OPEN ACCESS Australasian Journal of Computer Science

ISSN 2251-3221
DOI: 10.3923/aujcs.2018.13.25

Research Article
Modeling Differentiated Pricing Scheme for Heterogeneous Cloud Computing Environments

[1]Nankya Mariam, [1]Drake Patrick Mirembe and [2]Michael Okopa

[1]Makerere University Kampala, Uganda
[2]Faculty of Science and Technology, Cavendish University, P.O. Box 33145, Kampala (U), Uganda

Abstract

Background and Objective: Pricing in cloud computing environments assumes the same type of users and this constrains the performance and utilization of computing systems since requests are processed upon arrival even if they are delay-tolerant. The objective of this study was to model a pricing scheme in which cloud users 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 and those cloud users who are willing to tolerate delay are charged lower prices at the expense of delaying packet completion time. **Materials and Methods:** To overcome the above challenge, this study proposed a pricing scheme that charges different prices for different users depending on the time sensitivity of the request. The proposed pricing scheme is modeled using a multiserver system which is treated as an M/Mi/m queueing system, where M stands for Markovian and represents arrivals that follow a Poisson distribution; Mi stands for Markovian service time that follows an exponential distribution with multiservers, m represents the number of servers. The performance of the differentiated pricing scheme was compared to the pricing scheme with no differentiation using MATLAB. **Results:** Numerical results show that the derived models can provide price differentiation resulting into delay tolerant packets paying less while the delay sensitive packets result in paying more. The price differentiation was more pronounced at high load and high arrival rate values. It was further observed that increase in load and arrival rate increased revenue. For low load and low arrival rate values price differentiation had little effect on revenue. Additionally, it is observed that the more the servers, the more the revenue generated. **Conclusion:** It was concluded that the proposed scheme provided differentiated pricing in which real time packets result in paying more with less delay and non real time packets result in paying less at the expense of delaying its packets.

Key words: Cloud computing, delay sensitive, delay tolerant, pricing mechanism, high load and high arrival rate values

Citation: Nankya Mariam, Drake Patrick Mirembe and Michael Okopa, 2018. Modeling differentiated pricing scheme for heterogeneous cloud computing environments. Australasian J. Comp. Sci., 5: 13-25.

Corresponding Author: Michael Okopa, Faculty of Science and Technology, Cavendish University, P.O. Box 33145, Kampala (U), Uganda
Tel: +256772617705

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

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[1].

Cloud computing is defined 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[2-4].

Pricing is a critical factor for organizations offering cloud services or products[5]. Pricing is the process of determining what a service provider will receive from an end user in exchange for their services. Cloud computing success can be obtained only by developing adequate pricing techniques[6]. An agreement between a customer and a provider to receive a particular service provision is presented in a Service Level Agreement (SLA)[7]. The 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)[8].

Studies on pricing in cloud computing environments assume cloud servers are homogeneous i.e., they have the same characteristics[9-11]. Homogeneous cloud computing servers consist of same storage capacity, processing power, energy supply and same service rate. However, cloud server systems consist of heterogeneous servers with different service rates and storage capacities[12]. 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[13].

The performance and utilization of cloud computing systems are heavily constrained by the characteristics of jobs being served (e.g., their sensitivity to latency)[14]. 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.

Pricing is considered a critical factor for organizations offering services or products[15]. 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[16].

Yeoa *et al.*[17] 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 and Teo[18] 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.

Cao *et al.*[9] 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[14]. 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.*[11] scheduled the cloud resources among different service instances adaptively based on the dynamically collected information. 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 multi-server system was assumed to be homogeneous.

To accurately model practical deployment scenarios of cloud servers, Nansamba *et al.*[19] 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 FIFO

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.

In practice, both latency-critical and delay-tolerant requests coexist in the cloud[14]. Theoretical and experimental results showed that latency-sensitive requests induce a low resource utilization of cloud resources of between[20] 6 and 12%. Conversely, delay-tolerant requests tend to lead to a much higher utilization[20-22]. In cloud computing, differentiated service is used to give different services to different classes of customers. 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.

MATERIALS AND METHODS

Nansamba *et al.*[19] proposed a pricing scheme for heterogeneous multiserver system and modeled it using the M/Mi/m queue system. In this case the first M denotes Markovian and represents Poisson arrivals into systems, Mi 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 mean revenue G brought by a service provision in terms of mean response time is derived as[19]:

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

where, a is the service charge per unit amount of service and Po is as given in Eq. 2:

$$P_o = \left(\sum_{j=0}^{m-1}\frac{(m\rho)^j}{j!} + \frac{(m\rho)^m}{m!(1-\rho)}\right)^{-1} \quad (2)$$

In the same vein, the mean revenue G brought by a service provision in terms of mean slowdown is derived as[19]:

$$G = a\left(\frac{\frac{P_o(m_m)^m.\rho^{(m+1)}}{(\pi_{j=1}^m m_j)(1-\rho)^2}\cdot\frac{1}{\lambda}}{\frac{1}{\mu}}\right) \quad (3)$$

Where:

$$P_o^{-1} = \sum_{i=0}^{m-1}\left(\frac{\lambda^i}{\pi_{k=1}^i\left(\sum_{j=1}^k \mu_j\right)}\right) + \left(\frac{\lambda^m}{(1-\rho)\pi_{j=1}^m m_j}\right) \quad (4)$$

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 FIFO policy. The assumption of same type of users heavily constrains the performance and utilization of computing systems since requests are processed up on arrival even if they are delay-tolerant.

This study used analytical methodology to evaluate the performance of the pricing models. 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 provided faster and more computationally efficient methods of obtaining performance measures.

In particular, a multiserver heterogeneous queueing system in which the arrivals follow a Poisson process with mean arrival rate λ and exponentially distributed inter arrival times is considered. 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 does not 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/Mi/m queueing system. The M/Mi/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) 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.

The study considered a cloud service provider where user's requests are served by using a heterogeneous multiserver system which is constructed and maintained by an infrastructure vendor and rented by the service provider. 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 (Q1 and Q2). 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. The charging model is 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 was shown by the flow chart in Fig. 1 and the pricing scheme is represented in Eq. 5:

$$\text{Cost} = \begin{cases} ar & r \le T \\ ar(1-d) & r > T \end{cases} \quad (5)$$

Derivation of revenue for delay sensitive packets under heterogeneous multiserver system in terms of mean response time: Assume 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/Mi/m queue system. The delay sensitive packet experiences the following delays:

- Delay due to sensitive packets found in the queue
- Delay due to packets found in service (residual service time)

The service rate can be defined as shown in Eq. 6:

$$\mu_i = \begin{cases} 0 & i = 0 \\ \mu_1 & i = 1 \\ \mu_1 + \mu_2 & i = 2 \\ . & \\ . & \\ . & \\ \mu_1 + \mu_2 + ... + \mu_m & i \ge m \end{cases} \quad (6)$$

In this case, $\mu_1 \ge \mu_2 \ge ..., \mu_m$. Equation 6 can be formulated in two ways when the system contains less than m requests, in which μ_i is a variable and when there are m or more requests in the system, in which case μ_i is a constant.

Basing on the pricing model proposed by Nansamba *et al.*[19], the mean revenue G brought by a service provision in terms of mean response time is given as in Eq. 1 and P_o is as given in Eq. 2.

Derivation of revenue for delay sensitive packets under heterogeneous multiserver system in terms of mean slowdown: In this section, the expression for revenue in terms of mean slowdown is derived. Using the expression for the average request response time given in Nansamba *et al.*[19] and the definition of mean slowdown being the ratio of average task response time to the size of the request, the expression for mean slowdown can be given as: $\frac{\overline{T}}{x}$, where, $\overline{T}$ is the average task response time, x is the size of the request. x can also be expressed as the reciprocal of average service rate, $\frac{1}{\mu}$. Therefore mean slowdown is given as:

$$\overline{S} = \frac{\frac{P_o (m_m)^m . \rho^{(m+1)}}{(\pi_{j=1}^{m} m_j)(1-\rho)^2} \cdot \frac{1}{\lambda}}{\frac{1}{\mu}} \quad (7)$$

Basing on pricing model in Nansamba *et al.*[19], the mean revenue G in terms of mean slowdown brought by a service provision for a delay sensitive packet is given as in Eq. 3 and P_o is as given in Eq. 4.

Derivation of revenue for delay tolerant packets under heterogeneous multiserver system in terms of mean response time: In this study, models for the revenue generated for delay tolerant packets are derived. Assuming a tagged delay tolerant packet in the queue. This packet will be delayed by:

- Mean residual time of the packets found in service
- Mean waiting time of delay sensitive packets found in the queue
- Mean waiting time of delay tolerant packets found in the queue
- Mean waiting time of subsequent arrivals of delay the tagged is waiting in the queue for service

The expression for the mean response time for the delay tolerant packets can be derived as:

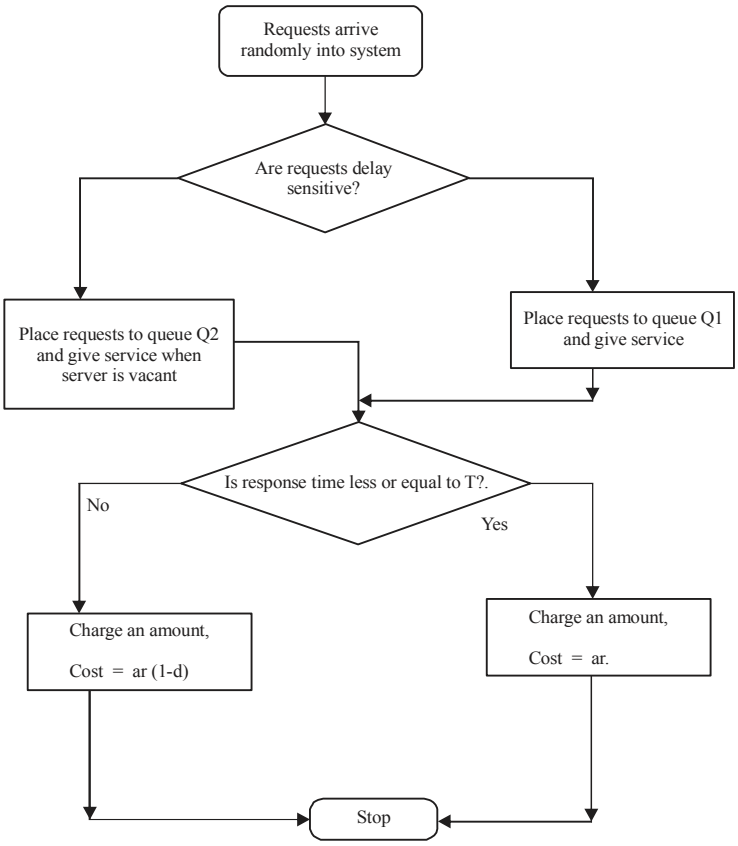

Fig. 1: Flow chart of the system model

$$W^T = R + \frac{1}{\mu}N^S + \frac{1}{\mu}N^T + \frac{1}{\mu}\lambda^S W^T$$
$$= R + \frac{1}{\mu}N^S + \frac{1}{\mu}N^T + \frac{1}{\mu}\lambda^S W^T$$
$$= R + \frac{1}{\mu}\lambda^S W^S + \frac{1}{\mu}\lambda^T N^T + \frac{1}{\mu}\lambda^S W^T$$
$$= R + \rho^S W^S + \rho^T W^T + \rho^S W^T$$

Where:

- W^T = Average waiting time of delay tolerant packets
- W^S = Average waiting time of delay sensitive packets
- N^S = Average number of delay sensitive packets in the queue
- N^T = Average number of delay tolerant packets in the queue
- R = Mean residual service time
- ρ^S = Load due to delay sensitive packets
- ρ^T = Load due to delay tolerant packets
- λ^S = Arrival rate of delay sensitive packets
- λ^T = Arrival rate of delay tolerant packets

$$W^T = \frac{\left(R + \rho^S W^S\right)}{1 - \rho^S - \rho^T} \quad (8)$$

From Eq. 1, W^S can be expressed as:

$$\overline{W^S} = \frac{P_o (m_m)^m . \rho^{(m+1)}}{(\pi_{j=1}^m m_j)(1-\rho)^2} \cdot \frac{1}{\lambda} \quad (9)$$

Where:

$$P_o^{-1} = \sum_{i=0}^{m-1}\left(\frac{\lambda^i}{\pi_{k=1}^i\left(\sum_{j=1}^k \mu_j\right)}\right) + \left(\frac{\lambda^m}{(1-\rho)\pi_{j=1}^m m_j}\right) \quad (10)$$

Substituting W^S from Eq. 9 in Eq. 8, the mean response time for delay tolerant packets can be expressed as:

$$W^T = \left(\frac{(R + \rho^S(\frac{P_o (m_m)^m . \rho^{(m+1)}}{(\pi_{j=1}^m m_j)(1-\rho)^2} \cdot \frac{1}{\lambda}))}{(1 - \rho^S - \rho^T)}\right) \quad (11)$$

Basing on the pricing model proposed in Eq. 5, the mean revenue in terms of mean response time G brought by a service provision for delay tolerant packets is:

$$G = \left(\frac{\alpha(1-d)(R + \rho^{S}(\frac{P_o(m_m)^m.\rho^{(m+1)}}{(\pi_{j=1}^{m} m_j)(1-\rho)^2}.\frac{1}{\lambda}))}{(1-\rho^{S}-\rho^{T})} \right) \quad (12)$$

where, α is the service charge per unit amount of service and P_o is as given in Eq. 10.

Derivation of revenue for delay tolerant packets under heterogeneous multiserver system in terms of mean slowdown: The expression for delay tolerant packets under heterogeneous multiserver system in terms of mean slowdown is derived. Using the expression for the average waiting time given in Nansamba *et al.*[19] and the definition of mean slowdown being the ratio of average task response time to the size of the request, the expression for mean slowdown can be given as: $\frac{\overline{T}}{x}$, where, x is the size of the request. The x can also be expressed as the reciprocal of average service rate, $\frac{1}{\mu}$. Therefore mean slowdown is given as:

$$\overline{S} = \frac{W^T}{1/\mu} \quad (13)$$

Using Eq. 13, the expression for the mean slowdown for the delay sensitive packet can be expressed as:

$$\overline{S} = \frac{(R + \rho^{S}(\frac{P_o(m_m)^m.\rho^{(m+1)}}{(\pi_{j=1}^{m} m_j)(1-\rho)^2}.\frac{1}{\lambda}))}{\mu(1-\rho^{S}-\rho^{T})} \quad (14)$$

Basing on pricing model proposed in Eq. 5, the mean revenue in terms of mean slowdown G brought by a service provision for delay tolerant packets is:

$$G = \left(\frac{\alpha(1-d)(R + \rho^{S}(\frac{P_o(m_m)^m.\rho^{(m+1)}}{(\pi_{j=1}^{m} m_j)(1-\rho)^2}.\frac{1}{\lambda}))}{\mu(1-\rho^{S}-\rho^{T})} \right) \quad (15)$$

where, a is the service charge per unit amount of service and P_o is as given in Eq. 10.

RESULTS

For numerical evaluation, the following hypothetical data was considered but there is conformity with the standard results available in literature. The parameters are in conformity with those used in literature[9,19]. The number of servers has been fixed to 5 with servers of capacities $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, $\mu_4 = 4$, $\mu_5 = 5$. The total service rate for 5 heterogeneous servers is 15 packets/second (1+2+3+4+5). The arrival rate of delay sensitive packets is varied from 0-7 packets sec^{-1} to ensure that the load due to delay sensitive packets is 0.5 with service rate of 15 packets sec^{-1}. The total load in the system is 0.95 which represents a highly utilized system. The load due to delay sensitive packets is varied from 0-0.45 to bring the maximum load to 0.95 when combined with load due to delay tolerant packets. The price constant for service instance, a is set at 10 cents. The monetary unit"cent" in this study may not be identical but should be linearly proportional to the real cent in US dollars[9].

Comparison of revenue for delay sensitive and delay tolerant packets in terms of load and arrival rate: In this section the variation of revenue with load and arrival rate for delay sensitive and delay tolerant packets in terms of response time are investigated. In doing this the effect on revenue of increasing load and arrival rate are investigated.

Variation of revenue with load: The effect of increasing load of delay sensitive packets on revenue for delay sensitive and delay tolerant packets are investigated. In doing this, Eq. 1 and 12 are used to plot the graph of revenue as a function of load due to delay sensitive packets as illustrated in Fig. 2.

In this case the heterogeneous multiserver system has five servers with service rates of 1 request sec^{-1}, 2 requests sec^{-1}, 3 requests sec^{-1}, 4 requests sec^{-1}, 5 requests sec^{-1}. It is observed that revenue increases with increase in load for both delay sensitive and delay tolerant packets. It is further observed that delay sensitive packets generate more revenue than delay tolerant packets for higher values of load due to delay sensitive packets. For low load due to delay sensitive packets the revenue generated for delay sensitive and delay tolerant packets are almost the same.

Variation of revenue with arrival rate: The effect of increasing arrival rate on revenue for delay sensitive and delay tolerant packets for heterogeneous multiserver systems in terms of mean response time is investigated.

Figure 3 showed a graph of revenue as a function of arrival rate of delay sensitive packets in terms of mean response time. In doing this, Eq. 1 and 12 are used to plot the graph of revenue as a function of arrival rate. It is

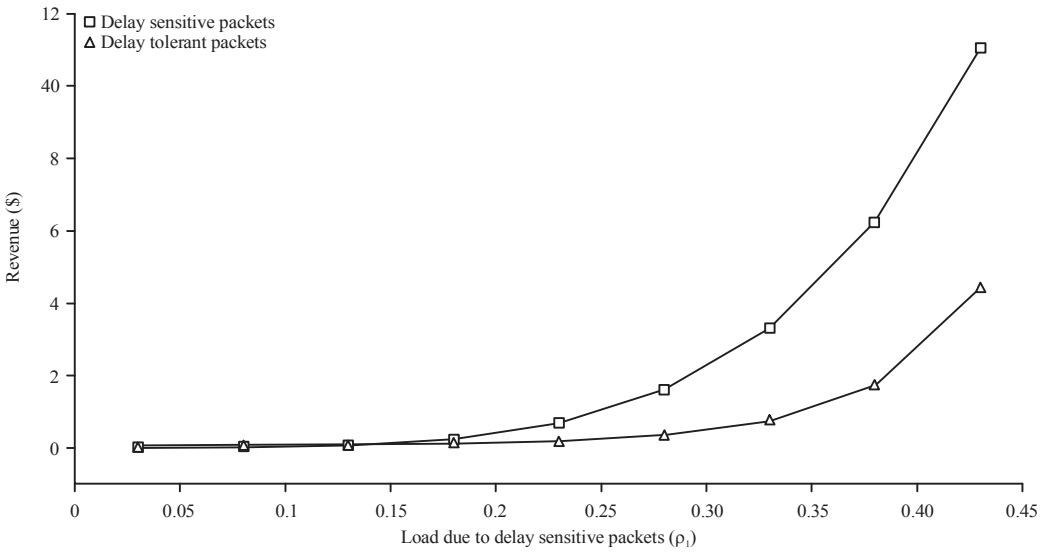

Fig. 2: Revenue vs. load of delay sensitive packets in terms of mean response time

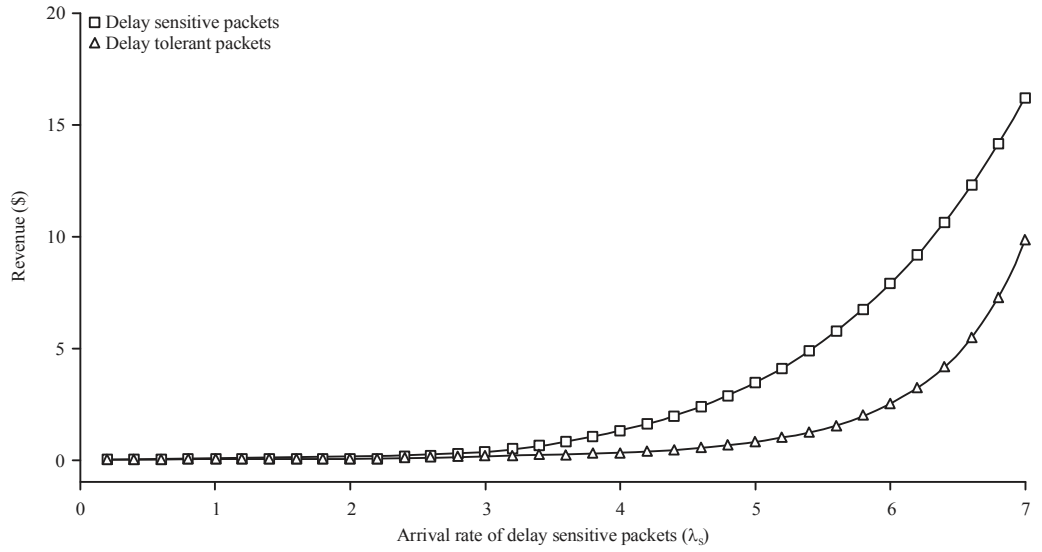

Fig. 3: Variation of revenue with arrival rate of delay sensitive packets in terms of mean response time

observed that revenue generally increases with increase in arrival rate for all considered types of packets. It is further observed that for low arrival rate values, the revenue generated for both delay sensitive and delay tolerant are the same, however as the arrival rate increases, delay sensitive packets generate more revenue than delay tolerant packets. The difference in revenue is more pronounced at higher arrival rate values. Therefore, service differentiation is more effective at higher arrival rate values of delay sensitive packets.

Comparison of revenue with load and arrival rate for different number of servers: The variation of revenue with load and arrival rate for different numbers of servers is investigated. In doing this, it investigated the effect of number of servers on revenue for delay sensitive and delay tolerant packets.

Variation of revenue with load for varying number of servers in terms of mean response time: The effect of increasing number of servers on revenue for a heterogeneous multiserver system in terms of mean response time is investigated.

Figure 4 showed a graph of revenue as a function of load of delay sensitive packets in terms of mean response time with varying number of servers. In doing this,

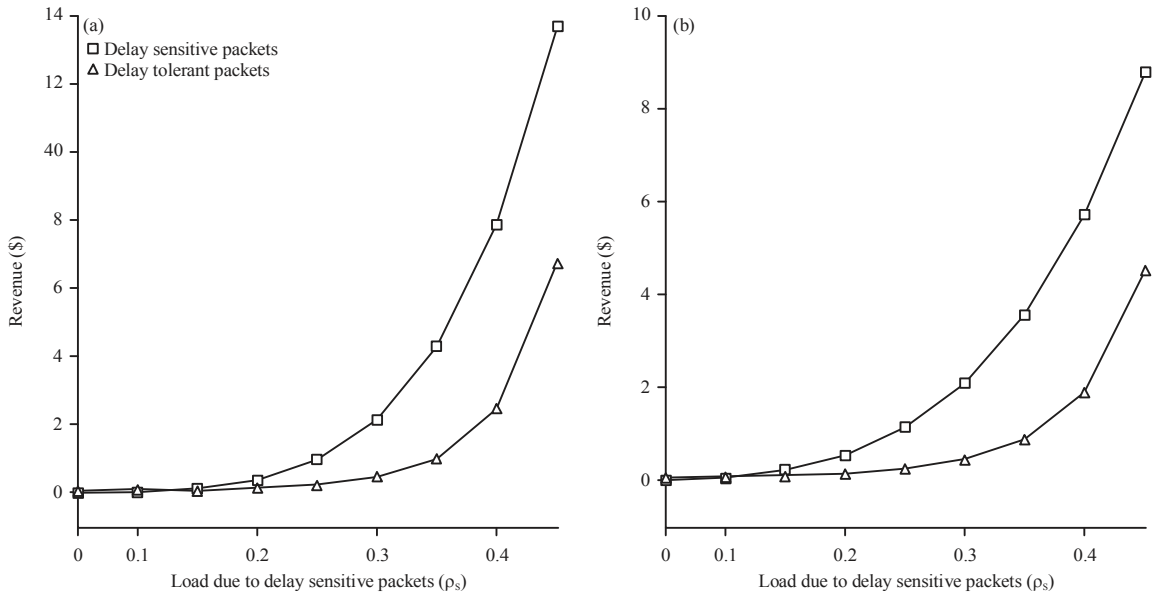

Fig. 4(a-b): Variation of revenue with load of delay sensitive packets in terms of mean response time with varying number of servers, (a) Revenue for 5 servers and (b) Revenue for 4 servers

Eq. 1 and 12 are used to plot the graph of revenue as a function of load. It is observed that revenue generally increases with increase in load for both considered number of servers. It is also observed that for low load values the revenue generated is almost the same for both delay sensitive and delay tolerant packets, however as the load increases the revenue generated is higher. It is also observed that for any considered load values, higher numbers of servers generate more revenue than lower number of servers as indicated in Fig. 4a, b. This observation implied that deploying higher number of servers is more effective at higher load values.

Variation of revenue with arrival rate for varying number of servers in terms of mean response time: The effect of increasing arrival rate for varying number of servers on revenue in terms of mean response time is investigated.

Figure 5 showed a graph of revenue as a function of arrival rate of delay sensitive packets with varying number of servers. In this case the revenue generated is determined for a heterogeneous multiserver system with five servers (with service rates of 1 request sec^{-1}, 2 requests sec^{-1}, 3 requests sec^{-1}, 4 requests sec^{-1} and 5 requests sec^{-1}) and a heterogeneous multiserver system with four servers (with service rates of 1 request sec^{-1}, 2 requests sec^{-1}, 3 requests sec^{-1} and 4 requests sec^{-1}). In doing this, Eq. 1 and 12 are used to plot the graph of revenue as a function of arrival rate. It is observed that revenue generally increases with increase in arrival rate for both considered number of servers. It is further observed that for low arrival rate values the revenue generated is almost the same for delay sensitive and delay tolerant packets; however as the arrival rate increases the revenue generated is higher for delay sensitive packets than for delay tolerant packets. It is also observed that higher number of servers generate more revenue than lower number of servers for the same arrival rate values as indicated in Fig. 5a, b. This observation implies that deploying higher number of servers is more effective at higher arrival rate values.

Comparison of revenue with load and arrival rate in terms of mean slowdown: The variation of revenue with load and arrival rate for in terms of mean slowdown is investigated. In doing this, the effect of arrival rate and load on revenue for delay sensitive and delay tolerant packets are investigated.

Variation of revenue with load in terms of mean slowdown: The effect of increasing load on revenue for delay sensitive and delay tolerant packets in terms of mean slowdown is investigated.

Figure 6 showed the graph of revenue as a function of load due to delay sensitive packets for heterogeneous multiserver system in terms of mean slowdown. In doing this, Eq. 3 and 15 were used to plot the graph of revenue as a

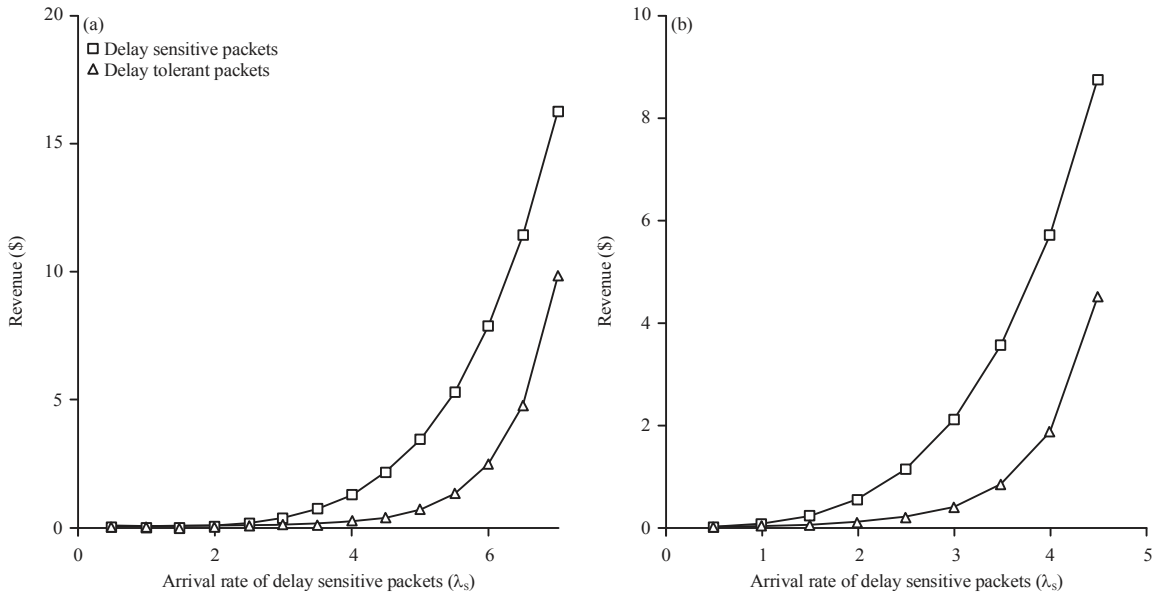

Fig. 5(a-b): Variation of revenue with arrival rate of delay sensitive packets for varying number of servers in terms of mean response time, (a) Revenue for 5 servers and (b) Revenue for 4 servers

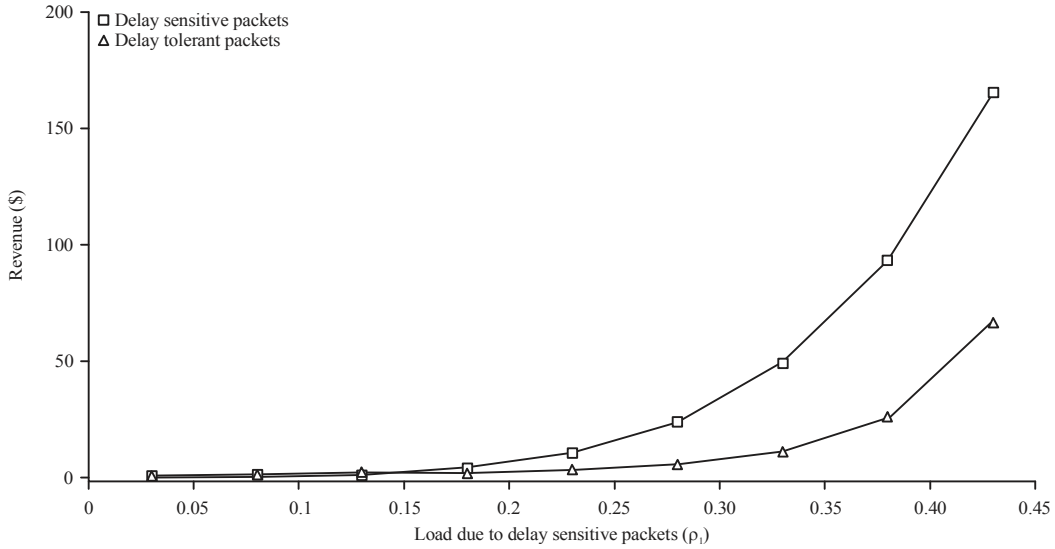

Fig. 6: Variation of revenue with load of delay sensitive packets in terms of mean slowdown

function of load. It is observed that revenue generally increases with increase in load for both considered types of packets. It is further observed that for low load values the revenue generated is almost the same for delay sensitive and delay tolerant packets, however as the load increases the revenue generated is higher for delay sensitive packets compared to delay tolerant packets. The difference in revenue between delay sensitive and delay tolerant packets is more pronounced at higher values of load.

Variation of revenue with arrival rate in terms of mean slowdown: The effect of increasing arrival rate on revenue for delay sensitive and delay tolerant packets in terms of mean slowdown is investigated.

Figure 7 showed a graph of revenue as a function of arrival rate for delay sensitive and delay tolerant packets in terms of mean slowdown. Equation 3 and 15 were used to plot the graph of revenue as a function of arrival rate. It is observed that revenue generally increases with increase in

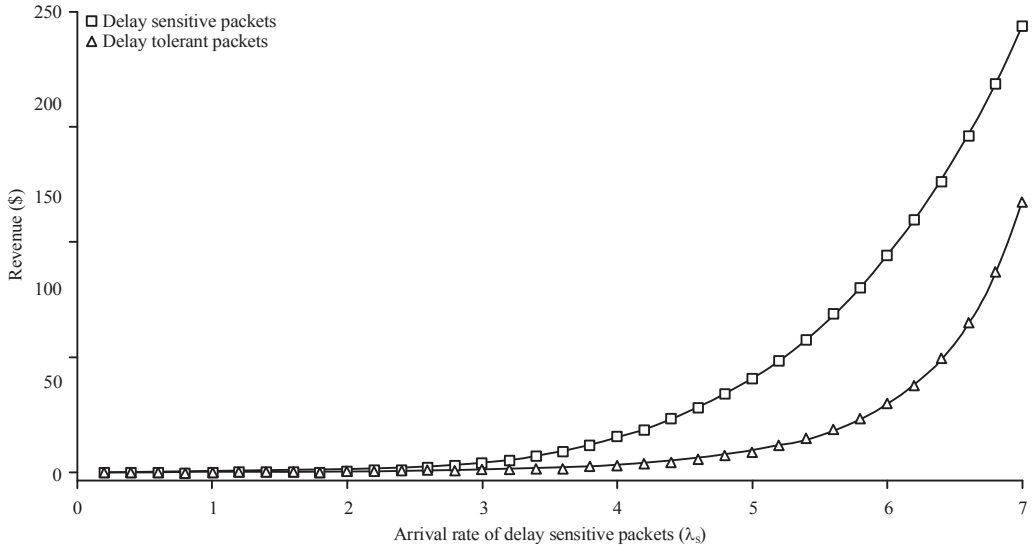

Fig. 7: Variation of revenue with arrival rate of delay sensitive packets in terms of mean slowdown

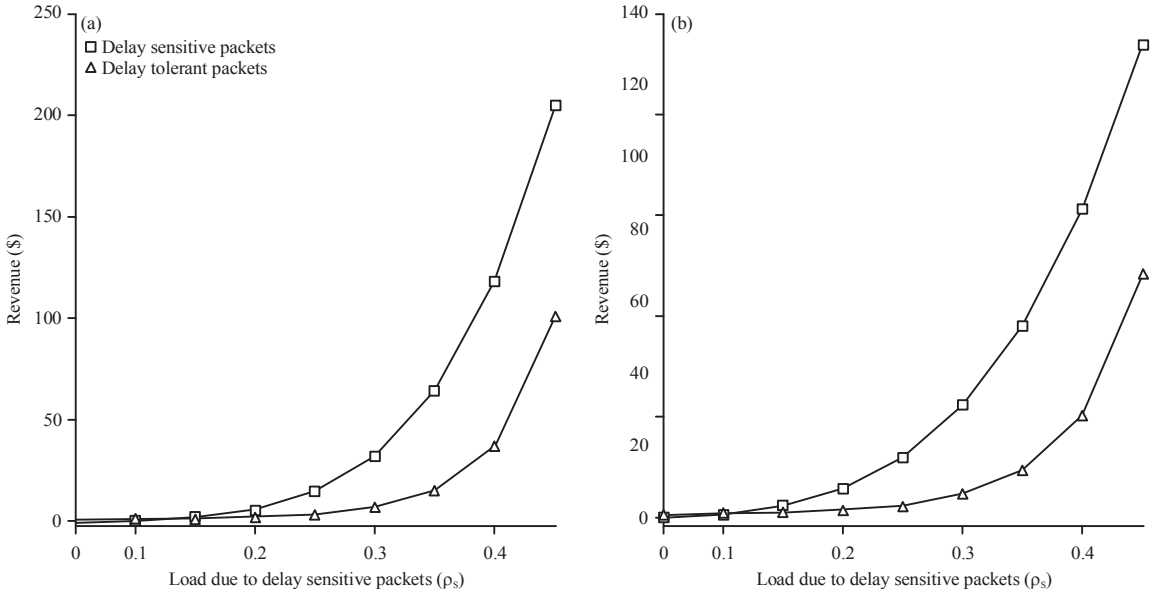

Fig. 8(a-b): Variation of revenue with load of delay sensitive packets with varying number of servers in terms of mean slowdown, (a) Revenue for 5 servers, (b) Revenue for 4 servers

arrival rate for both delay sensitive and delay tolerant packets. It is further observed that for low arrival rate values the revenue generated is almost the same for delay sensitive and delay tolerant packets, however as the arrival rate increases the revenue generated is higher for delay sensitive packets as compared to revenue generated for delay tolerant packets. The effectiveness of deploying service differentiation and price differentiation is observed at higher arrival rate values.

Variation of revenue with load for varying number of servers in terms of mean slowdown: The effect of variation of load on revenue for varying numbers of servers is investigated. In particular, the effect of increasing number of servers on revenue for delay sensitive and delay tolerant packets in terms of mean slowdown is investigated.

Figure 8 showed a graph of revenue as a function of load of delay sensitive packets with varying number of servers in terms of mean slowdown. Equation 3 and 15 are used to plot the graph of revenue as a function of load. It is observed that revenue generally increases with increase in load for both considered number of servers. It is further observed that for low load values the revenue generated is almost the same for both delay sensitive and delay tolerant packets, however as the load increases the revenue generated

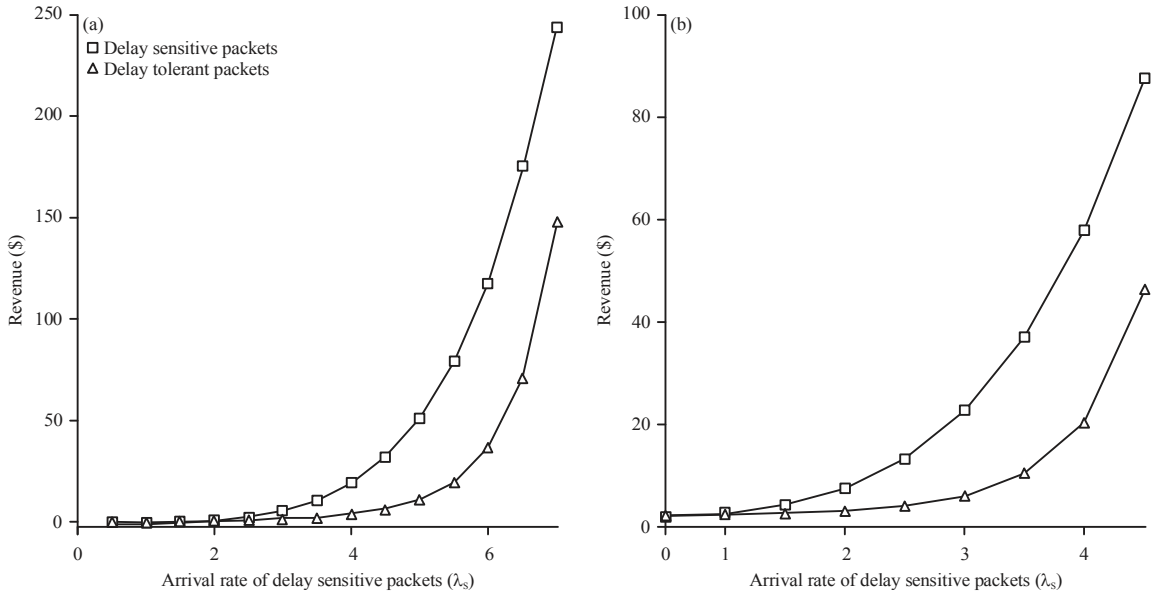

Fig. 9(a-b): Variation of revenue with arrival rate of delay sensitive packets with varying number of servers in terms of mean slowdown, (a) Revenue for 5 servers, (b) Revenue for 4 servers

is higher for delay sensitive packets as compared to delay tolerant packets. The revenue generated for higher number of servers is more than the revenue generated for lower number of servers as shown in Fig. 8a, b. This observation implies that higher number of servers generates more revenue at higher load values.

Variation of revenue with arrival rate for varying number of servers in terms of mean slowdown: The effect of variation of arrival rate on revenue for varying numbers of servers is investigated. In particular, the effect of increasing number of servers on revenue for delay sensitive and delay tolerant packets in terms of mean slowdown is investigated.

Figure 9 showed a graph of revenue as a function of arrival rate for delay sensitive and delay tolerant packets in terms of mean slowdown. Equation 3 and 15 are used to plot the graph of revenue as a function of arrival rate. It is observed that revenue generally increases with increase in arrival rate for both delay sensitive and delay tolerant packets. It is further observed that for low arrival rate values the revenue generated is almost the same for delay sensitive and delay tolerant packets, however as the arrival rate increases the revenue generated is higher for delay sensitive packets as compared to revenue generated from delay tolerant packets. It is also observed that higher number of servers generate more revenue than lower number of servers at higher arrival rate values as shown in Fig. 9a, b. This observation implies that deploying higher number of servers is more effective at higher arrival rate values.

DISCUSSIONS

Previous studies on pricing mechanisms assumed the same type of users and this constrains the performance and utilization of computing systems since requests are processed upon arrival even if they are delay-tolerant[9-11]. In practice, both latency-critical and delay-tolerant jobs coexist in the cloud[10]. Pricing is a critical factor for organizations offering cloud computing services because prices determine customer's behaviour and loyalty. The current pricing scheme proposed by Nansamba *et al.*[19] restricted the choice of cloud users who may want to pay less but are comfortable with incurring extra delay for their packets.

This study proposed a charging model in which cloud users 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 and those cloud users who are willing to tolerate delay are charged lower prices at the expense of delaying the completion time of packets.

Numerical results showed that the derived models can provide price differentiation with delay tolerant packets resulting in paying less while the delay sensitive packets result in paying more.

It is further observed that increase in load and arrival rate lead to increase in revenue. When load or arrival rate increases, the number of packets also increase leading to more packets being processed and hence increase in revenue. However, for low load and low arrival rate values price differentiation has little effect on revenue. Additionally when more servers are used, more revenue is generated as a result of increase in service rate.

The results further showed that the proposed pricing scheme provides differentiated pricing for real time and non real time which outperforms the pricing scheme proposed by Nansamba *et al.*[19], it also outperforms the distributed pricing scheme proposed by Wang *et al.*[23], the service provider oriented scheme proposed by Cao *et al.*[9] and customer oriented pricing scheme proposed by Feng *et al.*[11] that all provide the same service for all considered types of traffic. Providing differentiated pricing ensures that different types of traffic are given different pricing levels. This study considered two classes of traffic, that is, real time and non-real time. In the future the study can be extended to scenarios with more than two classes of traffic.

CONCLUSION

This study derived models of the differentiated pricing scheme in which prices are charged based on the time sensitivity of the packets. In this model cloud users who are not willing to tolerate any delay in the completion time of their requests are charged using a standard pricing model and those cloud users who are willing to tolerate delay are charged lower prices at the expense of delaying packet completion time. The performance of the proposed revenue models was analyzed against the single class models. The numerical results obtained from the derived models show that delay sensitive packets generate more revenue while the delay tolerant packets generate less revenue. The differences in the price is more pronounced for higher load and arrival rate values.

SIGNIFICANCE STATEMENTS

This study discovers the possible ways of modeling a scheme that provides differentiated pricing for heterogeneous cloud computing multiservers. It is expected that this study will help researchers to uncover possible ways of charging prices in cloud computing such that cloud users who are not willing to tolerate any delay in the completion time of their requests are charged differently from cloud users who are willing to tolerate delay but are charged lower prices.

REFERENCES

1. Garimella, S., N. Garg and Vikasdeep, 2012. Features, benefits, futuristic projections of cloud and intercloud extensions to the NET. Int. J. Innov. Eng. Technol., 1: 23-30.
2. Foster, I., Y. Zhao, I. Raicu and S. Lu, 2008. Cloud computing and grid computing 360-degree compared. Proceedings of the Grid Computing Environments Workshop, November 12-16, 2008, Austin, TX., USA., pp: 1-10.
3. Acharjya, D.P., S. Dehuri and S. Sanyal, 2015. Computational Intelligence for Big Data Analysis: Frontier Advances and Applications. Springer International Publishing, Switzerland, ISBN-13: 9783319165981, Pages: 267.
4. Uma, V. and V.J. Suseela, 2014. Current Practices in Academic Librarianship. Allied Publishers Pvt. Ltd., India, ISBN-13: 9788184249422, Pages: 268.
5. Dutta, S., M.J. Zbaracki and M. Bergen, 2003. Pricing process as a capability: A resource-based perspective. Strategic Manage. J., 24: 615-630.
6. Weinhardt, C., A. Anandasivam, B. Blau, N. Borissov, T. Meinl, W. Michalk and J. Stoβer, 2009. Cloud computing-a classification, business models and research directions. Bus. Inform. Syst. Eng., 1: 391-399.
7. Sahal, R., M.H. Khafagy and F.A. Omara, 2016. A survey on SLA management for cloud computing and cloud-hosted big data analytic applications. Int. J. Database Theory Applic., 9: 107-118.
8. Firdhous, M., S. Hassan and O. Ghazali, 2013. Monitoring, tracking and quantification of quality of service in cloud computing. Int. J. Scient. Eng. Res., 4: 112-117.
9. Cao, J., K. Hwang, K. Li and A.Y. Zomaya, 2013. Optimal multiserver configuration for profit maximization in cloud computing. IEEE Trans. Parallel Distrib. Syst., 24: 1087-1096.
10. Zhang, L. and D. Ardagna, 2004. SLA based profit optimization in autonomic computing systems. Proceedings of the 2nd International Conference on Service Oriented Computing, November 15-19, 2004, New York, USA., pp: 173-182.
11. Feng, G., S. Garg, R. Buyya and W. Li, 2012. Revenue maximization using adaptive resource provisioning in cloud computing environments. Proceedings of the 13th ACM/IEEE International Conference on Grid Computing, Volume 13, September 20-23, 2012, IEEE Computer Society Washington, DC, USA., pp: 192-200.
12. Narman, H.S., M.S. Hossain and M. Atiquzzaman, 2014. h-DDSS: Heterogeneous dynamic dedicated servers scheduling in cloud computing. Proceedings of the IEEE International Conference on Communications, June 10-14, 2014, Sydney, NSW., Australia, pp: 3475-3480.
13. Crago, S., K. Dunn, P. Eads, L. Hochstein and D. Kang *et al.*, 2011. Heterogeneous cloud computing. Proceedings of the IEEE International Conference on Cluster Computing, September 26-30, 2011, Austin, TX., USA., pp: 378-385.

14. Wu, X. and F.D. Pellegrin, 2018. On the benefits of QoS-differentiated posted pricing in cloud computing: An analytical model. J. Latex, Vol. 10.
15. Suri, P.K. and M. Sumit, 2012. A comparative study of various computing processing environments: A review. Int. J. Comput. Sci. Inform. Technol., 3: 5215-5218.
16. Al-Roomi, M., S. Al-Ebrahim, S. Buqrais and I. Ahmad, 2013. Cloud computing pricing models: A survey. Int. J. Grid Distrib. Comput., 6: 93-106.
17. Yeo, C.S., S. Venugopal, X. Chu and R. Buyya, 2010. Autonomic metered pricing for a utility computing service. Future Generat. Comput. Syst., 26: 1368-1380.
18. Mihailescu, M. and Y.M. Teo, 2010. Dynamic resource pricing on federated clouds. Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, May 17-20, 2010, IEEE Computer Society, pp: 513-517.
19. Nansamba, B., K.S. Kaawaase, M. Okopa and B.K. Asingwire, 2017. Pricing scheme for heterogeneous multiserver cloud computing system. Aust. J. Comput. Sci., 4: 32-43.
20. Lo, D., L. Cheng, R. Govindaraju, P. Ranganathan and C. Kozyrakis, 2016. Improving resource efficiency at scale with heracles. ACM Trans. Comput. Syst., Vol. 34. 10.1145/2882783.
21. Zheng, L., C. Joe-Wong, C.G. Brinton, C.W. Tan, S. Ha and M. Chiang, 2016. On the viability of a cloud virtual service provider. Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, June, 14-18, 2016, Juan-Les-Pins, France, pp: 235-248.
22. Daigle, J.N., 2005. The Basic M/G/1 Queueing System. In: Queueing Theory with Applications to Packet Telecommunication, Daigle, J. (Ed.)., Springer, US., ISBN: 978-0-387-22859-4, pp: 159-223.
23. Wang, H., Q. Jing, R. Chen, B. He, Z. Qian and L. Zhou, 2010. Distributed systems meet economics: Pricing in the cloud. Microsoft. https://www.usenix.org/legacy/event/hotcloud10/tech/full_papers/WangH.pdf