulation, analysts can see exactly which inputs had which values together when certain outcomes occurred. This is invaluable for pursuing further analysis.

5. Correlation of Inputs: In Monte Carlo simulation, it's possible to model interdependent relationships between input variables. It's important for accuracy to represent how, in reality, when some factors goes up, others go up or down accordingly.

Appendices

A. Research Time Frame

NO	ACTIVITY	PROPOSED START DATE	END DATE
1	Development of Concept paper	October 2017	December 2017
1	Proposal Writing and Approval	December 2017	February 2018
2	Methodology, Model development	March 2018	April 2018
3	Model Validation	April 2018	April 2018
4	Report Writing and Handing in	May 2018	May 2018

B. Budget (Uganda Shillings)

QTY	ITEM	UNIT COST	TOTAL
1	Laptop	$2,000,000$	$2,000,000$
1 Ream	Stationary	$10,000$	$10,000$
100	Printing	200	$200,000$
4	Hard cover Binding	$10,000$	$40,000$
Lump sum	Transport		$600,000$
	Subtotal		2,850,000
1	Contingency (10%)		$285,000$
	GRAND TOTAL		$3,135,000$

References

[1] E. Gorelik, Cloud Computing Models, Masters thesis, Management and the MIT Engineering Systems Division, Massachusetts Institute of Technology, 2013.

[2] I. Foster, I. Yong, Z. Raicu and S. Lu, ”Cloud Computing and Grid Computing 360-Degree Compared”, Grid Computing Environments Workshop, 2008.

[3] D.P. Acharjya, S. Dehuri, and S. Sanyal, Computational Intelligence for Big Data Analysis, Frontier Advances and Applications, Springer International Publishing, Switzerland, 2015, pages 201-203.

[4] V. Uma, V.J. Suseela, Current Practices in Academic Librarianship, publisher by Allied Publishers PVT, Ltd, 2014, pages 245-248.

[5] R. Buyya, D. Abramson, J. Giddy, and H. Stockinger, Economic models for resource management and scheduling in grid computing, in Proc. Concurrency and Computation: Practice and Experience, vol. 14, 2007, pp. 15071542.

[6] S. Dutta, M. Zbaracki, and M. Bergen, Pricing Process as a Capability: A Resource-Based Perspective, in Proc. Strategic Management Journal, vol. 27, no. 7, 2003.

[7] C. Weinhardt, A. Anandasivam, B. Blau, N. Borissov, T. Meini, W. Michalk, and J. Stossern, Cloud Computing-A Classification, in Proc. Business Models, and Research Directions, Bus. Models and Information System Engineering , vol. 1, no. 5, 2009.

[8] R. Sahal and M.H. Khafagy and F. A. Omara, A Survey on SLA Management for Cloud Computing and Cloud-Hosted Big Data Analytic Applications, International Journal of Database Theory and Application, Vol. 9. No. 3, 2016, pp. 107-118.

[9] M. Firdhous and S. Hassan and O. Ghazali, Monitoring, Tracking and Quantification of Quality of Service in Cloud Computing, Proc. International Journal of Scientific & Engineering Research, Vol.4 No. 5, 2013, pp.112-117.

[10] J. Cao, K. Hwang, K. Li, and A. Zomaya, Optimal Multiserver Configuration for Profit Maximization in Cloud Computing, inProc. IEEE Transactions on Parallel & Dis-tributed Systems , vol. 24, no. 6, Jun. 2013, pp. 1087-1096.

[11] L. Kleinrock, Queueing Systems, Volume I& II. Computer Applications, John Wiley & Sons, 1976.

[12] L. Zhang and D. Ardagna, SLA Based Profit Optimization in Autonomic Computing Systems, in Proc. the 2nd International Conference on Service Oriented Computing, 2004, pp. 173182.

[13] G. Feng, S. Garg, R. Buyya, and W. Li, Revenue Maximization Using Adaptive Re-source Provisioning in Cloud Computing Environments, pp. 192200, 2012.

[14] H. S. Narman, M. S. Hossain, and M. Atiquzzaman, h-DDSS: Heterogeneous Dynamic Dedicated Servers Scheduling in Cloud Computing, in Proc.IEEE International Conference on Communications, Jun. 2014, pp. 3475-3480.

[15] S. Crago, K. Dunn, P. Eads, L. Hochstein, D. Kang, M. Kang, D. Modium, K. Singh, J. Suh, J. P Walters, Heterogeneous cloud computing, Proc. in IEEE International Conference on Cluster Computing, 2011, pp. 378-385.

[16] X. Wu, F. De Pellegrin, On the Benefits of QoS-Differentiated Posted Pricing in Cloud Computing: An Analytical Model, Journal of Latex, Vol. 14, No. 8, 2015.

[17] P.K. Suri and S. Mittal, A Comparative Study of various Computing Processing Environments: A Review, Proc. International Journal of Computer Science and Information Technologies, Vol 3 No. 5, 2012, pp.5215-5218.

[18] D. Lo, L. Cheng, R. Govindaraju, P. Ranganathan, and C. Kozyrakis. ”Improving Resource Efficiency at Scale with Heracles.” In ACM Transactions on Computer Systems, 2016.

[19] L. Zheng, C. Joe-Wong, C. G. Brinton, C. W. Tan, S. Ha, and M. Chiang. ”On the Viability of a Cloud Virtual Service Provider.” In ACM SIGMETRICS, 2016.

[20] J. N. Daigle. ”The Basic M/G/1 Queueing System. Queueing Theory with Applications to Packet Telecommunication, pp. 159-223. Springer, 2005.

[21] B. Nansamba, K. S. Kaawaase, M. Okopa and B. K. Asingwire. Pricing Scheme for Heterogeneous Multiserver Cloud Computing System. Australasian Journal of Computer Science, Vol. 4, No. 1, 2017, pp 32-43.

[22] J. Rasley, K. Karanasos, S. Kandula, R. Fonseca, M. Vojnovic, and S. Rao. ”Efficient Queue Management for Cluster Scheduling.” In ACM EuroSys, 2016.

[23] S. K. Baruah, and J. R. Haritsa. Scheduling for Overload in Real-Time Systems. IEEE Transactions on Computers, 1997.

[24] J. Chen, C. Wang, B. B. Zhou, L. Sun, Y. C. Lee, and A. Y. Zomaya, Tradeoffs between profit and customer satisfaction for service provisioning in the cloud, in ACM HPDC, 2011.

[25] L. M. Vaquero, L Rodero-Merino, J. Caceres and M. Lindner, ”A Break in the Clouds”, Proc. Towards a Cloud Definition, Comput. Commun. Review, volume 39, No. 1, 2009.

[26] F. Etro, The Economic Impact of Cloud Computing on Business Creation, Employment and Output in Europe, pp. 179208, Jun. 2009.

[27] World Economic Forum, ”Exploring the future of cloud computing: Riding the next wave of technology-driven transformation.” Journal of World Economic Forum In partnership with Accenture Report, 2010, pp. 179-208.

[28] P. Mell and T. Grance, ”The NIST Definition of Cloud Computing”, Version 15, National Institute of Standards and Technology (NIST), Oct, 2009.

[29] W. Voorsluys, J. Broberg, and R. Buyya, ”Introduction to cloud computing,” Cloud Computing, pp.1-41, 2011.

[30] I. Ashraf, ”An Overview of Service Models of Cloud Computing.” International Journal of Multidisciplinary and Current Research, volume 2, August, 2014.

[31] D. Ma and J. Huang, ”The Pricing Model of Cloud Computing Services”, International Conference on Electronic Commerce, Singapore Management University, 2012, pp.6-8.

[32] S. Goyal, ”Public vs Private vs Hybrid vs Community-Cloud Computing: A Critical Review”, Proc. International Journal of Computer Network and Information Security, 2014.

[33] W. Jansen and T. Grance, ”Guidelines on security and privacy in public cloud computing”. NIST special publication 800-144, 2011.

[34] P. Jain, ”Cloud Computing and It’s Types in Mobile Network”, Proc. International Journal of Science and Research”, 2013, pp. 71-73.

[35] H. Wang, Q. Jing, R. Chen, B. He, Zh. Qian and L. Zhou, ”Distributed Systems Meet Economics: Pricing in the Cloud,” June , 2010.

[36] M. Al-Roomi, Sh. Al-Ebrahim, S. Buqrais and I. Ahmad, ”Cloud Computing Pricing Models: A Survey,” International Journal of Grid and Distributed Computing Vol.6, No.5, pp.93-106, 2013.

[37] , M. Al-Roomiand, S. Al-Ebrahim, S. Buqrais and I. Ahmad, ”Cloud Computing Pricing Models: A Survey”, Proc. International Journal of Grid and Distributed Computing, Volume 6, No. 5, 2013, pp. 93-106.

[38] B. Sharma, R. Thulasiram, P. Thulasiraman, S. Garg and R. Buyya, ”Pricing Cloud Compute Commodities: A Novel Financial Economic Model”, Proc. IEEE/ACM Int. Symp. on Cluster, Cloud and Grid Computing, 2012.

[39] Mazrekaj, Artan, I. Shabani and B. Sejdiu, ”Pricing Schemes in Cloud Computing: An Overview”, Proc. International Journal of Advanced Computer Science and Application, 2016, pp.80-86.

[40] E. Iveroth, A. Westelius, C. Petri, N. Olve, M. Coster and F. Nilsson, ”How to Differentiate by Price: Proposal for a Five-Dimensional Model,” Proc. European Management Journal, 2012.

[41] ”Amazon Web Services”, http://aws.amazon.com/. Accessed 14th January 2018.

[42] Google App Engine, https://appengine.google.com/. Accessed 19th January 2018.

[43] Windows Azure, http://www.windowsazure.com/en-us/. Accessed 14th February 2018.

[44] J. Jtmaa, ”Financial Aspects of Cloud Computing Business Models”, Journal of Information Systems Science, 2011.

[45] M. Mihailescu and Y. Teo, ”Dynamic Resource Pricing on Federated Clouds”, ”Proc. 10th IEEE/ACM International Symposium on Clustered Cloud and Grid Computing, 2010.

[46] J. Zhao, H. Li, C. Wu, Z. Li, Z. Zhang and F. C. M. Lau, ”Dynamic Pricing and Profit Maximization for the Cloud with Geo-distributed Data Centers”, Proc. INFOCOM, 2014.

[47] P.K. Suri and S. Mittal, ”A Comparative Study of various Computing Processing Environments: A Review,” Proc. International Journal of Computer Science and Information Technologies, Vol. 3. No. 5, 2012, pp. 5215-5218.

[48] S. Dutta, M. Zbaracki and M. Bergen, ”Pricing Process as a Capability: A Resource-Based Perspective”, Proc. Strategic Management Journal, volume 27, number 7, 2003.

[49] C. S. Yeoa and S. Venugopalb and X. Chua and R. Buyyaa, ”Autonomic Metered Pricing for a Utility Computing Service”, Future Generation Computer Systtem, volume 26, No. 8, 2010.

[50] W. Wang and P. Zhang and T. Lan and V. Aggarwal, ”Datacenter Net Profit Optimization with Individual Job Deadlines”, ”Proc. Conference on Inform. Sciences and Systems, 2012.

[51] K. Li, ”Optimal load distribution for multiple heterogeneous blade servers in a cloud computing environment”, Proc. the 25th IEEE International Parallel and Distributed Processing Symposium Workshops, May, 2011, 943-952.

[52] J. Sherwani, N. Ali, N. Lotia, Z. Hayat and R. Buyya, ”Libra: a computational economy-based job scheduling system for clusters”, Proc. Software-Practice and Experience, Volume 34, May, 2004, 573-590.

[53] J. Lagrange, ”Mcanique Analytiques”, sect. IV, 2 vols. Paris 1811 h.

[54] I. A. Akyildiz, Brandon. F. Lo and R. Balakrishnan, ”Cooperative Spectrum Sensing in Cognitive Radio Networks: A Survey”, Proc. Physical Communication Journal, Volume 4, 2011, 40-62.