Vol 8. No. 2 Issue 2 – August, 2015

African Journal of Computing & ICT

www.ajocict.net

Performance Modeling of Cloud E-Marketplaces Using Non Pre-Emptive Prioritized Multi Server Multi Stage Model

A.O Akingbesote
Department of Computer Science
Adekunle Ajasin University,
P. M. B 01,
Akungba-Akoko, Nigeria
alaba.akingbesote.@aaua.edu.ng

ABSTRACT

E-marketplaces have witnessed many evolutions, among which are: traditional, Internet, web services, Grid and Cloud E-marketplaces. The cloud E-marketplaces has revolutionized the E-markets by providing services based on pay per go. One major challenge in the marketplace is the cloud performance. For example, the performance challenge rose in 2008 from 63.1% to 82.9% in 2009. This is an increase of 19.8% as against the Security challenge of about 12.9%. Most existing literature that delved into performance impact on Cloud E-marketplaces focuses on the exogenous Non Priority model with emphasis on First Come First Serve and preemptive disciplines. With the increase in consumer demanding for different services, the First Come First Serve may not be suitable. Also, literature reveals that in practice, pre-emption and migration of virtual machines are costly. Second, pre-emption leads to increase in response time of consumers' requests especially when the requests are deadline constrained. This research proposes a Non Pre-emptive prioritized Multi Server Multi Stage Model. This model prioritizes services using multiple servers at each stage. Experiment is conducted and a comparative study of this model is done with the Non Pre-emptive multi stage model. Our results reveal a better performance in consumers' waiting time and on the issue server utilisation, the result proves better when the arrival rate increases.

Keywords: E-marketplace; Multi-stage mechanism; waiting time, Server;

African Journal of Computing & ICT Reference Format:
A.O Akingbesote (2015): Performance Modeling of Cloud E-Marketplaces Using Non Pre-Emptive Prioritized Multi Server Multi Stage Model Afr J. of Comp & ICTs. Vol 8, No. 2, Issue 2. Pp 15-24.

1. INTRODUCTION

The concept of E-marketplaces has been on for long with so many benefits see [3] [4] [5]. This concept started with the traditional E-marketplace. This is a Web portal where buyers and suppliers come together to explore new business opportunities [1]. This market allows the buying and selling of goods with both the buyers and the consumers having direct link. This market uses digital means to brand their products or logo. This is like people finding or getting a particular business through a referral or a network and eventually build a rapport with them [2]. While this market has some advantages, so too are major setbacks, among which is the lack of competiveness and the in ability to creating an air of excitement [3]. The evolution of the internet technology gives room to organizations to open their shops on the internet and also allows millions of consumers to participate in the global online marketplaces.

The Second evolution of the E-marketplace is the Web service E- marketplaces. This evolution is an update to object-oriented computing. This came as result of the emergence of Service Oriented Architecture (SOA), which is the paradigm of organizational models of systems, aimed at solving large business operations using existing services.

The web services E-marketplace is a market that belong to a community that allows products' producer to produce their product(s), advertise them on the web for the consumer to consume them. In other words, it is a local community of service providers and service consumers organized in vertical markets and gathering around portals [6]. While the use of the Web Service marketplace has been successful, especially with the business class, that of using this market for high computational power is a challenge coupled with others like the high costs of maintaining the equipment and human resources [7] [8]. This brought the concept of Grid E-market Technology. This is a market where computational power is purchased by consumers (Consumers/Applications) through the use of middleware or resource allocation broker.

Four major things distinguish the Grid marketplace from others as written in [9][10]. While these markets have been viable in the context of high performance, Exploitations of underutilized Resources, Resource balancing and wide- scale distributed computing[11] [12], this market has some challenges as stated in [13][14] [15] [16].

These challenges reveal the vision of scholars as far back as 1969 as written in [17] that:

> "*As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities' which, like present electric and telephone utilities, will service individual homes and offices across the country*".

This vision evolves cloud E-market. This is a market paradigm that allows consumers to shift from building of computer software, Infrastructure and platform to procurement. The emergence of cloud market allows many consumers and providers to participate by allowing the traditional/web and grid service providers to rebrand their services as cloud hosting. However, literature for example [18], reveals that issue of performance, security have been the top most challenges [18].

On the issue of performance, for example, in the International Data Corporation (IDC) report, the performance challenge rose in 2008 from 63.1% to 82.9% in 2009 as reported in [19] [20][21]. This is an increase of 19.8% as against the Security challenge of about 12.9% which is higher than that of Security [22]. In addition, As the markets grow, most providers are implementing various service offerings to their consumers; For example, Amazon Elastic Compute Cloud (EC2) offers three different offerings, the Reserved, Spot and the On-Demand [23].

Most existing literature that delved into performance impact on Cloud E-marketplaces focuses on the exogenous Non Priority model with emphasis on First Come First Serve discipline [24] [25][26]. As the server farms increases with consumers demanding different service disciplines, some scholars like [27][28][29],[30] use the Preemptive service discipline in the area of networking while other scholars (see [31]) use the Preemptive service discipline and migration in the context of Cloud E-marketplaces but literature reveals that in practice, pre-emption and migration of virtual machines are costly [32]. Second, pre-emption leads to increase in response time of consumers' requests especially when the requests are deadline constrained [33]. This work extends existing and widely adopted theories to the exogenous Non Pre-emptive Multi Server Multi Stage Model.

The remainder of this paper is organized as follows. Section II discusses the related work. Section III introduces our mathematical model description with the numerical and simulation set up. In Section IV, we have our results and discussion. The paper ended with the conclusion in Section V.

2. LITERATURE REVIEW

Much work has been done under cloud E-marketplaces. For examples, energy, privacy and security [34] [35][36]. However, little has been done in the area of optimization for resource management with regard to cloud performance[26]. The work in [37] model the cloud as series of queues with each service station model as M/M/1 for optimal resource allocation. In the work of these authors, the theoretical concept is based on three concatenated queues and the theoretical analysis is based on the relationship between the service response time and the allocated resources in each queuing system. In [38], the authors propose M/G/c to evaluate a cloud server firm with the assumption that the numbers of server machines are not restricted. The result of the work demonstrates the manner in which request response time and number of task in the system may be assessed with sufficient accuracy.

In [39], the authors use discrete time preemptive priority to analyze two classes; the authors consider two classes of customers which have to be served under high and low priority. This work is based on theoretical concept to show the influence of the priority discipline and service time distribution on the performance measure through numerical examples. The work of [40] considers the waiting time queue in the accumulating priority queue.

The use of pre-emptive policy in cloud E-marketplaces is proposed in [31]. The is based on the argument that when an urgent request arrives, it preempts the current request in service and such preempted request is then migrated to another virtual machine if it cannot meet the deadline for completion. In [31] the authors propose a Multi-dimensional Resource Integrated Scheduling (MRIS) which is an inquisitive algorithm to obtain the approximate optimal solution. This work removes the scheduling bottleneck from one dimensional to multi-dimensional resources. The work in [26] proposes an M/M/m queuing model to develop a synthetic optimization method to optimize the performance of services in an on Demand service. The simulation results show that the proposed method can allow less wait time, queue length and more customers to gain the service using synthetic optimization function when the numbers of servers increases.

In [41], the authors model the cloud using M/M/c/c model with different priority classes with the main goal of studying the rejection probability for different priority classes. The works of these preceding authors presented the researcher the opportunity to make the contribution in this paper. For example, the argument for using the queuing model is based on that of [37] [26] [41], also the work of [31] [41] are the fore-runners of the idea of viewing the cloud as networks of queue.

However, the opportunity to contribute derives from the solid foundations already laid in some of these works. For example, while the works of [37] [26] achieve better results toward accurate waiting time performance, these could only be used where the service provisioning discipline is the same. In addition, the work of [26] uses pre-emptive approach but these are costly and may lead to increase in consumers waiting time due to the significant amount spent when pre-emption occurs. Therefore this could not be used in in the context where the requests are deadline constrained. Another critical issue which is never considered by all these authors is when the incoming requests on the scheduler increases to a level where the queue length has reached a stage that will breach the Service Level agreement. To resolve this, this work proposes a solution by creating a Control able Prioritized Non Pre-emptive Multi Server Multi Stage Mechanism (CPNPM) that switches to a reservoir server when the main scheduler reaches it upper limit. This mechanism is used at the dispatcher-In and dispatcher-Out level. This mechanism differentiates this work from previous works. This, to the best of our knowledge has never appeared in the literature**.**

3. PROPOSED MODEL

The proposed model is shown in Figure 1. This model consists of three sub models which are sub model 1, 2 and 3 respectively. Sub model 1 is the arrival stage and it consists of the incoming web or consumer applications with two dispatching points (Dispatcher-In A and B). Dispatcher-In A serves as the main entry of all prioritised work. When the queue length of consumers in Dispatcher-In A reaches certain threshold say N, consumer applications switch to Dispatcher-In B. Sub model 2 is the second stage that consists of n service stations that are networked together. The processing of the applications takes place at these service stations. The third stage is the dispatcher-Out. This has two stages like the dispatcher-In. these are the Dispatcher-Out A and B respectively.

The idea is that when the incoming consumers C1, C2, and C3 arrive at the cloud market, it move to the dispatcher A under the control of the Controllable Prioritized Non Pre-emptive Model (CPNPM) and as the incoming number of consumers increase to a point say N, then the CPNPM switches to dispatcher-In B to reduce congestions. This process continues until the number of consumers in dispatcher-In A goes to N-1.

Sub-model 2 is made up of the web queue servers that acts as the real processors that provide the service based on prioritized Non pre-emptive policy. This sub model follows the same prioritized principle of all the sub models. The idea of this principle is that when an incoming request meets lower one on the queue, it takes over from that request but when lower request is currently under processing, that request is allowed to finish. When requests of the same priority are on the queue then the order is based on First Come First Serve (FCFS). The word consumer in this paper is referred to as an application requesting service from the provider [42].

The third stage is the Dispatcher-Out, when the number of consumers have been processed, it moves out through the Dispatcher-Out A and as soon as the length of queue reaches point N, the CPNPM moves the outgoing consumers to Dispatcher-Out B until the number of consumers in dispatcher—Out A comes back to N.

The analytical solution of this work is based on the use of queuing theory as the proof of concept. This concept is achieved by studying various literature. See [43] [44] [45][46][47][48][49] [50] [51] and [52]. The result of this search enables us to base the mathematical concept on the work of Moder in [44]. The Moder concept is adopted as the solution approach to solve the (CPNPM) sub-model 1 and 3 problems. On sub-model 2, the mathematical solution on the author's previous work is adopted. See [53][22]under the M/M/c/Pr model.

Modelling the Dispatcher-In and Dispatcher-Out

In [44], the author determines the measure of effectiveness through the following steps (see the full equation in Appendix A):

i. Determine the list of admissible states
ii. Determine the steady state equations
iii. Solve each of the steady state equations
iv. Calculate the measure of effectiveness.

The measure of effectiveness needed in this paper is the waiting time at the dispatcher-In ($W_{q_{din}}$) and dispatcher-Out

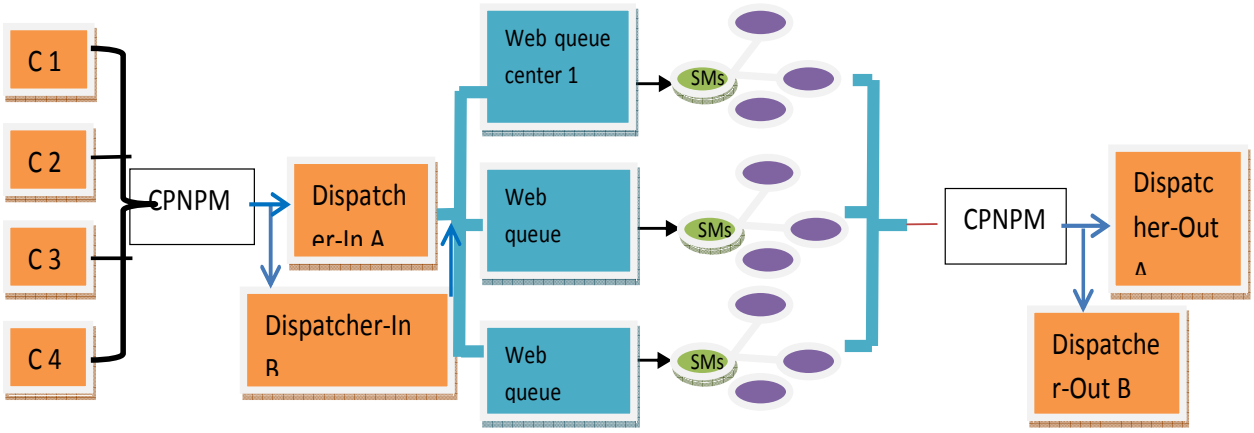

Figure 3: Proposed Prioritized Non Pre-emptive Multi Server Multi Stage Model

(Wq_{dout}) respectively. These are:

$$Wq_{din} = \frac{L_q}{\lambda}$$

and

$$Wq_{dout} = \frac{L_q}{\lambda}$$

but

$$L_q = \frac{P^{\delta}}{1-\gamma^{N-v}}\left[\frac{\rho^{\delta}[\gamma-(v+1)\gamma^{v+1}+v\gamma^{v+2}](1-\gamma^{v})}{\delta!(1-\gamma)^2}\right.$$
$$+\frac{\rho^{\delta}}{\delta!}\left\{\frac{\gamma^{N}(N\gamma-\gamma-N)+\gamma^{v+1}(v+1-v\gamma)}{(1-\gamma)^2}-\frac{1}{2}\gamma^{N-1}[N(N-1)-v(v+1)]\right\}$$
$$+\frac{1}{2}\gamma^{N-1}(1-\gamma)[N(N-1)-v(v-1)]\left\{\sum_{s=\delta}^{S-1}\frac{\rho^{s}}{s!}-\frac{\rho^{\delta}}{\delta!}\right\}$$
$$+\frac{\rho^{S}\gamma^{N-1}(1-\gamma)}{2S!(S-\rho)^3S^{N-2}}\{S^{N-1}(S-\rho)^2[(N-1))((N-2))-v(v-1)]-2S^{v}\rho^{N-v}[\rho(N-2)-S(N-1)]$$
$$\left.-2S^{N-v-1}\rho^{v+v}[Sv-(v-1)\rho]+2(S^{N-v}[\rho^{N}(S(N-1)-\rho(N-2))]]\right]$$

Where

N = queue length

s = number of busy channel

L = mean number of consumers in the system

L_q = mean number of consumers in the queue

M= number of (fixed) channels in the conventional multiple channel process

N = The shift up point, i.e the queue length at which additional channels are instantaneously opened if s < S

$P_{n\,s}$ = The steady state probability that n consumers are in the queue and s consumers are being serviced.

S = Maximum number of manned channels $\delta < S < \infty$

W= Mean wait time in the system

W_q = Mean wait time in the queue

δ = Maximum number of manned channels $\delta \geq 1$

v= The shift down point i.e the minimum queue length when $\delta \geq 1$

λ = The mean arrival rate

μ = The mean service rate per channel

$\rho = \lambda/\mu$ = Utilisation factor

$\gamma = \rho/\delta$

Modelling the Web queue stations

As earlier said, the previous concept [53][22] is adopted by using the M/M/c/Pr approach. The arrival and the service process are exponentially distributed for each priority at each of the c channels within a station. Also, due to large volume of consumers entering the market for request we assume an infinite population. To get our performance measure, we followed the six steps stated in [54] and the law of conservation of flow [55].

www.ajocict.net

ρ_k is defined as

$$\rho_k = \frac{\lambda_k}{c\mu_k} \quad (1 \leq k \leq r) \tag{4}$$

and

$$\sigma_k = \sum_{i=1}^{k} \rho_k \quad (\sigma_0 \equiv \rho = \lambda/c\mu) \tag{5}$$

Where the system is stationary for $\rho < 1$,and

$$W_q^{(i)} = \sum_{k=1}^{i-1} E[S_k^r] + \sum_{k=1}^{i} E[S_0] \tag{6}$$

Where S_k is the time required to serve n_k consumers of the kth priority in the line ahead of the consumer. S_k^r is the service time of the n_k^r consumers of priority k which arrive during $W_q^{(i)}$, S_0 is the amount of time remaining until the next server becomes available.

Therefore

$$E[S_0] = \Pr\left(\begin{matrix}all\ servers\ are\ busy\ within\ a \\ service\ station\end{matrix}\right).E[S_0|all\ server\ are\ busy\ within\ a service\ station$$

$$= \left(\sum_{n=c}^{\infty} P_n\right) \frac{1}{c\mu} =$$

$$P_0 \left(\sum_{n=c}^{\infty} \frac{c\rho^n}{c^{n-c}\, c!}\right) \frac{1}{c\mu} = \frac{P_0 (c\rho)^c}{c!(1-\rho)} \tag{7}$$

$$= \frac{(c\rho)^c}{c!(1-\rho)(c\mu)} \left[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]^{-1} \tag{8}$$

but

$$E[S_0] = \rho \sum_{k=1}^{r} \frac{1}{u_k} \frac{\rho_k}{\rho} = \sum_{k=1}^{r} \frac{\rho_k}{u_k} \tag{9}$$

$$W_q^{(i)} = \frac{E[S_0]}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{10}$$

$$= \frac{\left[c!(1-\rho)(c\mu)\sum_{n=0}^{c-1}(c\rho)\frac{n-c}{n!} + c\mu\right]^{-1}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{11}$$

The expected time taken in a service station is

$$W_{qst} = \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_{qst}^{(i)} \tag{12}$$

The overall average time taken in j service stations is

$$W(ave)_{qst} = \frac{\sum_{n=1}^{j}\left[\sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_q^{(i)}\right]}{j} \tag{13}$$

Our interest in this experiment is the waiting time experienced by consumers in all the sub models. This is given as:

$$Wq_{Tot} = Wq_{din} + W(ave)_{qst} + Wq_{dout} \tag{13}$$

Experimental Set-Up

The first thing that is done is to validate the mathematical solution with the simulation to ascertain the degree of correction. We measure the waiting time of consumers using the wolfram Mathematical 9.0 as the mathematical tool for our validation results and arena 14.5 as the simulator. This is done by setting both the simulation and the analytical parameter to the same values. We set λ (arrival rate) to 10, 20 ……60 and N= 10. The service time $(1/\mu)$ is set to 0.005 for the dispatcher-In/Out and 0.005 for each of the servers in the web Queue stations. The Average waiting time of the three prioritised consumer is recorded. The result is depicted in Figure 2. In this Figure, as the arrival rate increases so also the waiting time increase in both Analytical and simulation. Also, the degree of variation is hardly noticed in both simulation and the analytical approach except at the point where λ is = 40 and 50 respectively. However the coefficient of variation is very small and less than unity. This then implies that the simulation is in line with the formulated mathematical concept.

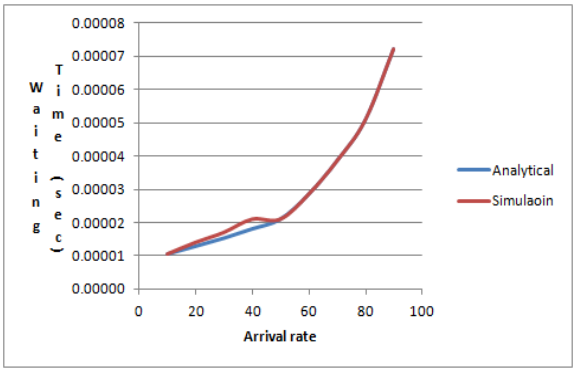

Figure 4: Analytical and Simulation results

www.ajocict.net

On the main experiment, The service rate is set to 0.005 for the dispatcher-In/Out and 0.005 for each of the A and B servers in the web Queue stations. N which is the upper bound is set to 10 and the arrival patter is constant. The experiments started with mean inter arrival time of 0.1 to 0.9. The total number of consumers entering the market for each experiment is set to 60,000. That is, each group of consumers (for example higher priority group is set to 20,000).
This simulation is run with replication length of 1000 in 24 hours per day and the base time in seconds. The experiment is replicated 10 times. The same Arena 14.5 is used as the simulator. In section IV, the results are analyzed and compared with the non multi stage model.

4. RESULTS AND DISCUSION

The results are based on three things: The first is the waiting time experienced by the prioritised consumers (C1…C3), second is the average waiting experienced by this model compared with existing model when the system is not multi staged. The third result is based on server utilisation of the multi stage system.

On the first result shown in Figure 3, it is observed that the consumer with higher priority (C1) experienced short waiting time while that of lower priority had the highest waiting time (C3). This occur between the inter arrival time of say 0.1to 0.6. Above 0.6 to 0,9 they both have the same waiting. What accounted for this equal waiting time because the arrival rate of consumers is getting slow such that the higher priority consumers never met the lower one. Therefore there was no need for priority during that period.

The second result is based on the comparison of this model (Multi Stage) with the existing model that did not experience multi stage (non-multi stage). This is depicted in Figure 4. It is observed that waiting time of consumers under the multi stage model performs better than that of non multi stage between 0.1 and 0.4. That is, when the number of consumers entering the market is high. However, they both have the same waiting time after this time. What accounted for this is that when the number of consumers is high, the multi stage model puts into operation the Distpacher-In B and Dispacther-Out B servers (See Figure 1) thereby reducing the consumers waiting time. As the number of consumers reduces, these two servers stop operating thereby behaving like a normal non multi stage model.

To substantiate this, the server utilisation experienced under the multi stage model is recorded in Table 1. It is observed that the Distpacher-In B stop working after the inter arrival time of 0.6. At this point, the consumers inter arrival is low such that Distpacher-In A is enough to process incoming request.

The same thing occurs to that of Dispacther-Out B after 0.4. What accounted for the difference between 0.6 in Distpacher-In A and Dispacther-Out B is based on the experimental set up where the server processing rate of the two are different.

The last experiment conducted is based on the comparison of the server utilisation of the proposed model (Multi Stage) and the non multi stage model. The experimental results are depicted in Table 2. One remarkable observation that makes the multi stage better than the non multi stage is when the inter arrival time is 0.1 (higher arrival rate). It is observed that the server utilisation experienced under the multi stage is .0000722767 while under the non multi stage the message "server error" indicated by "SE" in Figure 2 is recorded. This error is due to the high number of consumers on the queue therefore making the inter arrival rate higher that the service rate leading to error. Unlike the multi stage which allows the second server to work when such operation occurs. Between the inter arrival time of 0.2 to 07, the server utilisation of the non multi stage seems higher than that of multi stage because during these periods, only one server is working under the non multi server while two are working under the multi stage model. When the inter arrival time goes above 0.6, then the two model have the same server utilisation results. What accounted for this is because of the low number of consumers in the market as earlier discussed. At these points, the same number of servers is working.

In summary, the multi stage model had better performance in consumers' waiting time because of the Controllable Prioritized Non Pre-emptive Mechanism (CPNPM). On the issue server utilisation, the result proves better when the arrival rate increases.

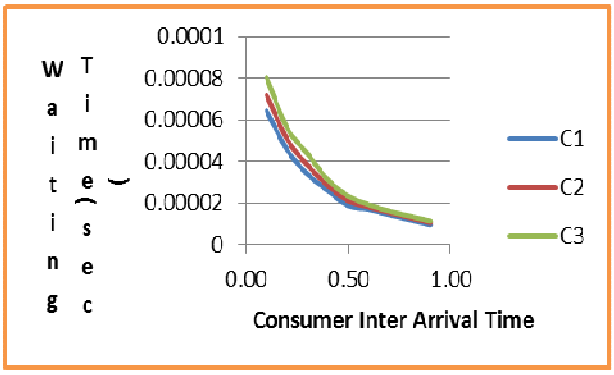

Figure 3: Waiting Time of the three prioritized Multi Server Multi Stage Model

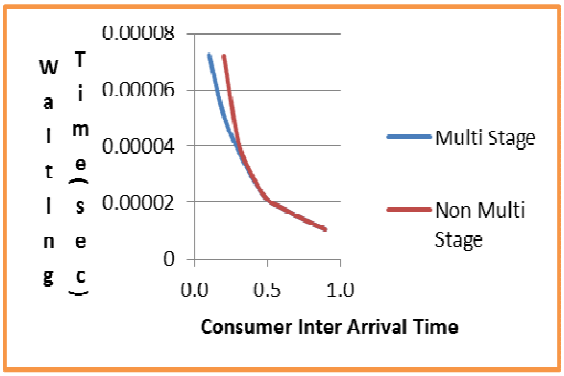

Figure 4: Waiting Time under Multi Stage and Non Multi Stage Model

Table 1: Server Utilization under the Prioritized Multi Stage Model

Instantaneous Utilization	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8	Experiment 9
Inter Arrival Time	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Dispatcher In-A	0.4223	0.2686	0.1929	0.1462	0.1172	0.0978	0.08387919	0.07336304	0.06521159
Dispatcher In-B	0.1701	0.0254377	0.00266676	0.00031252	0.00017501	0.00002083	0.00001786	0	0
Web queue center 1	0.04861852	0.02406283	0.01587552	0.01190677	0.00952548	0.0080629	0.0068932	0.00604718	0.00537527
Web queue center 2	0.05361791	0.0280004	0.01900063	0.01437563	0.01145057	0.00925046	0.00792897	0.00693785	0.00616698
Web queue center 3	0.05411836	0.0233753	0.01575055	0.01187553	0.00957548	0.0080004	0.00687534	0.00601593	0.00534749
Dispatcher Out-A	0.1319	0.07300096	0.04991834	0.03715787	0.02977649	0.02514709	0.02155465	0.01887594	0.01676473
Dispatcher Out-B	0.01224869	0.00181253	0.00008334	0.00003125	0	0	0	0	0
Ave. Resource Uti.	0.15	0.07375	0.04935	0.01513	0.02962	0.02471	0.02119	0.01854	0.01648

Table 2: Waiting Time under Multi Stage and Non Multi Stage Models

Waitng Time	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8	Experiment 9
Int.Arrival Time	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Multi Stage	0.0000722767	0.0000508667	0.0000385767	0.0000285133	0.0000209933	0.0000180567	0.0000152733	0.0000128467	0.0000104633
Non Multi Stage	"SE"	0.0000719633	0.0000418200	0.0000291400	0.0000210600	0.0000180600	0.0000152733	0.0000128467	0.0000104633

5. CONCLUSION

Cloud computing is a computing paradigm that allows consumers to shift from building of computer software, Infrastructure and platform to procurement. This has allowed cloud e-market providers to rebrand their services as cloud hosting.

One major challenge in the market is that of cloud performance especially in context of consumers' waiting time. This is because; keeping consumers waiting for long time will lead to consumers' dissatisfaction and loss of business. This paper tackles this problem by proposing a Non Pre-emptive prioritized Multi Server Multi Stage Model.

While the multi-server model prioritizes services at the second stage, the first and the third stage uses the Controllable Prioritized Non Pre-emptive mechanism (CPNPM) to monitor and control the server. When the incoming number of consumers increase to a point say N, then, the CPNPM switches to dispatcher-In/Out B. This process continues until the number of consumers in dispatcher-In/Out A is back to N-1. Experiment is conducted and a comparism of this model with the Non Multi Stage is carried out using the waiting time and server utilisation as the metrics of measurement. The results reveal a better waiting time under the Multi Server Multi Stage Model than the Non Multi Stage Model. Also, the Multi Server Multi Stage Model performs better under the resource utilization when the inter arrival time of consumers is low (high arrival rate) unlike the Non Multi Stage Model that recorded server error operation.

Acknowledgment

This work is based on the research supported in part by the National Research Foundation of South Africa-Grant UID: TP11062500001 (2012-2014). The authors also acknowledge funds received from industry partners: Telkom SA Ltd, Huawei Technologies SA (Pty) Ltd and Dynatech Information Systems, South Africa in support of this research.

REFERENCES

[1] D. R. Ferreira and J. Pinto Ferreira, "Building an e-marketplace on a peer-to-peer infrastructure," *Int. J. Comput. Integr. Manuf.*, vol. 17, no. 3, pp. 254–264, 2004.

[2] "Traditional Marketing versus Digital Marketing - Digital Marketing Strategies," *http://digital-marketing-strategy.weebly.com/digital-marketing.html*. [Online]. Available: http://digital-marketing-strategy.weebly.com/digital-marketing.html. [Accessed: 17-Apr-2015].

[3] D. N. Dholakia, R. Dholakia, D. Zwick, and M. Laub, "Electronic commerce and the transformation of marketing," *Internet-MarketingPerspektiven und Erfahrungen aus Deutschl. und den USA*, pp. 55–77, 1999.

[4] T. W. Malone, J. Yates, and R. . Benjamin, "The Logic of Electronic Markets - HBR," *Harvard Business Review*, 1989. [Online]. Available: https://hbr.org/1989/05/the-logic-of-electronic-markets. [Accessed: 24-Apr-2015].

[5] A. Pucihar and J. Gričar, "Environmental Factors Defining eMarketplace Adoption : Case of Large Organizations in Slovenia," pp. 1–13, 2005.

[6] P. M. Papazoglou, *Title Principle and Technology*. Edinburgh Gate, England: Pearson Education limited, 2008.

[7] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities," *Proc. - 10th IEEE Int. Conf. High Perform. Comput. Commun. HPCC 2008*, pp. 5–13, 2008.

[8] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," *2009 Int. Conf. High Perform. Comput. Simul.*, pp. 1–11, Jun. 2009.

[9] R. Wolski, J. S. Plank, and J. Brevik, "g-Commerce – Building Computational Marketplaces for the Computational Grid," 2000.

[10] L. . Kacsukne and T. Kiss, *Distributed and Parallel Systems*, vol. 777. Boston: Kluwer Academic Publishers, 2005.

[11] I. Foster and N. T. Karonis, "A Grid-Enabled MPI: Message Passing in Heterogeneous Distributed Computing Systems," *Proc. IEEE/ACM SC98 Conf.*, pp. 1–11, 1998.

[12] S. Pardeshi, C. Patil, and S. Dhumale, "Grid Computing Architecture and Benefits," *Ijsrp.Org*, vol. 3, no. 8, pp. 3–6, 2013.

[13] H. Casanova and D. Jack, "A Network Server for Solving Computational Science Problems," in *Proceedings of the 1996 ACM/IEEE Conference on Supercomputing (SC')*, 1996, pp. 1–14.

[14] A. S. Grimshaw, W. a Wulf, J. C. French, A. C. Weaver, P. F. R. Jr, and P. F. Reynolds, "Legion : The Next Logical Step Toward a Nationwide Virtual Computer e pluribus unum -- one out of many Technical Report CS-94-21, University of Virginia," 1994.

[15] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: The Condor experience," *Concurr. Comput. Pract. Exp.*, vol. 17, no. 2–4, pp. 323–356, 2005.

[16] K. Nadiminti and R. Buyya, "Enterprise grid computing: State-of-the-art TGrid Computing and Distributed Systems Laboratory, The University of Melbourne," 2005.

[17] L. A. Kleinrock, "vision for the internet.," *ST J. Res.*, vol. 2, no. 1, pp. 2–5, 2005.

[18] N. M. Ani Brow and K. Jayapriya, "An Extensive Survey on QoS in Cloud computing," *International Journal of Computer Science and Information Technologies, Vol. 5 (1) 2014, 1-5*, 2014. [Online]. Available: http://www.ijcsit.com/docs/Volume 5/vol5issue01/ijcsit2014050101.pdf. [Accessed: 16-Apr-2015].

[19] S. O. Kuyoro, "Cloud Computing Security Issues and Challenges," no. 3, pp. 247–255, 2011.

[20] K. Popovic and Z. Hocenski, "Cloud computing security issues and challenges," in *MIPRO, 2010 Proceedings of the 33rd International Convention*, 2010, pp. 344–349.

[21] "AWS | Amazon Elastic Compute Cloud (EC2) - Scalable Cloud Hosting." [Online]. Available: http://aws.amazon.com/ec2/. [Accessed: 23-Apr-2015]." .

[22] A. . Akingbesote, M. . Adigun, S. . Xulu, M. Sanjay, and I. . Ajayi, "Performance Analysis of Non-Preemptive Priority with Application to Cloud E-marketplaces," in *IEEE International Conference on Adative Technolgy (ICAST)*, 2014, pp. 1–6.

[23] "AWS | Amazon Elastic Compute Cloud (EC2) - Scalable Cloud Hosting." [Online]. Available: http://aws.amazon.com/ec2/. [Accessed: 23-Apr-2015].

[24] K. Xiong and H. Perros, "Service Performance and Analysis in Cloud Computing," *2009 Congr. Serv. - I*, pp. 693–700, Jul. 2009.

[25] H. Khazaei, J. Misic, and V. B. Misic, "Modelling of Cloud Computing Centers Using M/G/m Queues," *31st Int. Conf. Distrib. Comput. Syst. Work.*, pp. 87–92, Jun. 2011.

[26] L. Guo, T. Yan, S. Zhao, and C. Jiang, "Dynamic Performance Optimization for Cloud Computing Using M/M/m Queueing System," *J. Appl. Math.*, vol. 2014, pp. 1–8, 2014.

[27] J. Walraevens, B. Steyaert, and H. Bruneel, "A Packet Switch with a Priority Scheduling Discipline: Performance Analysis," *Telecommun. Syst.*, vol. 28, no. 1, pp. 53–77, Jan. 2005.

[28] J. Walraevens, B. Steyaert, M. Moeneclaey, and H. Bruneel, "Delay Analysis of a HOL Priority Queue," *Telecommun. Syst.*, vol. 30, no. 1–3, pp. 81–98, Nov. 2005.

[29] J. Walraevens, B. Steyaert, and H. Bruneel, "Performance analysis of priority queueing systems in discrete time," *Netw. Perform. Eng. , Vol. 5233, No. 1, pp.203–232.*, 2011.

[30] Q. Gong and R. Batta, "A Queue-Length Cutoff Model for a Preemptive Two-Priority M/M/1 System," *SIAM J. Appl. Math.*, vol. 67, no. 1, pp. 99–115, Jan. 2006.

[31] R. Santhosh and T. Ravichandran, "Pre-emptive scheduling of on-line real time services with task migration for cloud computing," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013, pp. 271–276.

[32] S. T. Maguluri, R. Srikant, and L. Ying, "Stochastic models of load balancing and scheduling in cloud computing clusters," in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 702–710.

[33] M. A. Salehi, B. Javadi, and R. Buyya, "Preemption-aware Admission Control in a Virtualized Grid Federation," in *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*, 2012, pp. 854–861.

[34] F. A. Alvi, B. S. Choudary, and N. Jaferry, "A review on cloud computing security issues & challenges," *I iaesjournal.com*, vol. 1 (2), 2012.

[35] J. Wang, "eRAID: A Queueing Model Based Energy Saving Policy," *14th IEEE Int. Symp. Model. Anal. Simul.*, no. 1, pp. 77–86, 2006.

[36] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking," in *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security - ASIA CCS '13*, 2013, pp. 71–82.

[37] X. Nan, Y. He, and L. Guan, "Optimal resource allocation for multimedia cloud based on queuing model," *2011 IEEE 13th Int. Work. Multimed. Signal Process.*, pp. 1–6, Oct. 2011.

[38] K. Popović and Z. Hocenski, "Cloud computing security issues and challenges," in *proceedings of the 33rd international convention,* 2010, pp. 344–349.

[39] J. Walraevens, B. Steyaert, and H. Bruneel, "Analysis of a discrete-time preemptive resume priority buffer," *Eur. J. Oper. Res.*, vol. 186, no. 1, pp. 182–201, Apr. 2008.

[40] D. a. Stanford, P. Taylor, and I. Ziedins, "Waiting time distributions in the accumulating priority queue," *Queueing Syst.*, vol. 77, no. 3, pp. 297–330, Dec. 2013.

[41] W. Ellens, M. Živković, J. Akkerboom, R. Litjens, and H. Van Den Berg, "Performance of cloud computing centers with multiple priority classes," in *Proceedings - 2012 IEEE 5th International Conference on Cloud Computing, CLOUD 2012*, 2012, pp. 245–252.

[42] H. Goudarzi and M. Pedram, "Maximizing Profit in Cloud Computing System via Resource Allocation," *2011 31st Int. Conf. Distrib. Comput. Syst. Work.*, pp. 1–6, Jun. 2011.

[43] J. Romaní, "Un modelo de la teoria de colas con numero variable de canales," *Trab. Estad.*, vol. 8, no. 3, pp. 175–189, Oct. 1957.

[44] J. J. Moder, "Queuing with fixed and variable channels," *Oper. Res.*, vol. 10, no. 2, pp. 218–232, 1962.

[45] M. Yadin and P. Naor, "Queueing Systems with a Removable Service Station†," *J. Oper. Res. Soc.*, vol. 14, no. 4, pp. 393–405, 1963.

[46] E. Khmelnitsky and Y. Gerchak, "Optimal control approach to production systems with inventory-level-dependent demand," … *Control. IEEE Trans.*, pp. 1–12, 2002.

[47] H. Li and T. Yang, "Queues with a variable number of servers," *Eur. J. Oper. Res.*, vol. 124, no. 3, pp. 615–628, 2000.

[48] A. I. Pazgal and S. Radas, "Comparison of customer balking and reneging behavior to queueing theory predictions: An experimental study," *Comput. Oper. Res.*, vol. 35, no. 8, pp. 2537–2548, 2008.

[49] M. M. Systems, "Queuing system with variable server number," no. 4, pp. 63–65, 2007.

[50] S. Stidham and R. R. Weber, "Monotonic and Insensitive Optimal Policies for Control of Queues with Undiscounted Costs," *Oper. Res.*, vol. 37, no. 4, pp. 611–625, 1989.

[51] M. Yamashiro, "A system where the number of servers changes depending on the queue length," *Microelectron. Reliab.*, vol. 36, no. 3, pp. 389–391, 1996.

[52] A. i A. N. Dudin, "Optimal assignment of the rate for service of customers in a multilinear two-rate service system," *Telemekh., no. 11*, 1981. [Online]. Available: http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=at&paperid=6044&option_lang=eng. [Accessed: 10-Jun-2015].

[53] A. O. Akingbesote, M. O. Adigun, S. Xulu, and E. Jembere, "Performance Modeling of Proposed GUISET Middleware for Mobile Healthcare Services in E-Marketplaces," *J. Appl. Math.*, vol. 2014, p. 9, 2014.

[54] S. T. Maguluri, R. Srikant, and L. Ying, "Stochastic models of load balancing and scheduling in cloud computing clusters," *2012 Proc. IEEE INFOCOM*, pp. 702–710, Mar. 2012.

[55] L. Kleinrock, *Queueing Systems*. John Wiley & Sons, 1975.