MODELLING THE PERFORMANCE OF WEB SERVICES IN CLOUD E-MARKETPLACES BASED ON CONSUMER WAITING TIME AND PROVIDER COST

AKINGBESOTE ALABA OLU

M.Sc. (FUTA), B.Sc. (OOU)

A thesis submitted to the Faculty of Science and Agriculture in fulfillment of the requirement for the award of Degree of Doctor of philosophy (PhD) in Computer Science at the University of Zululand

PROFESSOR M.O ADIGUN

PROMOTER

NOVEMBER 2015

DECLARATION

This dissertation represents the author's original work, conducted at the University of Zululand. It is submitted for the award of the degree of Doctor of Philosophy (PhD) in Computer Science in the Faculty of Science and Agriculture, University of Zululand.. No part of this research has been submitted in the past, or is being submitted for a degree or examination at any other University. I declare that this thesis is based on some of the author's work published in scientific journals, conferences and symposium/workshops listed under the list of publications. These papers serve as the backbone and the foundation of this thesis.

I declare that some of these papers are contained *verbatim* in this thesis and all other sources used in this dissertation have been duly acknowledged.

Akingbesote Alaba Olu

Signature: _____ Date: _____

DEDICATION

This research is dedicated to Prof. Mathew Oluwasegun Adigun who brought me from grass to grace in the field of computer research and in the area of financial assistance throughout the period of my research at the University of Zululand.

ACKNOWLEDGEMENTS

My special appreciation goes to the Almighty God who opens the door when man shuts it. I thank Him for his mercy and grace given to me to start and complete this programme peacefully. Baba I say syabonga kakhulu.

There are a number of people without whom this thesis might not have been written and to whom I am greatly indebted.

First, I would like to express my sincere gratitude to my supervisor Prof. Mathew Olusegun Adigun who gave me all I needed to study at the University of Zululand and also for his constructive feedback throughout my stay with him. Special thanks to my co-supervisor Prof. S.S Xulu who was like my father. I thank you for your full support and understanding throughout my stay in South Africa.

I owe a lot to my family, particularly to my loving wife, Adebola Olubunmi Akingbesote who made time to see to the affairs of the home for the number of years I left her and the children Boluwatife, Jesulonimi and my Ogooluwa. I also want to thank her for making time to edit this thesis prior to the main editor's. I thank my siblings Ven and Mrs C.A Akingbesote, Mr and Mrs Lucas Akingbesote , Mrs. Stella Fajimolu, Mrs Taiwo olisa and Mr. and Mrs. Balogun for their full support throughout this programme. Special thanks to my Lord Bishop and his wife and the Rt. Rev and Mrs Akiode for their prayers and support throughout the period of my stay in South Africa.

I have been lucky enough to have the support of many good friends within and outside the University who have been the main sources of inspiration, laughter and joy. Let me mention but a few: Edgar, Bethel Pregason, Muthanga, Paul, ijeoma, Zaki, Nombuso, Kayode, Tosin, Dominic, Fakuade and my office partner Mr Ndlovu. I also like to thank the secretary Mrs. Nelli and the administrative staff of the department Miss Thabile and Ncnene. I am grateful to you all. I will definitely miss my two friends Mr. Akpagu Francis and Dr. Ogunyinka. I say thank you all.

I thank all my Pastors and members of Grace Covenant Centre Ugbe Akoko for their prayers and support throughout my stay in South Africa. Let me use this opportunity to appreciate and remember my late Pastor Debo Ajagunna for his support and the encouragement I got from him when I was about to come for this programme. May God continually keep your family.

I wish to acknowledge Adekunle Ajasin University, Akungba-Akoko for their support. I thank the management, the Dean of science and the Head (Computer Science) for their support and understanding throughout the period of my stay. I say thank you all.

I must also express my deep appreciation to my late mother, Chief Grace Olatunwon Akingbesote who died in 2011 while I was away. Mama, you have been the strong pillar of my life. Several times you did not eat because of my education. I sincerely thank you, may your soul rest in perfect peace. Amen.

TABLE OF CONTENTS

DECLARATION ii

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF PUBLICATIONS xi

LIST OF FIGURES ix

ABSTRACT xiv

CHAPTER ONE 1

INTRODUCTION 1

1.1 PREAMBLE 1

1.2 PROBLEM STATEMENT 4

1.3 RESEARCH QUESTIONS 5

1.4 RESEARCH PURPOSE OF STUDY 6

1.4.1 RESEARCH GOAL 6

1.4.2. RESEARCH OBJECTIVES 6

1.5 MOTIVATION 6

1.6 INTENDED CONTRIBUTION TO THE BODY OF KNOWLEGDE 7

1.7 RESEARCH HYPOTHESES 8

1.8 RESEARCH METHOLOGY 8

1.8.1 ANALYTICAL APPROACH 9

1.8.2 SIMULATION 10

1.9 OUTLINE AND ORGANISATION OF THE THESIS 11

1.10 CHAPTER SUMMARY 12

CHAPTER TWO 13

BACKGROUND - PERFORMANCE OF CLOUD E-MARKETPLACES 13

2.1 INTRODUCTION 13

2.2 EVOLUTION OF E-MARKETPLACES 15

2.2.1 TRADITIONAL MARKETPLACES 16

2.2.2 INTERNET MARKETPLACE 17

2.2.3 WEB SERVICES MARKETPLACES 19

2.2.4 GRID MARKETPLACES 22

2.2.5 CLOUD E-MARKETPLACES 24

2.3 STATE OF THE ART IN CLOUD E-MARKETPLACES PERFORMANCE 27

2.4 RESEARCH OPPORTUNITY 29

.....31
2.5 CHAPTER SUMMARY 32
CHAPTER THREE.....33
PERFORMANCE MODELLING OF CLOUD E-MARKETPLACE BASED ON NON PRIORITY FOR COST MINIMISATION 33
3.1 INTRODUCTION.....33
3.2 PROPOSED NON- PRIORITY MODEL.....34
3.2.1 MODELLING THE DISPATCHER-In USING M/M/1/K 36
3.2.2 MODELLING THE WEB SERVICE STATIONS.....40
3.3 COST MODEL FORMULATION.....42
3.4 SIMULATION I 43
3.4.1 NUMERICAL VALIDATION AND SIMULATION 44
3.5 RESULTS AND DISCUSSION I 44
3.6 ASPIRATION LEVEL MODEL.....50
3.7 SIMULATION II 52
3.8 RESULTS AND DISCUSION II.....52
3.9 CHAPTER SUMMARY 58
CHAPTER FOUR.....59
PERFORMANCE MODELLING OF CLOUD E-MARKETPLACE BASED ON TWO PRIORITY NON PREEMPTIVE MODEL.....59
4.1 INTRODUCTION.....59
4.2 THE ANALYTICAL MODEL DESCRIPTION OF TWO PRIORITY MODEL.....61
4.2.1 MATHEMATICAL MODELLING OF THE DISPATCHER-IN QUEUE 62
4.2.2 MODELLING OF WEB QUEUE STATION 65
4.2.3 MODELLING THE DISPATCHER-OUT.....67
4.3 NUMERICAL VALIDATION AND SIMULATION 68
4.4 RESULTS AND DISCUSION.....69
4.5 CHAPTER SUMMARY 75
CHAPTER FIVE.....76
PERFORMANCE MODELLING OF CLOUD E-MARKETPLACE BASED ON GENERALISED NON PREEMPTIVE MODEL FOR BALACING SERVICE LEVEL AND CONSUMERS' WAITING TIME 76
5.1 INTRODUCTION.....76
5.2 GENERALISED NON PREEMPTIVE MODEL.....77

5.2.1 MATHEMATICAL MODELLING OF DISPATCHER-IN QUEUE AS M/M/1/Pr 79

5.2.2 MATHEMATICAL MODELLING OF DATABASE QUEUE AS M/M/1 82

5.2.3 MATHEMATICAL MODELLING OF THE PRIORITISED WEB STATION QUEUE AS M/M/c/Pr 84

5.2.4 MATHEMATICAL MODELLING OF PRIORITSED DISPATCHER-OUT AS M/M/1/Pr 86

5.3 FORMULATING THE NON PRE-EMPTIVE COST MODEL 87

5.4 SIMULATION AND NUMERICAL VALIDATION 88

5.5 RESULTS AND DISCUSION 90

5.6 CHAPTER SUMMARY 96

CHAPTER SIX 98

PERFORMANCE MODELLING OF CLOUD E-MARKETPLACE USING DYNAMIC CONTROL MODEL 98

6.1 INTRODUCTION 99

6.2 ARCHITECTURE OF THE DYNAMIC CONTROL SYSTEM (DCS) 100

6.3 MEASURES OF EFFECTIVENESS 102

6.4 PROFITABILITY OF THE DCS 104

6.5 EXPERIMENTAL SETUP 107

6.6 RESULTS AND DISCUSSION 108

6.7 CHAPTER SUMMARY 114

CHAPTER SEVEN 115

CONCLUSION AND FUTURE WORK 115

7.1 SUMMARY 115

7.3 LIMITATIONS OF THE RESEARCH ..... **Error! Bookmark not defined.**

7.4 FUTURE WORK (why is formulation of SLA not outlined here See page 92) ..... **Error! Bookmark not defined.**

APPENDIX A 122

REFERENCES 126

LIST OF FIGURES

Fig. 1.1: A Queuing Performance Model for Computer Service in Cloud 8
Fig. 1.2: Proposed Model 9
Fig. 2.1: Evolution of Marketplaces 16
Fig. 2.2: Architecture of a Typical Traditional E-Marketplace 18
Fig. 2.3: An Illustration for Request for Services in Cloud E-Marketplaces 25
Fig. 3.1: Non-Priority E-cloud Marketplace Model 35
Fig. 3.2: Optimal Service Level Algorithm 43
Fig. 3.3: Analytical and Simulation 46
Fig. 3.4: Input and Output of Web Applications 47
Fig. 3.5: SLA- Waiting Time 48
Fig. 3.6: Expected Total Cost: Service Level 48
Fig. 3.7: SLA with Consumers Waiting Time 49
Fig. 3.8: Idle Period- Service Level 55
Fig. 3.9: Waiting Time of Consumers 56
Fig. 3.10: Aspiration Level 1 (ASL 1) 56
Fig. 3.11: Aspiration Level 2 (ASL 2) 57
Fig. 3.12: Combined Aspiration Levels 57
Fig. 4.1: Analytical and Simulation 71
Fig. 4.2: Two Class Non Preemptive Priority 72
Fig. 4.3: Two Class Priority with Total Waiting Time 73
Fig. 4.4: Two Class Non Priority with Total Waiting Time 74
Fig. 4.5: Performance of the Non Priority /Non Preemptive Priority Classes .74
Fig. 5.1: Simulation and Analytical 91
Fig. 5.2: Performance of the Five Classes Waiting with the Total Waiting Time 92
Fig. 5.3: Performance of Non Pre-emptive Priority and Non Priority.....93
Fig. 5.4: Total Consumers Processed and Randomly Generated Requests 95
Fig. 5.5: Total Cost - Service Level.....96

LIST OF TABLES

Table 2-1: Cloud E-Marketplaces: Some Definitions 14
Table 2-2: Meaning and Definitions of queuing theory terms 31
Table 2-3: Some Performance Measures and their meanings 31
Table 3-1: SM-Waiting Time-ETC 47
Table 3-2: Overall Analysis 49
Table 3-3: Waiting Time and Idle Period Distributions 55
Table 4-1: Result of the Simulation and Analytical 71
Table 4-2: Non Priority System (FCFS) 73
Table 5-1: Simulation and Analytical 91
Table 5-2: Detail Waiting time Results of the Non Pre-emptive Priority and the Non priority 92
Table 5-3: Service Level (SM) and the Waiting Time of Five Non Pre-emptive Classes 95
Table 5-4: Total Cost 96
Table 6-1: Detail results obtained based on the used parameters under DSC and the classical fixed methods 113
Table 6-2: Cost-Benefit Analysis of DCS and classical methods 113

LIST OF PUBLICATIONS

i. A.O. Akingbesote, M. O. Adigun, J . Oladosu and E. Jembere and I.Kaseeram, "MODELLING the cloud E-Marketplaces for cost minimisation using queuing model" Australian Journal of Basic and Applied Science, 8(4), 59-67, 2014.

ii. Alaba Olu Akingbesote, Mathew Olusegun Adigun, Sibisuso Xulu, and Edgar Jembere, "Performance Modelling of Proposed GUISET Middleware for Mobile Healthcare Services in E-Marketplaces," Journal of Applied Mathematics, vol. 2014, Article ID 248293, 9 pages, 2014. doi:10.1155/2014/248293

iii. A.O. Akingbesote, M. O. Adigun , J. Oladosu, E. Jembere, "A Quality of Service Aware Multi- Level Strategy for Selection of Optimal Web Service," in proceedings of the 5th IEEE ICAST International Conference on Adaptive Science and Technology, Pretoria, South Africa, pp.25-27, Nov. 2013.

iv. A.O. Akingbesote, M. O. Adigun, J. Oladosu and E. Jembere and I.Kaseeram, "Modelling the cloud E-Marketplaces for cost minimisation using queuing model," in proceedings of International Conference on Information Technology (ICIT) Bali, Indonesia, 12-13 Dec. 2013.

v. A. O. Akingbesote, M. O. Adigun, J. Oladosu and E. Jembere[1] and I. Kaseeram, " The Trade-off between consumer's satisfaction and resource service level by E-Market providers in E-Marketplaces" in proceedings of International Conference on Electrical Engineering and Computer Science (EECS), Hong Kong, pp395-404, 19-20 Dec. 2013.

vi. A.O Akingbesote, M.O Adigun, E. Jembere, M.A.Othman, I.R Ajayi. "Determination of Optimal Service level in Cloud E-Marketplaces Based

on Service Offering Delay" In proceedings of IEEE International Conference on Computer, Communication and Control Technology (I4ct), Malaysia. pp 283-288, 2-4 Sept. 2014.

vii. A.O Akingbesote, M.O Adigun, E. Jembere M. Sanjay, I.R Ajayi. "Performance Analysis of Non-Preemptive Priority with Application to Cloud E-Marketplaces" In proceedings of the 6^{h} IEEE ICAST International Conference on Adaptive Science and Technology, Ota, Nigeria, pp 1-6, Oct 29th-31st 2014.

viii. A.O Akingbesote, M.O Adigun, S.S.Xulu, E. Jembere, "Performance Evaluation of Cloud E-Marketplaces using Non Preemptive Queuing Model "In the proceedings of IEEE World Congress on Sustainable Technologies (WCST-2014) proceeding, London,UK, December 8-10, 2014.

ix. A.O Akingbesote, M.O Adigun, X. Xulu. "Comparative Analysis of Cloud E-Marketplaces Performance under Non-Priority and Non Pre-emptive models using Queuing Theory in 3rd Faculty of Science International Conference, Nigeria, p. 8, May 11-15, 2015.

- **SYMPOSIA/WORKSHOPS ATTENDED**

i. A.O Akingbesote, M.O Adigun, E. Jembere. A Quality of Service Aware Multi- Level Strategy for Selection of Optimal Web Service, in 7th Annual Faculty of Science and Agriculture University of Zululand Symposium, KwaDLagenzwa, South Africa, 2012.

ii. A.O Akingbesote, M.O Adigun, E. Jembere Optimal selection of server machines in cloud E-Marketplaces using Queuing model in 8th Annual Faculty of Science and Agriculture University of Zululand Symposium, KwaDLagenzwa, South Africa, p.20 1st Nov. 2013.

iii. A.O Akingbesote, M.O Adigun, E. Jembere , J Oladosu "MODELLING the Cloud E-Marketplaces for Cost Minimisation using Queuing Model " in 5th IEEE ICAST Symposium on Adaptive Science and Technology, Pretoria, South Africa, 28th Nov. 2013.

iv. A.O Akingbesote, M.O Adigun, X. Xulu . "Balancing Service Level and Consumers' Waiting Time in Cloud E-Marketplaces Using Aspiration Model" in 9th Annual Faculty of Science and Agriculture University of Zululand Symposium, KwaDLagenzwa, South Africa. p. 8 28st Oct. 2014.

v. Akinola, A.T., Adigun M.O, Akingbesote A.O "QoS-Aware Single Service Selection Mechanism for Ad-Hoc Mobile Cloud Computing" in IEEE proceedings on International Conference on Computing, Communication and Security (ICCC), Dec. 4th -6th 2015. Mauritius.

- **AWARDS**

i. Best PhD oral presentation 2nd Prize awards 2012 in the 7th Annual Faculty of Science and Agriculture University of Zululand Symposium South Africa 2012.

ii. Best paper award presenter on Cloud Computing in the IEEE 6th International Conference on Adaptive Technology (ICAST). Nigeria. 2014

ABSTRACT

Cloud E-Marketplaces are virtualised global network markets that allow the exchange of digital information through a broker for the purpose of conducting and delivering cost effective business services. Prior to the Cloud, E-Marketplaces were the Traditional Internet web service and Grid E-Markets. All these Marketplaces experienced challenges which led to the creation of Cloud E-Marketplaces for service delivery in form of Software, Platform and Infrastructure.

Research on Cloud E-Marketplaces has concentrated on sub-domains including security [1], energy [2] and privacy [3] but little has been done with regard to optimization for resource management in Cloud performance [4]. One major parameter that is required to bring the cost down is the number of server machine(s) used by a service provider during service provisioning. A second one is the performance measures like waiting time which are used to determine the effectiveness of Cloud E-Marketplaces' performance.

In order to establish the thesis that minimizsation of server machine cost and consumer waiting time in the context of non-priority and non-pre-emptive priority policy is imperative, the study accomplished the following:

i. It extensively reviewed the existing body of knowledge on the performance of E-Marketplaces.
ii. It identified the need to re-engineer the existing Cloud E-Marketplace architecture as networks of queues with parallel web stations with a feedback from scheduler in the context of non-priority and non-pre-emptive policy without dedicating any web station to any class to achieve optimal service level.
iii. It evaluated the Non Priority First Come First Serve, FCFS service discipline and the Non-Preemptive model in order to gauge the performance impact on consumers' waiting time and providers' cost.

iv. It formulated a cost structure that balances the server machine (Service Level) and consumers' waiting time on both non-priority and non-preemptive models.

v. It formulated a dynamic waiting time optimisation control mechanism that further addressed the issues of service over and under-provisioning.

The contributions are:

i. the evaluative study of non-priority queues in series against the generalised approach that uses a single point of entry as proposed by others in the literature. This was used to determine the optimal service level and consumer waiting time.

ii. the exhaustive evaluation of a novel non-preemptive architectural model of the Cloud E-Marketplace with each of service stations modeled as M/M/c/Pr against the M/M/1 proposed in the literature. This model was unique in that it:

i. explored a different mathematical and simulation concept and also;

ii. resolved the challenge of dedicating or allocating servers to a particular consumer class thereby reducing consumers' waiting time.

iii. investigated E-Marketplaces under the non-priority and also the two service non pre-emptive and the generalised models;

iv. introduced the novel concept of profitability and Cost Benefit Ratio by using the Dynamic Control Model (DCM) over the Fixed Server Model (FSM).

CHAPTER ONE

INTRODUCTION

This chapter describes the general concept of E-Marketplaces and Cloud E-Marketplaces as the new paradigm of computing. The author identifies the problem statement, research questions, goal and objectives. The motivation, research methodology and the intended contributions are also explained. The chapter concludes with an outline and organisation of the thesis.

1.1 PREAMBLE

E-Marketplaces are local communities of service providers and requestors (service consumers) organised in vertical markets and gathering around portals [5]. These E-Marketplaces allow consumers to shop for services from anywhere in the world based on a pay-as-you-go model [6]. The concept of using Internet and Electronic media has revolutionised E-Marketplaces, thereby increasing the number of macro and micro enterprises participating in the market [7]. Because the number of participants is increasing daily, especially the providers of services, new marketing strategies and business models of selling and buying are also increasing [8]. Among the expected benefits of E-Marketplaces are increased exposure to global markets [9], enhanced communication [10] and reduced transaction costs [11] due to the aggregation of needy buyers generated by the E-Marketplaces.

As new cloud E-Market providers are emerging every month and many Traditional service providers rebrand services as cloud hosting, so is there an increase in the number of consumers in these Marketplaces. Three things are paramount in these markets: the Cloud implementations, the performance, and the optimal provisioning of these server machines that will minimise both cost and consumer waiting time. Although researchers are working on these issues,

the literature reveals that most research being done is on Cloud E-Market implementation, with little work on performance [12][13]. The issues of performance as well as that of optimal provisioning of server machines to maximis profit and minimise the consumers' waiting time have been a great challenge [14] [5] [15], [16]. On the issue of performance, for example, the performance challenge rose in 2008 from 63.1% to 82.9% in 2009, as reported in the International Data Corporation (IDC) report [17] [18][19]. This is an increase of 19.8% as against the Security challenge which only increased by 12.9%.

As the markets grow, firstly, most Cloud E-Marketplace providers are implementing various service offerings to their consumers. For example, Amazon Elastic Compute Cloud (EC2) offers three different services: Reserved, Spot, and On-Demand [20]. The Reserved offering requires paying in advance with the assurance of no or minimal delay. This will be good for higher priority service consumers. Under the Spot offering, services are allocated in advance. The On-Demand offering has no facility for advance payment or reservation and there is no commitment. Secondly, Cloud E-Marketplace providers deliver different application performance results based on certain parameters like, geographical location, cloud platform architecture and the service provisioning being offered. While many researchers concentrate their efforts on how this will work, the performance impact of service provisioning under different disciplines is yet to be fully explored in the context of Cloud E-Marketplaces.

One important aspect of performance is the time taken by E-Cloud Marketplace providers to respond to consumers' requests [21]. This waiting time, which is a key source of competitive advantage, has long been recognised as very important. For example, in the 1990s Traditional Marketplaces like McDonalds offered consumers their meal free of charge if the order was not served within 2 minutes [22].

Most existing literature that delves into performance impact on Cloud E-Marketplaces focuses on the exogenous Non Priority model with emphasis on First Come First Serve discipline [12] [23][4]. As the number of server farms increases, with consumers demanding different service disciplines, some scholars [24][25][26],[27] have used the Preemptive service discipline in the area of networking, while other scholars [28][13] use the Preemptive service discipline and migration in the context of Cloud E-Marketplaces. However, the literature reveals that in practice, pre-emption and migration of virtual machines are costly [29][14]. It also shows that pre-emption leads to an increase in response time to consumers' requests especially when the requests are deadline constrained [30]. The researcher's work extends existing and widely adopted theories to the exogenous Non Pre-emptive Priority model.

The third challenge, as mentioned earlier, is the optimal provisioning of these server machines in order to minimise both the cost and consumers' waiting time. This is because over- provisioning of servers may increase the cost incurred by the cloud E-Market providers in terms of both electrical energy cost and carbon emission [31], [32], while under-provisioning may cause long waiting time for consumers, which may lead to a breach of the Service Level Agreement (SLA) where cloud E-Market providers may pay for such waiting costs. Therefore, optimal provisioning of these server machines is imperative to maximis profit and minimise consumers' waiting time [15][6]. Determining the optimal number of server machines (resource service level) to be instantiated from each pool that can maximis the data centre profit without violating the SLA is a challenge [12],[33],[23]. Overall, this research models the Performance of web service E-Marketplaces in the Cloud based on consumers' waiting time and providers' cost.

1.2 PROBLEM STATEMENT

It has been mentioned that most research efforts concerning Cloud E-Marketplaces have been on the implementation, while less attention has been given to performance related issues [34],[35]. Due to the dynamic and virtualised nature of cloud environments, diversity of users' requests and time dependency of load, providing expected quality of service while avoiding over-provisioning is not a simple task [12],[31], [32],[36],[37],[38].

In a typical cloud market response time is of interest to every consumer and is also a key source of competitive advantage for any cloud E-Market provider. A low level of service may be inexpensive, at least in the short run, but may incur high costs of consumer dissatisfaction, such as loss of future business and actual processing costs of consumer complaints. A high level of service will cost more to an E-cloud provider but will result in lower dissatisfaction costs [39]. Balancing the trade-off between resource service level and consumers satisfaction in terms of waiting time is a challenge [40].

Most works on performance in Cloud E-Marketplaces, see[33] and [15] for example, are based on performance analysis and profit maximisation of IaaS, with less attention in the context of SaaS hosted application. Therefore, looking for solutions that will minimise cost without adversely affecting the consumers in the context of SaaS is imperative [16].

In order to ensure that the QoS perceived by end clients is acceptable the providers must exploit techniques and mechanisms that guarantee a minimum level of QoS. Also, accurate prediction of service performance to consumers based on systematic statistics allows a service provider not only to guarantee good QoS but to avoid over-provisioning to meet SLA [12]. Some of the QoS measures as identified by Garcia et al [41] are response time, throughput, availability, reliability, security and cost. The primary aspect of QoS considered in this work is related to response time and cost.

With the increase in the number of clients operating in the Cloud E-Marketplaces several experiments are being carried out by researchers, for instance, the cloudsim developed in [42] yielded a positive result, but a topic for future research is the issue of how pricing, provisioning policies and expenditure incurred by the service provider can be incorporated into the CloudSim [42]. Also, service availability and response time are two important quality measures in Cloud E-Market from users prospective. Quantifying and characterising such performance measures requires appropriate Modelling [35]. This research envisages good cost minimisation of the server machines and better waiting time performance.

1.3 RESEARCH QUESTIONS

1.3.1 How can optimal service level with better consumers' waiting time be achieved without the breach of SLA in E-Marketplaces?

1.3.2 How can a better performance be achieved to improve consumers' satisfaction, especially in the context of different service provisioning?

1.3.3 What strategy could be adopted to optimise the combined costs and performance to create better capacity planning without breaching the SLAs?

1.3.4 How can a better strategy be put in place to minimise Infrastructural and platform costs without adversely affecting consumers in the context of SaaS providers hosted software services?

1.3.5 What prescriptive measure should be in place to avoid service over- provisioning and under- provisioning in time dependent Cloud E-Marketplaces?

1.4 RESEARCH PURPOSE OF STUDY

1.4.1 RESEARCH GOAL

To optimise the management of a typical Cloud E-Marketplace architecture based on consumers' waiting time and providers' costs.

1.4.2. RESEARCH OBJECTIVES

The following are the research objectives for this study:

1.4.2.1. Investigate the current trend in Cloud E-Marketplaces based on consumers' waiting time and providers' cost.

1.4.2.2. Analyse the performance impact of Cloud E-Marketplaces on consumers' waiting time under the exogenous Non priority discipline and other service disciplines.

1.4.2.3. Develop a mechanism that will strike a balance between cost of offering a service and cost of waiting experienced by consumers without breaching the SLA.

1.4.2.4. Formulate an optimisation strategy based on existing knowledge with the aim of achieving the accurate measuring of performance by studying the significant time spent when variations occur.

1.5 MOTIVATION

The Cloud E-Market has been the subject of a lot of research interest in recent years. Although much attention has been on the implementation issue only a small portion of the research has been focused on performance. Response time, or Waiting time, is of great interest in any Cloud E-Marketplace. Fast response time or low waiting time leads to a perception of high availability of web services, while slow response time or high waiting time degrades the performance of web services.

Cost minimisation which is a better way of profit maximisation, is crucial to any Cloud service provider. Balancing the trade-off between cost minimisation and consumers' agreeable SLA is not a simple task.

With the increasing trend of consumers moving to the Cloud, the need to study and evaluate the performance of Cloud E-Marketplaces to see how this could further improve consumer-provider satisfaction motivated this research study.

1.6 INTENDED CONTRIBUTION TO THE BODY OF KNOWLEDGE

Most research efforts [12],[43], [44] ,[45], [4] have gone into adding novel ideas to the Cloud E-Marketplaces body of knowledge by Modelling the cloud as queue with only one point of entry, processing and dispatching. These efforts have been extended by other scholars [37],[23] as queue in series. The current study is intended to further extend contemporary thinking about Cloud E-Marketplaces as networks of queues with feedback from the database. The research approach is to propose a model which addresses the issue of variation of parameters like the service rate and virtual machine configurations. It will also systematically bring in other statistical information relevant to Cloud E-Marketplaces when necessary.

This work acknowledges the efforts of others [42], [23], who have worked on the performance impact on consumers' waiting time, especially in the context of IaaS. The contributions of these authors opened up the opportunity to extend the body of knowledge in the area of SaaS Cloud hosted software services by considering the trade-off between consumer waiting time and provider service level. In order to answer the research questions this work explored existing and widely adopted theories from the exogenous Non Priority model with its emphasis on First Come First Serve discipline [12] [4] [28] [4] and the Preemptive service discipline [28] to the Non Preemptive model.

1.7 RESEARCH HYPOTHESES

1.7.1 The total waiting time in the queue by consumers is independent of the service discipline but the waiting time distributions of the classes in the Non Preemptive priority differ while those of Non Priority have equal distributions.

1.7.2 Increasing the number of servers reduces consumers' waiting time but this may have an adverse effect on the cost incurred by Cloud E-Market providers.

1.7.3 Designing a queuing system to reduce expected waiting time for a particular group of consumers implies giving such group a higher priority over other groups, especially if the other group has higher mean service time.

1.8 RESEARCH METHOLOGY

The idea for this research was born from the simple operational mode of computing services in the cloud as being a contention problem where consumers are contending for services and a queue is built up as shown in Figure 1.1.

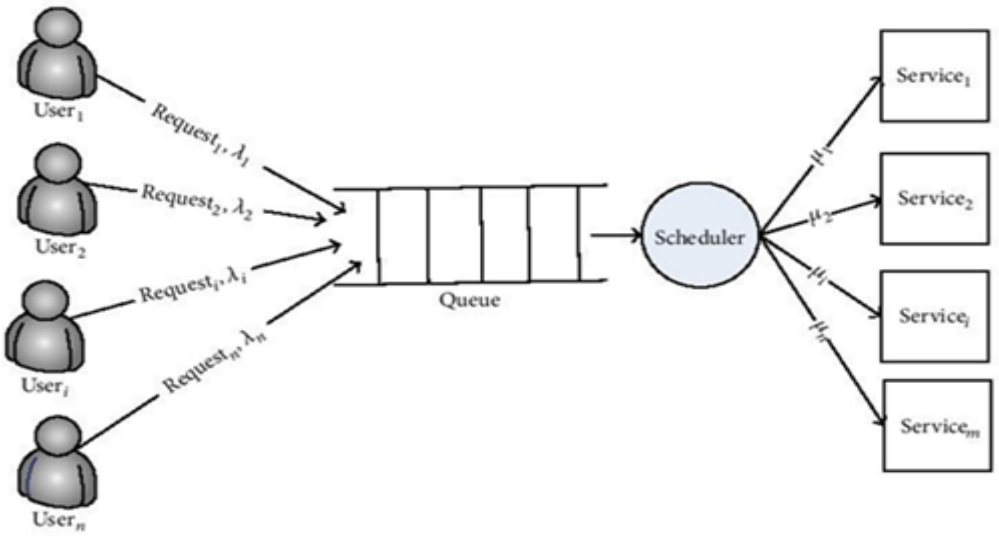

Fig. 1.1: A Queuing Performance Model for Computer Service in Cloud
Source: Guo et al, JAM 2014

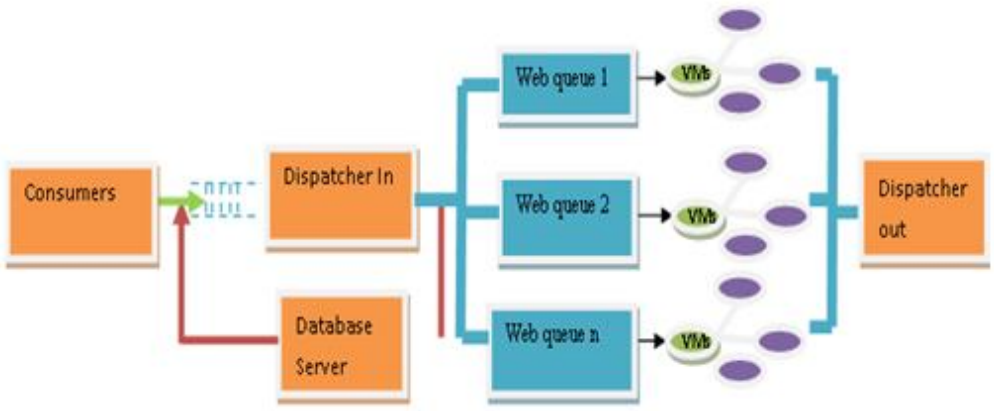

Fig. 1.2: Proposed Model

This idea gave birth to the proposed model shown in Figure 1.2 that will be used to accomplish the research objectives. Here, the Cloud E-Marketplace is modeled as a network of queues with a feedback loop from the database. All requests are sent to the Dispatcher-In server where they are then distributed to the web queue stations for service provisioning. The processed requests then move out through the Dispatcher-Out.

Because of this contention, the solution approach in this study will be based on the Analytical technique using queuing theory and the use of Simulation as the analysis tool.

1.8.1 ANALYTICAL APPROACH

The process of consumers or clients entering the Cloud E-Marketplace is usually in the form of a queue, and most performance problems are related to a queuing delay caused by contention for resources. Therefore, this study uses the queuing analogy to formulate the mathematical model. The justification for using this analogy is based on the work reported in [43]. Furthermore, the queuing models predict the performance behaviours of systems that attempt to provide services for randomly arriving demands [47]. In this study the cloud served as the system under the control of the cloud e-Provider that did the service provisioning, and

the web applications were the consumers that were making random demands for service. Two input parameters were of importance to both the analytical and Simulation solutions, namely the arrival rate and the service rate of the consumers. The investigation was systematically phased.

In the first phase, the researcher applied appropriate Kendall notation to model the Cloud E-Marketplace as a non-priority system. The performance impact was evaluated based on the consumers' waiting time. This was based on the systematic cost function formulated using operational and service cost. The quality measure result obtained was then used as part of the parameters in the cost function to be minimised.

The second phase was based on the priority model with appropriate Kendal notation. The idea was that a real –life queuing situation contains priority considerations. The third phase was a comparative study of the first and the second phase. The impact of the results from the comparison on performance was evaluated and then analysed.

The last phase is the dynamic control mechanism that considered a situation where there were certain numbers of virtual machines, say c_1, that will always be available. If the number of consumers in queue exceeds a certain critical value, say M_1, then additional servers were added but there was a limit of, say C_2, to how many servers could be added.

The analytical results in each of the phases were compared with corresponding simulation results to determine the degree of accuracy.

1.8.2 SIMULATION

This research used simulation to mimic the real E-cloud Marketplace behaviour. This was because simulation models have been found to be closest to reality when MODELLING and analysing business processes. It has always been the best

mechanism to dynamically model different samples of parameter values such as arrival rates or service intervals, especially those parameters that point to process bottlenecks when investigating suitable business alternatives. Therefore, the Arena Rockwell software tool was used as Discrete Event Simulator (DES) to investigate and analys the performance of Cloud E-Marketplace based on selected performance measures, for example, the waiting time. Experiments were conducted to check the validity of models, and the results obtained from the simulation were compared to the analytical solution.

1.9 OUTLINE AND ORGANISATION OF THE THESIS

The remaining part of the thesis is organised as follows. Chapter 2 covers the trend in the E-Marketplaces and the current state of the art in Cloud E-Marketplaces. The shortcomings of the current state are identified and the contributions of this research to the body of knowledge are clarified whilst the chapter closes with a brief summary.

Chapter 3 examines the Cloud E-Marketplace under the exogenous non priority model to achieve optimal service level and better waiting time without the breach of Service Level Agreement (SLA). A further extension was carried out using the aspiration model to formulate a strategy that was adopted to optimise the combined costs and the waiting time performance to form a better capacity planning in an environment where optimal solution was difficult or almost impossible. Experiments were conducted and results were analysed and discussed. This chapter wraps up with a brief summary.

In chapter 4, a study of how a better performance could be achieved to improve consumers' satisfaction especially in the context of two different service offerings is presented using the two priority non preemptive approach. The queuing theory was used to formulate the mathematical model while a

simulation of the model demonstrated the real life scenario. Results from the experiments conducted are analysed and discussed.

The model discussed in chapter 5 was an extension of chapter 4 such that a generalised model was used to expose many service offerings in contrast to two using a non-preemptive queuing theory. Results of the experiment conducted are presented.

The generalised approach is taken further in chapter 6 by adding a dynamic control mechanism that puts in place a prescriptive mechanism that avoids service level over-provisioning and under-provisioning in time dependent Cloud E-Marketplaces. The analysis of the relevant experiments is presented.

Chapter seven, being the final chapter, is used to convey the achievements of the thesis. Some of the limitations of the reported results are outlined and a number of future research suggestions are made for interested parties to consider.

1.10 CHAPTER SUMMARY

In chapter one, the background of this research is discussed. The concept of the research which is centred on the performance of web service Cloud E-Marketplaces based on consumers' waiting time and providers' cost is introduced. The problem statement, research questions, goal and objectives are identified. The motivation, research methodology and the intended contributions are also overviewed. This chapter closes with the outline and organisation of the thesis.

CHAPTER TWO

BACKGROUND - PERFORMANCE OF CLOUD E-MARKETPLACES

This chapter presents the current trends in E-Marketplaces. A detailed literature review is presented on the work done in the area of performance evaluation in E-Marketplaces. An overview of the key features is given and a systematic strategy towards addressing the shortcomings is then introduced. Section 2.1 discusses the trend in Cloud E-Marketplaces. In Section 2.2 the evolution of Cloud E-Marketplaces is discussed and the state of the art in Cloud E-Marketplaces discussed in section 2.3. Section 2.4 further discusses the contribution of this research. This is followed by a brief chapter summary in Section 2.5.

2.1 INTRODUCTION

The concept of the E-Marketplace started in the early nineteen seventies where some systems were developed in the area of airline reservation systems, for example, United Airlines' Apollo or American Airlines' Sabre [48]. In this system consumers were able to book flights through an agent which today is referred to as a broker. One major issue was the accessibility to the system which required a specialised expert broker. Another early example was that of J.C. Penney's Telaction Home-shopping System (see in [49]). This was an electronic home-shopping system that allowed consumers to shop via a cable television channel and a push-button phone. These E-Marketplaces had major drawbacks, among which were the lack of competiveness and the inability to create an air of excitement [48]. From these humble beginnings came the evolution of the E-Marketplace that now allows organisations to open their shops on the Internet and also enables millions of consumers to participate in the global online Marketplaces.

Several E- Marketplace definitions have been suggested by various authors as shown in Table 2-1.

Table 2-1: Cloud E-Marketplaces: Some Definitions

Author/Reference	Year	Definition
Malone, Yates, and Benjamin [49]	1989	Networks that let customers compare and order offerings from competing suppliers.
Bailey & Bakos [50]	1991	A market system that allows buyers and sellers to exchange information about market prices and product offerings, thus representing an investment in multilateral information sharing.
Archer and Gebauer [51]	2000	The E-Marketplace is a virtual Marketplace where buyers and suppliers meet to exchange information about product and service offers, and to negotiate and carry out business transactions.
Russ [52]	2001	A Web-based information system, where multiple suppliers and multiple buyers can undertake business transactions via the Internet.

All the definitions in Table 2-1 have provided the researcher with either the functions or characteristics of E-Marketplaces. For the remainder of this thesis E-Marketplace is defined as a virtualised global network market that allows the exchange of digital information, sometimes through a broker, for the purpose of conducting and delivering effective business services.

2.2 EVOLUTION OF E-MARKETPLACES

The idea of the E-Marketplace did not just come from a vacuum but originated from the Traditional Marketplaces. The Traditional Marketplaces dealt primarily with goods produced or distributed personally by the merchants themselves and were not adaptable to modern mass production and distribution systems [48], [53]. The removal of trade barriers , Industrialization in most parts of the world, the emergence of global markets and the use of Information and Communication Technology (ICT) changed the role of Traditional markets from its dominant position to supplementary as E-Marketplaces come into the main stream.

E-Marketplaces facilitate trading transactions for buyers and sellers through the use of electronic means [54]. Although the idea of E-Marketplaces started in early nineteen seventies, the conceptualisation of this idea evolved in the mid-1940s when Selevision was the E-Market system used in Florida to remote E-Market citrus fruits [55] [54]. In [56], but the impact of the E-Market in improving E-Marketplace transactions started with the initiation of computer-based pilot projects in the 1970s. In sections 2.2.1 to 2.2.5 the researcher traces the evolution of Marketplaces, starting with the Traditional Marketplaces and various E-Marketplace concepts such as the Internet market, Web Service, Grid and Cloud E-Marketplaces as shown in Figure 2.1.

Fig. 2.1: Evolution of Marketplaces

2.2.1 TRADITIONAL MARKETPLACES

Traditional Marketplaces offer places for wholesalers and retailers to co-operate and also offer the consumers the benefit of where to go directly for a given item. These markets allow the buying and selling of goods with both the buyers and the consumers having a direct link. However, a broker or intermediary will sometimes be involved to mediate in price bargaining between the buyer and the consumer, but such an intermediary has to be trusted by both the buyers and the sellers. These markets have some advantages in that: `

i. They require no special medium before it they can operate. For example, rather than customers going online for a business transaction, such business or service could be brought to potential customers with the use of some Traditional methods.
ii. They give room for price negotiation and bargaining.
iii. In addition, both literates and illiterates can participate in the market.

However, some major limitations of the Traditional Marketplace are:

i. Lack of a good Traditional marketing strategy.
ii. Inability to globalise the market.
iii. Slow uptake of cashless policy.
iv. High risk of products transfer/movement.

The emergence of the Traditional E-Marketplace came about as a result of some of these limitations. The Traditional E-Marketplace is a Web portal where buyers and suppliers come together to explore new business opportunities [57]. These markets allow the buying and selling of goods, with both the buyers and the consumers having direct link. This market uses digital means to brand products or logos. The idea promoted by the Digital markets strategy [58] toward the Traditional E-Marketplace is similar to people finding or getting a particular business through a referral or a network and eventually building a rapport with them.

In [57], the author attributed the challenge of Traditional E-Marketplace to that of supporting only a single business model, which is ineffective in dealing with all but the simplest kinds of exchanges. As a result, the Traditional concept of an E-Marketplace having broker mediating between buyer and supplier is not as suitable for every kind of product transactions as was initially expectedTo overcome this barrier, the concept of the Internet Marketplace was introduced.

2.2.2 INTERNET MARKETPLACE

In [59], electronic Marketplaces is defined as the notion of paperless exchanges of business information using EDI (electronic data interchange), electronic mail (E-mail), and electronic bulletin boards, electronic funds transfer (EFT), and other similar technologies. The Internet E-Marketplace allows the full range of using internet technology to fulfill its goal. This goal is to attract the biggest possible number of consumers and providers who will become members of that Internet E-Marketplace. This is done by matching consumers' needs against providers' selling offers [60] [50].

In [61][62], the authors identify three main elements in the structure of an E-Marketplace; these are the owner or the operator of the market place, the type of transaction being offered and the resources being offered to the consumers.

Three main types of Internet E-Marketplaces are identified in [61] [57] [63]: the seller-driven market, the buyer-driven market and the open market. In [57], the author defined the seller-driven market as an E-Marketplace promoted by a consortium of suppliers who place offers within the same industry or service sector. The buyer-driven market is maintained by a group of buyers who aggregate purchase needs so as to achieve advantageous conditions when buying from suppliers and the open market is an E-Marketplace owned by an independent third-party. In addition to these, in [64], the author identifies the fourth one as the technology driven market. This is similar to the independent market but the motive behind the set up may be different from that of Independent market. A typical diagrammatical structure of these is shown in Fig. 2.2 with detailed explanation given in [57]. The supplier and buyer are similar to the researcher's consumer and producer in this research.

Fig. 2.2: Architecture of a Typical Traditional E-Marketplace
Source: IJCM 17(3) 2004

Apart from the benefits offered by Traditional Marketplaces, E-Marketplaces have some characteristics that make them better than Traditional Marketplaces. Some of these characteristics are richness, cost effectiveness and provisioning of extra value [65]. Also, E-Marketplaces expand the choices available to consumers

and give providers access to new consumers, thereby reducing the transaction costs for all participants [66].

Internet E-Marketplaces differ from Traditional Marketplaces in that the business transactions occur in communication networks without the necessity of the clients and the producers seeing each other. This virtual, dynamic and real time platform allows the consumers or the clients and the producers to communicate through the use of Internet technologies. Hence, E-Marketplaces are regarded as one important part of E-business solutions in the process of enabling supply chain integration to maintain business value and growing competitive necessity [67]. The idea of this new paradigm is to reshape the business process by making available various types of products to consumers. It is a paradigm for building distributed computing applications over the Internet [68]. This idea was not fully exploited until the emergence of Service Oriented Architecture (SOA).

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

2.2.3 WEB SERVICES MARKETPLACES

In [69] Web services are defined as self-contained, self-describing, loosely-coupled computational components designed to support machine interaction via a distributed or centralised network. The web services E-Marketplace is a community that allows producers to advertise their products on the web for the consumer to use. In other words, it is a local community of service providers and service consumers organised in vertical markets and gathered around portals [5].

A number of scholars have worked on specific features of web service Marketplaces, including web service discovery [70], [71], selection [69], [72], [73] [74] [75] 69] [76] [77] [78] [79] [80] and composition[81] [82] [68]. The researcher is interested in service selection. For example in service selection, a lot of work has been done on the selection of optimal web service. In [75] the authors proposed a new analysis of service selection and evaluated the proposed algorithm. This was done by designing a mixed linear integer program for optimising service compositions based on service response time and energy consumption. In [83], the author use the QoS-based service selection approach to compose web applications by discovering feasible web services based on functionalities and QoS criteria of user requirements. The results show that the algorithm performs well and increases system availability and reliability. The use Particle Swam Optimisation (PSO) method was adopted to select the optimal service in [74] . This was done by defining the position, and the velocity equation. While this achieved a considerable solution global convergent ability was a challenge. This was further improved by using the Niche Particle Swarm Optimisation (NPSO) algorithm that integrates the Simulated Annealing (SA) and niche technique into the Particle Swarm Optimisation algorithm [84]. In [76] the authors proposed a genetic based service selection algorithm where the developed algorithm is compared with the heuristic algorithm based on both time complexity and the non-functional characteristics called reliability rate. The significant contribution to optimal service selection in [73] was made by comparing the two service selection algorithms i.e, the GA and the PSO using response time as the metric with multiple users [77]. The end results indicate that PSO performs better over GA for single and multi user service selections. While optimal selection is achieved in the context of the given set in the domain by these algorithms, there was silence on what happens when there are ties in that set based on service consumers' requests. Furthermore, the associative classification algorithm has been used to classify candidate web services into

different QoS levels. Semantic matching is then used to rank the most qualified web services based on their functional quality. In [78], these authors proposed a Multi Criteria Approach for Web Service Discovery. In their work, the authors introduced QoS parameters which allow the user to find relevant services that correspond to his/her preferences and enable him/her to also gain in terms of time by minimizing his/her search space. Review some of the techniques in the context of the QoS based approach was investigated in [79]. This was extended by Priya [80].

One issue that was not addressed in [74], [76][72] is when the competitive differentiation is zero among the selected optimal web services i.e. when there are ties. When this occurs, scholar like [78] only allow client the freedom to choose their own scenario and to gain in terms of processing time. To resolve this, the work in [85] propose a QoS based multi level selection algorithm for a situation where there are ties between optimal web services. The authors consider the Information services and use the non-deterministic Quality of Service metrics. An algorithm was formulated and an experiment conducted using a web service data set Quality of Service information as the input parameters. The experimental results show that the proposed model satisfies service consumers' requests based on non-functional requirements.
This justification for this approach was based on the evaluation report in [79] that the QoS based approach requires less expensive middleware, allows dynamic service selection, is fair to the clients and gives room for QoS extension.

These web service E-Marketplaces provide both the providers and consumers with the following benefits:

i. The use of modern technologies to improve communication between the producers and consumers is allowed.

ii. Purchasing operations are improved. Also, a purchasing community is established and consumer demands are satisfied in more integrated ways.
iii. The opportunity is created for a service provider to deliver value-added, integrated (packaged) services by composing existing E-services possibly offered by different enterprises [86].
iv. Providers get the enabling environment to generate extra sales by providing a way to reach new customers that are difficult to get through Traditional marketing methods.
v. A competitive market is established thereby allowing the springing up of many products of the same function and therefore giving the consumers the opportunity to have a multi-level based selection strategy for selecting services [85].
vi. Room is given for market Service Discovery, Selection and Composition.

While these benefits have had a positive impact on both consumers and providers, there are several challenges faced by this Marketplace. Among these are:

i. The need to see computing as service that is delivered to consumers over the internet from large-scale data centres - or the clouds, Rather than purchase of products.
ii. The need for consumers to invest heavily in building and maintaining complex IT infrastructure [87] and
iii. The high costs of maintaining the equipment and human resources.

2.2.4 GRID MARKETPLACES

While the use of the Web Service Marketplace have been successful, especially with the business class, using this market for high computational power is a

challenge, as are the high costs of maintaining the equipment and human resources [87] [88]. This led to the creation of the Grid E-Market Technology. This is a market where computational power is purchased by consumers (Consumers/Applications) through the use of middleware or a resource allocation broker. Four major features distinguish the Grid Marketplace from others [89][90]. These are:

i. Collaboration among members of the grid community with the use of powerful middleware.
ii. Integration of different heterogeneous hardware infrastructure.
iii. The distributed paradigm and
iv. Secure access through the use of a powerful security mechanism to grant the right delegation.

The vision of the Grid architects is for consumers to draw computation power from a distributed pool of resources in a way similar to that in which household appliances draw electrical power from a power utility seamlessly and ubiquitously. Basically, Grid market technology consists of clusters of computers under the control of powerful middleware.

Although these Grid markets have been viable in terms of high performance, exploitation of underutilised resources, resource balancing and wide- scale distributed computing[91] [92], this market has some challenges, among which are:

i. Lack of a distributed and robust resource allocation mechanism [93][94].
ii. Inability to provide a good accounting mechanism[95].
iii. Inability of the centralised system to scale in proportion to the potential computation power that will be available as high performance networking becomes available[89].

iv. Inability of the architecture to fully cope with the business world[96].

These challenges were already foreseen as far back as 1969, as quoted in [97] : "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".

2.2.5 CLOUD E-MARKETPLACES

The vision of Kleinrock [97] and others like Herb Grosch in the 1950s and John McCarthy in the 1960s laid the foundation of the current trends in the computing known as cloud computing by envisioning the transformation of computer usage from a Traditional in-house power generation model into a model that consists of services provided in a manner similar to utilities such as electricity, gas, and water [98],[99]. Three models have been identified for service delivery, namely, IaaS, Saas and PaaS [100]. Also, four deployment modes have been identified by scholars [101][99]. These are: Public, Private, Community and Hybrid. The definitions and explanation are given in [102], [103]. The researcher focuses on SaaS. An illustration of the request for services is depicted in Fig. 2.3 where Cloud consumers send different requests to the cloud through different applications for service provisioning. The basic characteristic of this provisioning model is that the users consume resources and are billed according to their personal demands.

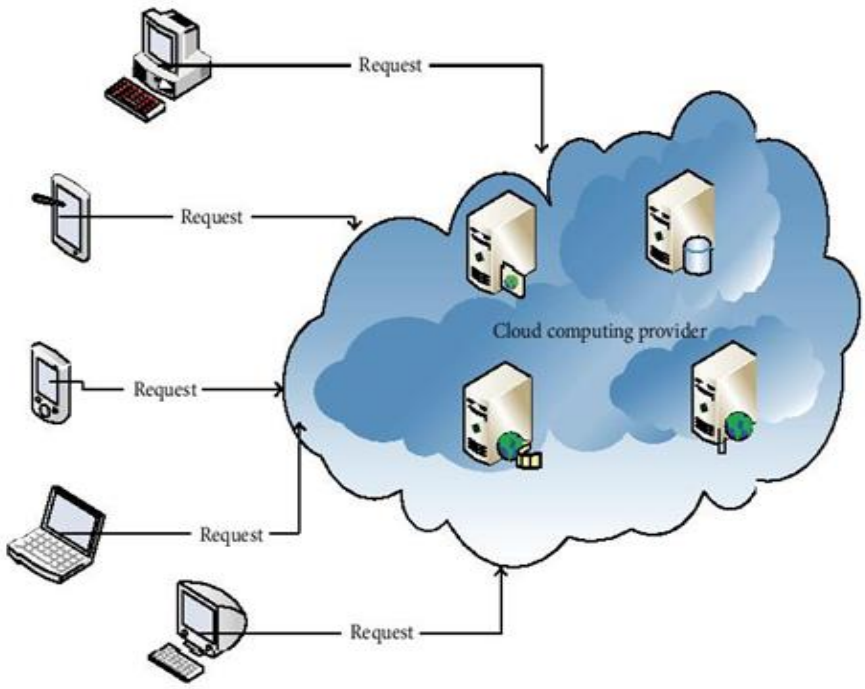

Fig. 2.3: An Illustration for Request for Services in Cloud E-Marketplaces
Source: JAM Volume(2014, Article ID 756592

The recent innovations in virtualisation and distributed computing, as well as improved access to high-speed Internet is one significant factor that has contributed to the high interest in Cloud E-Marketplaces. There are three major areas in which Cloud E-Marketplaces are different from other E-Marketplaces [23]. These are:

i. Provisioning of on demand services. This may be in minutes or hours.
ii. Elasticity that allows users to have as much or as little of a service as they want at any given time and
iii. The management of the services by Cloud providers

While the features Cost Efficiency, Almost Unlimited Storage, Backup and Recovery, Automatic Software Integration, Easy Access to Information and Quick Deployment have been the benefits of using services from the Cloud E-Marketplaces, the issues of Performance, Security and the vulnerability to

external hack attacks and threats have been the major challenge and have not been fully addressed [18][104].

In addition, the geographically distributed nature of data centres in Cloud E-Marketplaces and the architectural shift to container-based data centres have added to the challenges in the design, deployment and management of cloud computing platforms. These challenges are closely related to the performance of Cloud E-Marketplace web services. These include the distribution and migration of large volumes of data, the reduction of operational costs, the multi-dimensional allocation of available resources, accurate monitoring and prediction of service qualities, and flexible data centre network architectures. Therefore, for effective delivery of better Quality of Service (QoS) to consumers by the providers in Cloud E-Marketplaces, it is important to understand and improve the performance of cloud computing platforms, so that the performance needs of hosted applications can be satisfied. [100][105].

With the shift to the Cloud E-Marketplace paradigm scholars have concentrated on the issues of security and Privacy [3],[106], Energy [44][107] and Implementation [108]. However, performance has received less attention [100],[34]. The idea behind the Cloud E-Marketplace is that users pay for their used services without the need to spend massive amounts on integration, maintenance, or management of the IT infrastructure. With the growth in cloud E-Marketplaces there is still a shortage of performance evaluation and special measures are required to make it work, especially in the context of consumers' waiting time [100].

This research is aimed at optimising and evaluating the performance of a typical Cloud E-Marketplace based on consumers' waiting time and providers' costs.

2.3 STATE OF THE ART IN CLOUD E-MARKETPLACE PERFORMANCE

This section studies the state of art of the performance evaluation of Cloud E-Marketplaces with emphasis on consumers' waiting time (response time) and providers' costs. Much work has been done in the area of performance analysis. See [12] [34] [43] [37] [109][110]. For example, in [12], the authors use a generalised idea to address three things, namely, the level of QoS that can be guaranteed given service resources, the number of service resources that are required to ensure that customer services can be guaranteed in terms of the percentile of response time and the number of customers to be supported to ensure that customer services can be guaranteed.

Pakbaznia and Pedram [34] proposed the M/G/c to evaluate a cloud server firm with the assumption that the number of server machines is not restricted. The authors demonstrate the manner in which request response time and the number of tasks in the system may be assessed with sufficient accuracy. In Chen and Li [43], the authors model the cloud as M/M/S/k for performance management where web applications are modeled as parallel queues and the service centre as the virtual machines. This work extends that of [44], and [12] by removing the bottleneck of live migration in the packing algorithm based method.

In [37], the author uses the M[x]/G/m/m+r to describe a new approximate analytical model for performance evaluation of Cloud data centres with batch task arrivals and shows that important performance indicators such as mean request response time, waiting time in the queue, queue length, blocking probability, probability of immediate service and probability distribution of the number of tasks in the system can be obtained in a wide range of input parameters. This work was based on the so called On-Demand Service.

In [28], the authors propose a pre-emptive Cloud E-Market policy. The idea is 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. But in practice, preemption and migration of virtual machines are costly [29]. In [111], the author removes the scheduling bottleneck from one dimensional to multi-dimensional resources. This is done with the use of Multi-dimensional Resource Integrated Scheduling (MRIS) which is an inquisitive algorithm to obtain the approximate optimal solution. But [4] propose an M/M/m queuing model to develop a synthetic optimisation method to optimise the performance of services in an on Demand service. The simulation result shows that the proposed method can allow less wait time and queue length and more customers to gain the service using a synthetic optimisation function when the numbers of servers increases. In [112], the authors model the Cloud using the M/M/c/c model with different priority classes with the main goal of studying the rejection probability for different priority classes. But[113] extend Kleinrock's analysis to derive the stationary waiting distribution for each class in a single server accumulating priority queue with Poisson arrival and general distribution service time. In the opinion of the researcher, the M/M/m or the M/G/1 approach may not reflect a typical Cloud E- market because Cloud requests come through a point of entry and they go from there to various service stations for processing which then return the processed requests. Part of this research is closely related to [109][110] where these authors model the Cloud as a series of queues. What differentiates the researcher's work are:

i. Each of the service stations is modeled as M/M/c/Pr as against the M/M/1 proposed by the authors mentioned above which requires a different mathematical concept.

ii. No dedicated server is given or allocated to any class, thereby reducing consumers' waiting time.

optimisation Many researchers have worked on the issue of cost optimisation[16][28],[114][115] [116][12][31], [32]. For example, in [116][12] the

author proposes three meta-scheduling online heuristics, namely Min_Min Cost Time Trade-off (MinCTT), Sufferage Cost Time Trade-off (SuffCTT), and Max-Min Cost Time Trade-off (Max- CTT), to manage the trade-off between overall execution time and cost and to minimise them simultaneously on the basis of a tradeoff factor in the context of the utility grid. Also, [31] and [32] propose an algorithm based on convex and resource allocation optimisation methods using IaaS provisioning. But [16] use SLA-Based algorithms to extensively analys and demonstrate how to minimise the Software as a Service(SaaS) provider's cost and the number of SLA violations. The SLA-Based algorithms proposed in Buyya et al and Toosi et al [6] [15] extensively analys and demonstrate how to minimise the Software as a Service (SaaS) provider's cost and the number of SLA violations.

2.4 RESEARCH OPPORTUNITY

The work of these preceding authors presented the researcher with the opportunity to make his own contributions. For example, the argument for using the queuing model is based on that of [12][43] [10]. The works of [109][44][37][4] [28] are the fore-runners of the idea of viewing the Cloud as networks of queues. For example, while the authors of [12] [10], generalised the Cloud like any other Traditional system, the approach looks at it rather as networks of queues. This is because viewing the cloud as a single queue may not reflect a typical Cloud E-Market. The claimed response time could not represent the real cloud E-Market response time. Apart from the aforementioned facts that differentiate this work from [109][110], the work reported in [13] is based only on simulation and extend the idea of these scholars using the queuing theory. the reason for this is that the queuing models predict the theoretical performance behaviours of systems that attempt to provide services for randomly arriving demands [47]. Therefore, ascertaining the degree of correctness of any simulation requires theoretical proof to back it up.

In addition, the issue of considering only On Demand Service by [37] may not reflect the true picture of today's Cloud. This is because most providers are offering different services based on consumer demand. For example, Amazon.com offers three services: the On Demand, Spot, and Reserved [20]. This research considers both On Demand and others with the use of Non Priority and Non Pre-emptive Priority policies, which further differentiates the researcher's work from these authors.

Other related works, for example [113] [4][25] implement their Non Pre-emptive policy at the first point of entry alone, whereas the researcher model the Non Pre-emptive model at every point of queue. This is because at every point of queue there is a likely tendency that higher priority will arrive when lower one is on the queue.

Furthermore, on the issue of cost optimisation algorithms proposed by [31] and [32], the maximisation objective was subjected to many constraints which may become complex to understand. Second, the concept of cost model is based on operating cost rather than on both operational costs and fixed costs. Part of this research takes into consideration both operating costs and fixed costs to determine SaaS provisioning using queuing theory.

Finally, the significance of a dynamic optimisation control mechanism in effective server management further differentiates this work from others. In summary, this research successfully explored the application of existing and widely adopted theories of the Non Preemptive queue model to the design of a dynamic optimisation control mechanism for effective server management in the context of Cloud E-Marketplaces. All these to the best of researcher's knowledge have not been reflected in the literature in the context of Cloud E-Marketplaces. Throughout this thesis the definitions in Table 2-2 hold, as explained in [31][32][117][47] .

Table 2-2: Meanings and Definitions of queuing theory terms

Name	Meaning
Consumers:	An application requesting service from the provider
M/M/1/k	A Queue system in which consumers arrive at random rate, exponential service rate, one server and having limited buffer size with FCFS discipline
M/M/c/k	A Queue system in which consumers arrive at random rate, exponential service rate, more than one server and having limited buffer size with FCFS discipline
M/M/1/Pr	A Queue system in which consumers arrive at random rate, exponential service rate, one server and having unlimited buffer size based on Priority
M/M/c/Pr	A Queue system in which consumers arrive at random rate, exponential service rate, more than one server and having unlimited buffer size based on Priority

The steady measure performance (measure of effectiveness) shown in Table 2-3 below will be used in some of the chapters in this thesis.

Table 2-3: Some Performance Measures and their meanings

ρ	= Server utilisation
P_0	= Probability of zero consumer in the system
P_n	= Probability of n consumer in the system
L_s	= Expected number of Consumers in the system
L_q	= Expected number of Consumers in the queue
W_s	= Expected waiting time in the system
W_q	= Expected waiting time in the queue
c	= Expected number of busy servers

2.5 CHAPTER SUMMARY

This chapter has presented the trends in Marketplaces, which form the background to this research. This researcher has investigated the Traditional, Internet, Web service, Grid and the Cloud E-Marketplaces. The current state of the art in Cloud E-Marketplaces has been studied and this has given the researcher the chance to contribute based on the opportunity derived from the solid foundations already laid in some of these works by scholars such as [12][43] and [10].

CHAPTER THREE

PERFORMANCE MODELLING OF CLOUD E-MARKETPLACES BASED ON NON PRIORITY FOR COST MINIMISATION

In this chapter, the researcher models of the cloud E-Marketplace under the non- priority system and the performance impact are evaluated based on the consumers' waiting time. A systematic cost function was formulated based on operational cost and service cost. The result obtained was used as part of the parameters in the cost function to be minimised to get the optimal result. Sometimes it is very difficult or almost impossible to get an optimal solution because of the difficult task involved in determining the cost of waiting. In order to overcome this work explored the use of the Aspiration model. This was used to achieve an acceptable range using waiting time and percentage of server idleness as the conflicting effective measures. These, to the best of the researcher's knowledge have not appeared in the literature in the field of Cloud E-Marketplaces. The adopted solution approach used the M/M/1/K and M/M/c/K queuing models.

3.1 INTRODUCTION

Cloud E-Marketplaces are becoming perfect competitive markets. These competitive markets consist of two major participants, the cloud E-Market consumers and the cloud E-Market providers. Though the markets are virtualised, both share the same simple idea of exchanging goods for services. The researcher assume goods in this context are the costs, in terms of waiting time to access a server, paid by the consumers or clients. In these markets, providers must decide what level of service to offer. A low level of service may be inexpensive, at least in the short run, but may incur high costs of consumers' dissatisfaction, such as loss of future business and actual processing costs of complaints. A high level of service will cost more to the E-Cloud provider but will

result in lower dissatisfaction costs. Though some scholars have worked in this area, for example , [12] used a generalised approach in a single queue form, viewing the cloud as a single queue may not reflect a typical cloud E-Market. The acclaimed response time was not able also represent the real cloud E-Market response time. Apart from the aforementioned facts that differentiate this work from [109][110], the work reported in [13] is based only on simulation and extend the idea of these scholars using the queuing theory.

Two things will be addressed in this chapter, the first being the optimal level of service to be provided by a cloud provider that will minimise cost while at the same time satisfying consumers in terms of waiting time. The second is the aspiration level required to satisfy the consumer and provider in an environment where the waiting cost is difficult to measure or more parameters are required. To achieve this, the use of a queuing system is applied to get the performance measure. The cost and aspiration models are then designed using the performance results.

The remainder of this chapter is organised as follows: Section 3.2 overviews the proposed work and discusses the mathematical MODELLING. In section 3.3, the cost structure is described. The first simulation is discussed in section 3.4 with the results and discussion in section 3.5. The Aspiration model is the second part of the study. This is covered in section 3.6 with the simulation, result and discussion of the experiment in section 3.7 and 3.8 respectively. The chapter closes with a summary.

3.2 PROPOSED NON- PRIORITY MODEL

The model is divided into two sub models. These are sub_model 1 and 2 respectively. Sub_model 1 consists of the incoming web or consumer applications with dispatcher and database feedback, and the sub model 2 consists of three service stations that are networked together. The processing of

the applications takes place at these service stations. The model is represented by the diagram in Fig. 3.1.

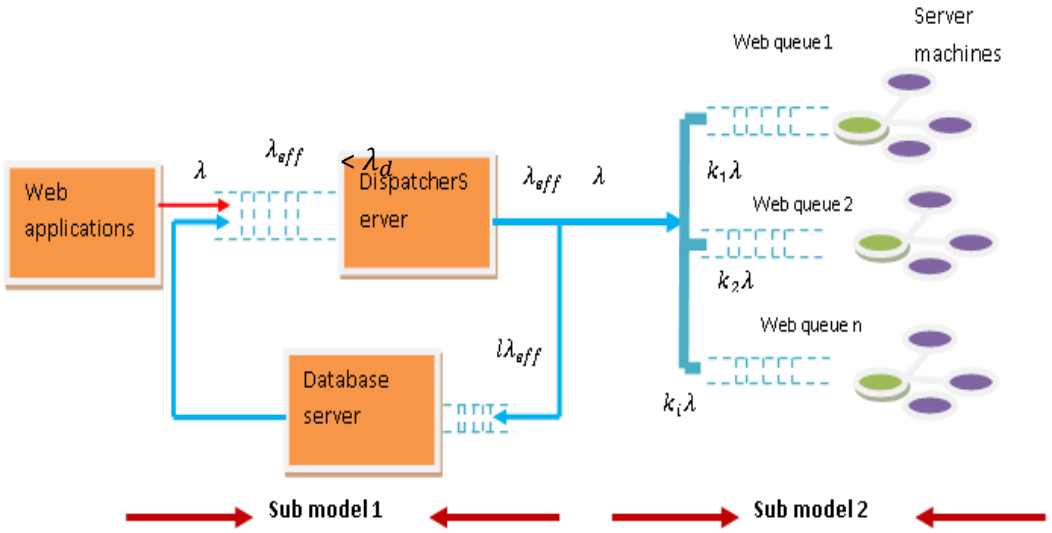

Fig. 3.1: Non-Priority E-cloud Marketplace Model

The dispatcher-In receives all incoming requests λ_d from both the consumers (λ) and the database feedback $(l\lambda_{eff})$ which are then scheduled to the web queue server in the web queue stations.

The number of requests λ_d coming to the dispatcher is derived by adding λ and $l\lambda_{eff}$ ($\lambda_d = \lambda + l\lambda_{eff}$). This is due to the fact that not all consumers can join the dispatcher-In as a result of the limited server capacity (M/M/1/k) of the researcher's model. Consequently, the real effective arrival rate as shown in the proposed model is defined as λ_{eff} (where $\lambda_{\text{eff}} < \lambda_d$). Therefore, $\lambda + l\lambda_{eff} = \lambda_{\text{eff}}$. λ_d is first applied in the mathematical formulation and later revert to λ_{eff} .

The web queue servers act as the real processors that provide the service based on Non-Priority First Come First Served (FCFS). Each of the web queue stations has c identical parallel servers (c= 1,2,…c) with equal probability distribution $k_i\lambda$ of requests to each web queue station. Where $k_1 = k_2 = k_i$. Also, arrival and

the service rate of the requests follow a Poisson process. One other assumption in this chapter is in line with [109] , that the latency of internal communication between the Dispatcher server, database server and the web queue service stations is insignificant. The general idea is to derive $\mathrm{P}(i)_{\mathrm{n}}$ as a function of $\lambda(i)_{\mathrm{n}}$, $\rho(i)$ and $\mu(i)_{\mathrm{n}}$ where *i = 1,2,3*= disp (Dispatcher-In), dbase, and each of the service stations respectively. For example, $P(disp)_{\mathrm{n}}$ represents the steady state probability of *n* consumers in the dispatcher-In queue while $\lambda(\text{web queue } 1)_{\mathrm{n}}$ represents the number of consumers arriving at the web queue station 1. Also, $\mu(disp)_{\mathrm{n}}$ represents the departure or service rate in the dispatcher server given *n* numbers of consumers in the system and $\rho(i)$ is the server utilisation of the i^{th} server machine. These probabilities are then used to determine the conflicting measures of performance which are the average queue length, average waiting time and the average utilisation of the facility.

The researcher models the dispatcher-In and the database servers as M/M/1/k queue respectively and that of web queue stations as M/M/c/k queue. The conflicting measure of performance is derived using the six steps stated in [118] and the law of conservation of flow in [47][14]. The dispatcher-In and the database are first modeled followed by the web queue stations.

3.2.1 MODELLING THE DISPATCHER-In USING M/M/1/K

The dispatcher-In is modeled as M/M/1/K. The justification for using the finite buffer (k) is that there is always a limit to what the server can contain [119]. The server utilisation ρ_1(for dispatcher-In) and ρ_2(database) of the two servers are given as

$$\rho_1 = \frac{\lambda_d}{\mu_1} \text{ and } \rho_2 = \frac{\lambda_d}{\mu_2} \tag{3.1}$$

The assumptions for this work are follows:

1. $\lambda_d \leq \mu_1$ and $l\lambda_d \leq \mu_2$ where μ_1 *and* μ_2 are the dispatcher and database service rate respectively.
2. Expected rate of flow into a state = Expected rate of flow out of that state in line with [118][120].

This is in contrast to some authors for example [11] which assumes $\lambda_{eff} < \mu_1$ for steady state condition. The researcher first model the dispatcher queue as

$$(\ \boldsymbol{\lambda_d + \mu_1})\ \boldsymbol{P_n} = \ \boldsymbol{\lambda_d P_{n-1} + \mu_1\ P_{n+1}} \qquad (\mathbf{3.2})$$

and for the database server, it is given as

$$\boldsymbol{l\lambda_d + \mu_2})\ P_n = \ l\ \lambda_d P_{n-1} + \mu_2\ P_{n+1} \qquad (\mathbf{3.3})$$

Therefore the probability of having one (n=1) consumers in the dispatcher P(disp) and database P(dbase) servers are

$$\mathbf{P(disp)_1} = \frac{\boldsymbol{\lambda_d}}{\boldsymbol{\mu_1}}\ \mathbf{P_0}\ \mathbf{and}\ \ \mathbf{P(dbase)_1} = \ \frac{\boldsymbol{l\lambda_d}}{\boldsymbol{\mu_2}}\ \boldsymbol{P_0} \qquad (\mathbf{3.4})$$

and

$$\boldsymbol{P}(\mathbf{disp})_n = \ \left(\frac{\boldsymbol{\lambda_d}}{\boldsymbol{\mu_1}}\right)^{\boldsymbol{n}} \boldsymbol{P_0}\ \mathbf{and}\ \mathbf{P(dbase)}_n = \ \left(\frac{\boldsymbol{l\lambda_d}}{\boldsymbol{\mu_1}}\right)^{\boldsymbol{n}} \boldsymbol{P_0} \qquad (\mathbf{3.5})$$

$$\boldsymbol{P}(\mathbf{disp})_n = \ \ \boldsymbol{\rho_1}^{\boldsymbol{n}}\boldsymbol{P_0}\ \ \mathbf{and}\ \mathbf{P(dbase)}_n = \ \ \boldsymbol{\rho_2}^{\boldsymbol{n}}\boldsymbol{P_0} \qquad (\mathbf{3.6})$$

since the total probability = 1, then

$$\sum_{i=0}^{N} P_i \ = \sum_{i=0}^{N} \ \ {\rho_1}^n P_0 \ \ = {\rho_1}^n \left[\frac{1-{\rho_1}^{N+1}}{1-\rho_1}\right]^{-1} = 1 \qquad (3.7)$$

and for the database server it is given as

$$\rho_2{}^n \left[\frac{1-\rho_2{}^{N+1}}{1-\rho_2}\right]^{-1} = 1 \qquad (3.8)$$

Eq. 7 and 8 hold when $\lambda_d = \mu_1$ and $l\ \lambda_d = \mu_2$. but when $\lambda_d \neq \mu_1$ and $l\ \lambda_d \neq \mu_2$ then

$$P(disp)_0 = \lim_{\rho_1 \to 1} \left[\frac{1-{\rho_1}^{N+1}}{1-\rho_1}\right]^{-1} \qquad (3.9)$$

Using L'Hospital's rule [118] it follows that

$$P(disp)_0 = \lim_{\rho_1 \to 1} \left[\frac{-(N+1){\rho_1}^N}{-1}\right]^{-1} = \left[\frac{N+1}{1}\right]^{-1} \qquad (3.10)$$

$$P(dbase)_0 = \lim_{\rho_2 \to 1} \left[\frac{-(N+1){\rho_1}^N}{-1} \right]^{-1} \tag{3.11}$$

Combining the two situations when $\lambda_d = \mu_1$, $l\,\lambda_d = \mu_2$ and when $\lambda_d \neq \mu_1$ and $l\,\lambda_d \neq \mu_2$ for the dispatcher and the database servers then

$$P(disp)_0 = \begin{cases} \left[\frac{1-{\rho_1}^{N+1}}{1-\rho_1} \right]^{-1} & if \rho_1 < 1 \; or \;\; \lambda_d \neq \mu_1 \\ \left[\frac{N+1}{1} \right]^{-1} & if \; \rho_1 = 1 \; or \; \lambda_d = \mu_1 \end{cases} \tag{3.12}$$

and

$$P(dbase)_0 = \begin{cases} \left[\frac{1-{\rho_2}^{N+1}}{1-\rho_2} \right]^{-1} & if \rho_2 < 1 \; or \;\; l\lambda_d \neq \mu_1 \\ \left[\frac{N+1}{1} \right]^{-1} & if \; \rho_2 = 1 \; or \;\; l\lambda_d = \mu_1 \end{cases} \tag{3.13}$$

and

$$P(disp)_n = \begin{cases} \left[\frac{1-{\rho_1}^{N+1}}{{\rho_1}^n(1-\rho_1)} \right]^{-1} & if \rho_1 < 1 \; or \;\; \lambda_d \neq \mu_1 \\ \left[\frac{N+1}{1} \right]^{-1} & if \; \rho_1 = 1 \; or \; \lambda_d = \mu_1 \end{cases} \tag{3.14}$$

and

$$P(dbase)_n = \begin{cases} \left[\frac{1-{\rho_2}^{N+1}}{{\rho_2}^n(1-\rho_2)} \right] & if \rho_2 < 1 \; or \;\; l\lambda_d \neq \mu_1 \\ \left[\frac{N+1}{1} \right]^{-1} & if \; \rho_2 = 1 \; or \;\; l\lambda_d = \mu_1 \end{cases} \tag{3.15}$$

This implies that for all values of n, n = 0,1,2,3,….,N

$$P(disp)_n = \begin{cases} \left[\frac{1-{\rho_1}^{N+1}}{1-\rho_1} \right] \rho^n & if \rho_1 < 1 \\ \left[\frac{N+1}{1} \right]^{-1} & if \; \rho_1 = 1 \end{cases} \tag{3.16}$$

and

$$P(dbase)_n = \begin{cases} \left[\frac{1-{\rho_2}^{N+1}}{1-\rho_2} \right] \rho^n & if \rho_2 < 1 \\ \left[\frac{N+1}{1} \right]^{-1} & if \; \rho_2 = 1 \end{cases} \tag{3.17}$$

In this work, $\rho_{1,}$and ρ_2 are less than 1. Therefore, the expected number of consumers in dispatcher-In (E(web_{disp})) and database ($\mathrm{E}(\mathrm{webdbase})$) system are:

E(web_{disp}) = $\sum_{n=0}^{N} nP_n$ = $\sum_{n=0}^{N} n{\rho_1}^n P_0$

$$= \left[\frac{1-{\rho_1}^{N+1}}{1-\rho_1}\right]^{-1} \rho_1 \left[\frac{(1+{\rho_1}^{N+1})-(N+1){\rho_1}^N\ (1-\rho_1)}{[1-\rho_1]^2}\right] \quad (3.18)$$

$$\mathrm{E}(\mathrm{webdbase}) = \left[\frac{1-{\rho_2}^{N+1}}{1-\rho_2}\right]^{-1} \rho \left[\frac{(1+{\rho_2}^{N+1})-(N+1){\rho_2}^N\ (1-\rho_2)}{[1-\rho_2]^2}\right] \quad (3.19)$$

There are two things that the researcher has done in the re-engineering process. The first is the modification of Little's formula to determine the expected number of web applications in the dispatcher and database queues. This is because the expected number of the web applications/consumers in dispatcher queue for example

$\mathrm{E}(\mathrm{disp.queue}) = \sum_{n=0}^{N}(n-1)P_n$

$$= \sum_{n=0}^{N} nP_n - \sum_{n=0}^{N} P_n = \mathrm{E}(web_{disp})\text{-}(1\text{-}P_0) \quad (3.20)$$

but using Little's formula then $\mathbf{E(disp.queue)} = \mathrm{E}(web_{disp})$ - $\frac{\boldsymbol{\lambda_d}}{\boldsymbol{\mu_1}}$

This is only true when the mean arrival rate is $\boldsymbol{\lambda_d}$ as assumed by [109] [12]. However,

1-$\boldsymbol{P_0} < \frac{\boldsymbol{\lambda_{eff}}}{\boldsymbol{\mu_1}}$. This because the mean arrival rate is $\boldsymbol{\lambda_d}$ when there is vacancy in the queue and zero when the system is full. This gives us the motivation to define the real effective arrival rate as $\boldsymbol{\lambda_{eff}}$. Therefore applying Eq.18 and Little's formula as $\frac{\boldsymbol{\lambda_{eff}}}{\boldsymbol{\mu_1}}$ =1-$\boldsymbol{P_0}$ or $\boldsymbol{\lambda_{eff}} = \boldsymbol{\mu_1}$(1-$\boldsymbol{P_0}$).

Thus, this can be written Eq. 18 as

$$\mathbf{E(disp.queue)} = \mathbf{E(webdisp)} - \frac{\boldsymbol{\lambda_{eff}}}{\boldsymbol{\mu_1}} \quad (\mathbf{3.21})$$

This also applies to the database queue which is then written as

$$\mathrm{E(dbase.queue} = \mathrm{E(webdbase)} -) \frac{l\,\lambda_{\mathrm{eff}}}{\mu_2} \tag{3.22}$$

The second issue in the re-engineering process is the calculation of the average waiting time in both the queue and system of the dispatcher and the database where most authors like [11] multiply ${\lambda_{\mathrm{eff}}}^{-1}\ by\ \mathrm{E(webdisp)}$ as the waiting time in the dispatcher system or ${\lambda_{\mathrm{eff}}}'^{-1}\ by\ \mathrm{E(disp.queue)}$ as the waiting time in the dispatcher queue.

The waiting time both in dispatcher system ($\mathrm{Ws_{disp}}$) and the queue ($\mathrm{Wq_{disp}}$) are represented as

$$\mathrm{Ws_{disp}} = \frac{\mathrm{E(LS_{disp})}}{\lambda_{\mathrm{eff}}} * \mathrm{Ex_{visitdisp}} \tag{3.23}$$

$$\mathrm{Wq_{disp}} = \left(\mathrm{Ws_{disp}} - \frac{1}{\mu_1}\right) * \mathrm{Ex_{visitdisp}} \tag{3.24}$$

$$\mathrm{Wq_{dbase}} = \left(\mathrm{Ws_{dibase}} - \frac{1}{\mu_2}\right) * \mathrm{Ex_{visitdbase}} \tag{3.25}$$

$$\mathrm{Ws_{dbase}} = \frac{\mathrm{E(LS_{disp})}}{\lambda_{\mathrm{eff}}} \mathrm{Ex_{visitdbase}} \tag{3.26}$$

Where $\mathrm{Ex_{visitdisp}}$ represents the number of visit(s) to the dispatcher-In which is given as

$\mathrm{Ex_{visit\ disp}} = \frac{1}{1-\lambda_{\mathrm{eff}}'}$ and that of the database as $\mathrm{Ex_{visit\ dbase}} = \frac{1}{1-\lambda_{\mathrm{eff}}}$

3.2.2 MODELLING THE WEB SERVICE STATIONS

As earlier mentioned, each web queue station is modeled as M/M/c/k with equal service distribution $k_i\lambda_d$ as shown in Fig. 3.1 where i= 1,2,3,….j represents the number of web queue service stations and each station has equal or identical servers (c) with the same service rate μ. For example, web queue service station 1 whose arrival rate is $k_i\,\lambda_d$ with c servers has a total service

rate of 3μ. Therefore, for each web queue service station, the mean arrival rate is given by

$$k_i\lambda_{\mathbf{eff_n}} = \begin{cases} \mathbf{k_i}\ \lambda_d & \mathbf{for\ n = 0, 1, 2 \ldots . N - 1} \\ \mathbf{0} & \mathbf{for\ n = N, N + 1, \ldots \ldots} \end{cases} \quad (\mathbf{3.27})$$

and

$$\boldsymbol{\mu_n} = \begin{cases} \mathbf{n\mu} & \mathbf{for\ n = 0, 1, 2 \ldots . c - 1} \\ \mathbf{c\mu} & \mathbf{for\ n = c, c + 1, \ldots \ldots} \end{cases} \quad (\mathbf{3.28})$$

where $\mathbf{1} < \boldsymbol{c} < \boldsymbol{N}$

Given the steady- state probabilities $\boldsymbol{P_n}$ and $\boldsymbol{P_0}$, then

$$\boldsymbol{P_n} = \frac{\mathbf{k_i}\ \lambda_{d0}\ \mathbf{k_i}\ \lambda_{d1}, \ldots \ldots, \mathbf{k_i}\ \lambda_{dn-1}}{\mu_1\mu_2 \ldots \ldots \ldots \ldots \ldots . \mu_n} \boldsymbol{P_0} \quad (\mathbf{3.29})$$

$$\boldsymbol{P_0}^{-1} = 1 + \sum_{n=1}^{\infty} \left[\frac{\mathbf{k_i}\ \lambda_{d0}\ \lambda_{d1} \ldots . \mathbf{k_i}\ \lambda_{dn-1}}{\mu_1\mu_2 , \ldots \ldots \ldots \ldots , \mu_n} \right] \quad (\mathbf{3.30})$$

substituting the value $\boldsymbol{k_i\lambda_d}$ and $\boldsymbol{\mu_n}$

$$\boldsymbol{P_0} = \left[\sum_{\mathbf{n=0}}^{\mathbf{c-1}} \frac{\mathbf{1}}{\mathbf{n!}} \left(\frac{\mathbf{k_i}\ \lambda_\mathbf{d}}{\boldsymbol{\mu}} \right)^{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{c!}} \left(\frac{\mathbf{k_i}\ \lambda_\mathbf{d}}{\boldsymbol{\mu}} \right)^{\mathbf{c}} \sum_{\mathbf{n=c}}^{\mathbf{k}} \left(\frac{\mathbf{k_i}\ \lambda_\mathbf{d}}{\boldsymbol{\mu}} \right)^{\mathbf{n-c}} \right]^{-1} \quad (\mathbf{3.31})$$

and

$$\mathbf{P_n} = \begin{cases} \frac{\mathbf{1}}{\mathbf{n!}} \left(\frac{\mathbf{k_i}\ \lambda_\mathbf{d}}{\boldsymbol{\mu}} \right)^{\mathbf{n}} \mathbf{P_0} & \mathbf{for\ n \leq c} \\ \frac{\mathbf{1}}{\mathbf{c!c^{n-c}}} \left(\frac{\mathbf{k_i}\ \lambda_\mathbf{d}}{\boldsymbol{\mu}} \right)^{\mathbf{n}} \mathbf{P_0}\ \mathbf{for\ c < n \leq k} \\ \mathbf{0} \qquad\qquad \mathbf{for\ n \leq k} \end{cases} \quad (\mathbf{3.32})$$

Therefore the expected number of consumers in the queue of each service station i is given by

$$\boldsymbol{E}\ (\boldsymbol{web\ queue_i}) = \sum_{n-c}^{N} \boldsymbol{n - c} \frac{\mathbf{1}}{\mathbf{c!c^{n-c}}} \left(\frac{\mathbf{k_i}\ \lambda_d}{\boldsymbol{\mu}} \right)^{\boldsymbol{n}} \boldsymbol{P_0} \quad (\mathbf{3.33})$$

but the server utilisation in each web queue service station i is $\boldsymbol{\rho_i} = \frac{\mathbf{k_i}\boldsymbol{\lambda_d}}{c\boldsymbol{\mu_1}}$.

Substituting $\boldsymbol{\rho_i}$ in Eq. 3.31 and differentiating $\frac{d}{d\rho_i}\left[\frac{1-\rho_i{}^{N-c+1}}{1-\rho_i}\right]$ **then**

$$\boldsymbol{E}\ (\boldsymbol{web\ queue_i}) =$$

$$\boldsymbol{P_0}\frac{\mathbf{k_i}\boldsymbol{\lambda_d}}{\boldsymbol{\mu}}\frac{\boldsymbol{\rho_i}}{\boldsymbol{c!}(\mathbf{1}-\boldsymbol{\rho_i})^2}[\mathbf{1}-\boldsymbol{\rho_i}{}^{N-c}-(\boldsymbol{N}-\boldsymbol{c})(\mathbf{1}-\boldsymbol{\rho_i})\boldsymbol{\rho_i}{}^{N-c}]\ \ (\mathbf{3.34})$$

The expected number of web applications in the system is given as

$$\boldsymbol{E}\ (\mathbf{web\ system_i}) = \sum_{n=0}^{c-1}\boldsymbol{nP_n} + \sum_{n=c}^{N}\boldsymbol{nP_n} \qquad (\mathbf{3.33})$$

Therefore, the modified Little's formula then becomes

$$\boldsymbol{E}\ (\boldsymbol{web\ system_i}) = \boldsymbol{E}\ (\boldsymbol{web\ queue_i}) + \frac{\mathbf{k_i}\boldsymbol{\lambda_{eff}}}{\boldsymbol{\mu}} \qquad (\mathbf{3.35})$$

Where $\mathbf{k_i}\boldsymbol{\lambda_{eff}}$ is the real effective arrival rate given as

$$\mathbf{k_i}\boldsymbol{\lambda_{eff}} = \boldsymbol{\mu}\left[\boldsymbol{c} - \sum_{n=0}^{c-1}(\boldsymbol{c}-\boldsymbol{n})\boldsymbol{P_n}\right]$$

The web system and queue waiting time are:

$$\mathbf{Wsystem_i} = \left\lfloor \mathbf{k_i}\boldsymbol{\lambda_{eff}}\right\rfloor^{-1} * \boldsymbol{E}(\mathbf{web\ system_i}) \qquad (\mathbf{3.36})$$

$$\mathbf{Wqueue_i} = \left\lfloor \mathbf{k_i}\boldsymbol{\lambda_{eff}}\right\rfloor^{-1} * \boldsymbol{E}\ (\mathbf{web\ queue_i}) \qquad (\mathbf{3}.37)$$

The average mean waiting time in the queue and system of all the web queue service stations are given as

$$\mathbf{Wqueue_{ave}} = \frac{\mathbf{1}}{\mathbf{j}}\sum_{i=0}^{j}\mathbf{Wqueu_i} \qquad (\mathbf{3.38})$$

$$\mathbf{Wsystem_{ave}} = \frac{\mathbf{1}}{\mathbf{j}}\sum_{i=0}^{j}\mathbf{Wsystem_i} \qquad (\mathbf{3.39})$$

Therefore, the total queue waiting time in all the service stations is given as

$$\mathbf{Ws\_total}\ = \mathbf{Wq_{disp}} + \mathbf{Wq_{dbase}} + \mathbf{Wqueue_{ave}} \qquad (\mathbf{3.40})$$

3.3 COST MODEL FORMULATION

The cost function uses the waiting time result derived in eq. 3.40. This is defined based on the works of the scholars in [121][122][120][123] as

$$\mathrm{ETC(x)} = \text{Variable Cost} + \text{Fixed Cost} = \mathrm{EWC(x)} + \mathrm{EOC(x)} \qquad (3.41)$$

$$\mathrm{ETC(x)} = \mathrm{k} * \mathrm{Ws\_total}\ +\ (\mathrm{c}/10 + 1) \qquad (3.42)$$

where ETC(x) is equal to the expected total cost per unit time and EOC(x) is the expected cost of operating the cloud E-Market servers per unit time and EWC(x) is the expected cost of waiting by web application per unit time. k = Cost value of waiting in the queue which is $5 in this experiment. $\mathrm{EOC(x)} = (\mathrm{c}/10) + 1)$ and c is the number of server machines working. A total of $1 is assigned as the operating cost of energy and $(\mathrm{c}/10)$ as servicing cost.

Knowing the upper bound of the experiment to be N. The generalised algorithm for optimal service level (optslev) is given in Figure 3.2. Where $\mathrm{ETC(x)_i}$ represents the quantity (x) of servers used in the *ith* experiment to calculate the Expected Total Cost and $\mathrm{sm(x)_i}$ represents the quantity (x) of server machines used in the *ith* experiment.

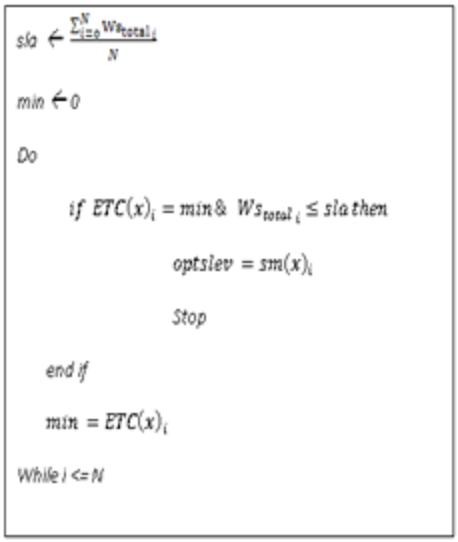

Fig. 3.2: Optimal Service Level Algorithm

3.4 SIMULATION I

The simulation was performed using Arena v 14.5. The process is carried out by setting the inter arrival time to .33 seconds, and the service time for the dispatcher and for the web queue server to 1.2 seconds. The buffer capacity of 1000 is used for the dispatcher to reduce balking. In the other servers 400 was used as the maximum buffer capacity and allocated $5 to cost of waiting. The

base time unit was set to minutes. Each experiment was conducted with 10 replications for an average of 49949,0000 seconds. The experiment started with a total of six server machines with each application queue having two server machines. At the end of each experiment the server machines was increased in each web application queue by one while the record of the waiting time in the system was kept. This performance measure was then used to as part of the cost function.

3.4.1 NUMERICAL VALIDATION AND SIMULATION

The researcher first validated the mathematical solution with the simulation to ascertain the degree of correction. This is done by setting the simulation and the analytical parameter to the same value. That is, $c = 4, 6, 8, 10, \ldots, 20$ respectively in each of the service stations and λ to a constant value. Wolfram Mathematical 9.0 is used as the mathematical tool for the validation of the results. This simulation was run with replication length of 1000 in 24 hours per day with base time in hours and replicated 5 times. The service rate was set to 0.001 for the dispatcher-In and 0.0005 for each of the servers in the web queue stations and the dispatcher-Out. A Server of low service rate of .0002 was used for the database server because of its randomness. The results and the explanation are given under the results and discussion section in section 3.5 of this work.

3.5 RESULTS AND DISCUSSION I

The comparison of the analytical and the simulation results to ascertain the degree of correctness is first presented. The degree of variation is not significant enough to be noticed in neither the analytical nor the simulation when 5 to 10 server machines were used. However, a noticeable variance of less than unity is observed after ten machines, giving the assurance that both the Analytical and the simulation results converge as depicted in Fig. 3.3

In the experiment a total of 150,000 consumers (web applications) arrived and these were processed (Tot Web In/Out) by web queues 1 and 2 servers. The database server automatically generated 5985.50 requests for collecting statistical data, demonstrating the re-engineered aspect of the model as shown in Fig. 3.4. This agrees with the model in Fig. 3.1. The result in Table 3.1 shows the mean waiting time in the system and the total server machines used. The ws_{total} in Table 3.1 was used to get the Expected Total Cost (ETC). SLA was benchmarked as $\sum_{i=0}^{N} Ws_{total\,i}/N$. From Table 3.1 it was observed that as the server machine increased the waiting time reduced while the SLA remained constant, as shown in Fig. 3.5. The waiting time intercepted the SLA time on the 11-12 server machines interval. Any point above this interval represented a breach of agreement and any point below would be to the advantage of the consumers as service is delivered at a lower waiting time but with the provider having to operate at an unnecessarily higher cost. This particular situation is expressed in Fig. 3.6 showing the ETC and the service level results. A higher service level is achievable but at the expense of increasing total cost. The minimum total cost was incurred when 18 server machines were used. This number represents the optimal service level required to minimise cost in this context. However at this level, the researcher needs to know how long the consumer has to wait; is the waiting time above or below the service level agreement? The answer to this question is found in Fig. 3 7. The answer was that using 18 server machines brought down the waiting Time by 0.1495 (1.706 – 1.5565) seconds, with the added advantage that it was achieved at a cost that is optimal for the cloud E-Market provider. It will be recalled that the service level agreement time was already achieved at 10-12 server machine costs to the provider. However, the consumer is happier that service is delivered faster than the SLA time. How this is achieved is the next subject of discussion.

The overall analysis is given in Table 3.2. The analysis shows that using between 6-9 server machines will reduce cost but will be a breach of the service level

agreement; this is the case because service was delivered at a consumer waiting time of between 0.2087 and 1.0404 longer than the SLA time. The next possibility is 12-15 servers. Although this is not a breach of SLA and the consumer is not worse off in terms of waiting time the provider is not operating at the minimum optimal Expected Total Cost, ETC. In other words, the consumer waiting time could still have been reduced further without raising the ETC to the provider. Using 18 server machines yielded the optimal cost and reduced the waiting time by 0.1495 seconds below the SLA time; and a much better gain in consumer waiting time is achieved. This means that before the optimal level was reached the provider was less competitive than others and it was just a matter of time before they started to lose knowledgeable customers.

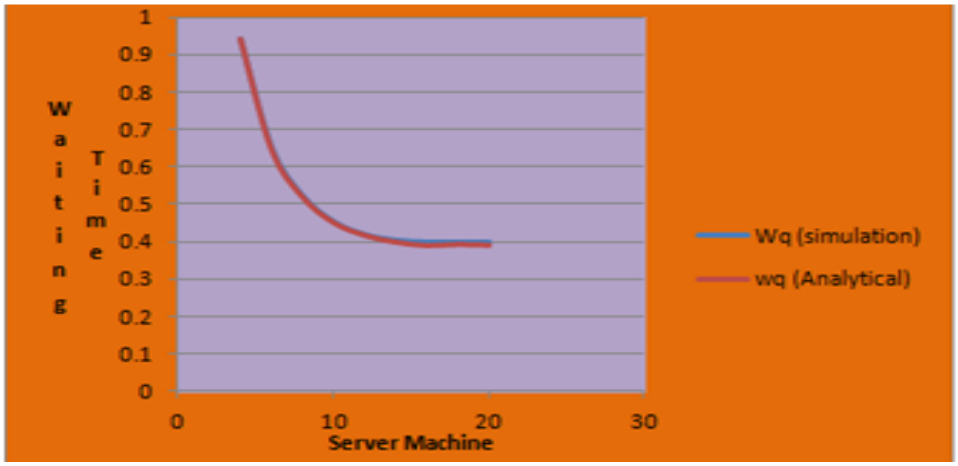

Fig. 3.3: Analytical and Simulation

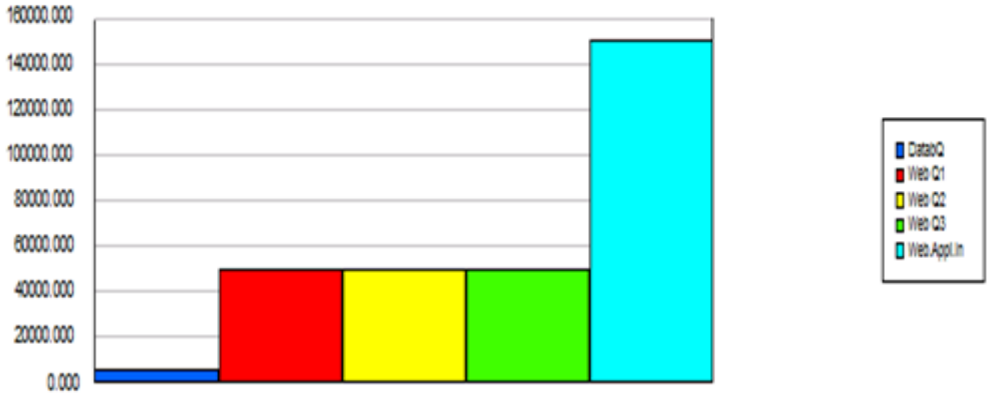

Fig. 3.4: Input and Output of Web Applications

Table 3-1: SM-Waiting Time-ETC

Experiment	Server machine	Average waiting Time in secs.(P)		Cost of waiting	Operating Cost
(i)	(Q)		ETC(x)	EWC(x)	EOP(x)
1	6	2.7467	15.33	13.73	1.6
2	9	1.9147	11.47	9.57	1.9
3	12	1.6963	10.68	8.48	2.2
4	15	1.6249	10.62	8.12	2.5
5	18	1.5565	10.58	7.78	2.8
6	21	1.5083	10.64	7.54	3.1
7	24	1.4948	10.87	7.47	3.4
8	27	1.4148	10.77	7.07	3.7
9	30	1.4003	11	7	4
	SLA	1.706			

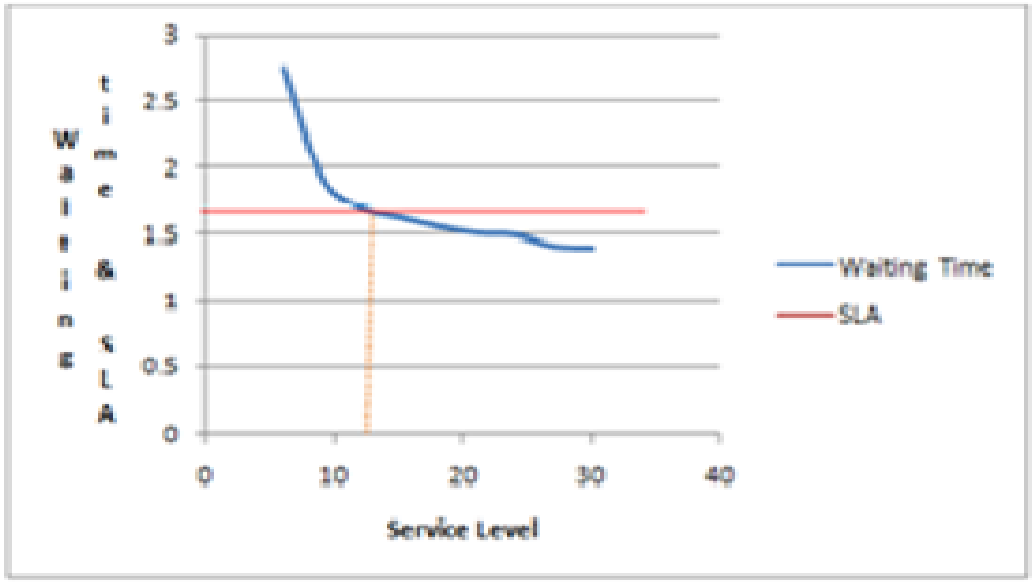

Fig. 3.5: SLA- Waiting Time

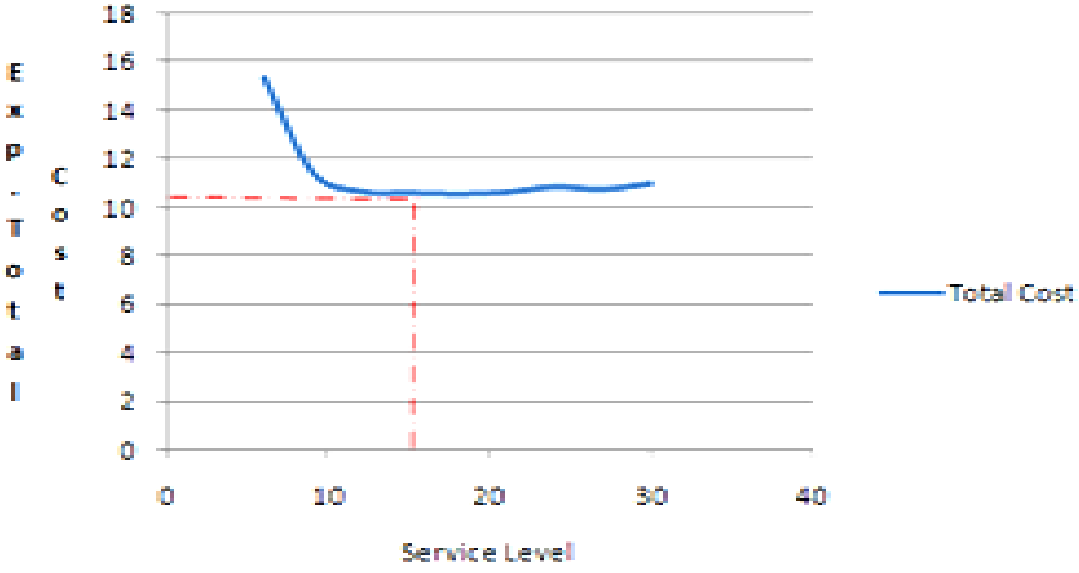

Fig. 3.6: Expected Total Cost: Service Level

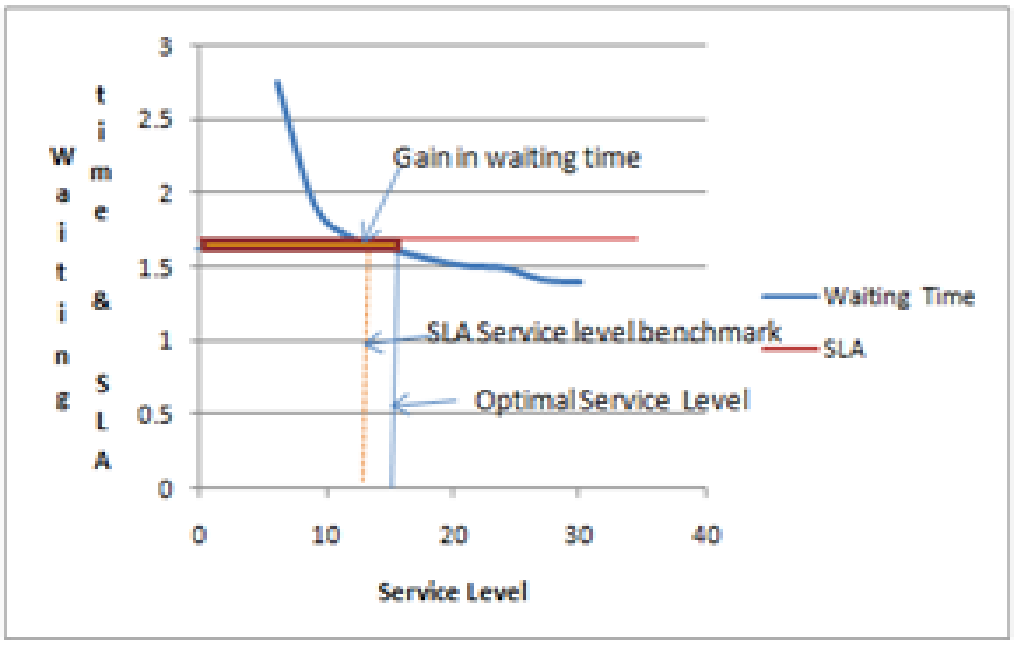

Fig. 3.7: SLA with Consumers Waiting Time

Table 3-2: Overall Analysis

Server Machine	Consumer Time (Loss/Gain in secs.)	Consumer - SLA	Cloud e-Provider
6 - 9	-1.0404 to -0.2087	Breach of SLA	Cost Reduction Non-competitive
12 - 15	0.0097 to 0.0811	Gain in Service Time Happy customer	Cost Reduction Less competitive
18	0.1495	Gain in Service Time Happier Customer	Optimal Cost Competitive
21 - 30	0.1977 to 0.3057	Gain in Service Time	Cost Increase Price war situation

		Cheaper than competitors May engender loyalty Larger market share

A further increase in server machines (from 21-30) increased cost and reduced waiting time to a level where the waiting time is almost constant. This may point to a price war situation in which a provider is implementing a strategy to become cheaper than competitors in order to engender customer loyalty and thereby gain a larger market share.

3.6 ASPIRATION LEVEL MODEL

The viability of the cost model proposed depends on how well the cost parameters can be estimated, which indeed is also difficult [120]. By viability the researcher means the likelihood of having a reasonable chance of success. For example, to determine the cost of waiting for x web applications (clients or consumers) will require so many assumptions by human being as a result of the dynamic change and uncertainty of the waiting time. This and other fundamental obstacles make it difficult to apply them to many real world problems. Even previous research in experimental economics and cognitive psychology has shown that human decision makers often do not adhere to fully rational behaviour [124]. The work of [125] also showed that individuals often deviate from optimal behaviour as prescribed by Expected Utility Theory. In addition, decision makers do not know the quantitative structure of the environment in which they act, as a result of lack of complete information. Even when people act rationally they cannot always compute the optimal solution for a given problem, as they lack the required facts for the computation to succeed. Therefore the

difficulty involved makes the expectation unrealistic. Also, some consumers sometimes base their choices on more than one conflicting measure thereby making it imperative to have a model that will solve such problem. To resolve this, this study extends the existing and widely adopted aspiration theory to Cloud E-Marketplaces. The idea is not to solve for optimal solution but to reveal an acceptable range for the service level by specifying reasonable limits the provider wishes to reach on the conflicting measure of performance.

This thesis proposes to extend existing and widely adopted aspiration theory to Cloud E-Marketplaces. This investigation uses the aspiration model based on the consumers' average waiting time in the system ($\mathrm{Ws\_total}$) derived from Eq. 3.40 and the percentage of the servers' idleness (S_Idle) in both dispatcher and web queue stations derived from the same equation as shown in Fig. 3.1. These two conflicting measures of performance serve as the control mechanism designed to regulate the service level. This mechanism translates dynamically to the voice of the decision makers depending on their choice. These are

$$\mathrm{Ws\_total} = \mathrm{Wq_{dispIn}} + \mathrm{Wq_{dbase}} + \mathrm{Wqueue_{ave}} + \mathrm{Wq_{dispOut}}$$

and

$$S\_Idle = \frac{\sum_{i=1}^{j} \frac{\left(\mathrm{Wsystem_i} - E\ (web\ queue_i)\right)}{c} * 100}{j}$$

$$= \frac{\sum_{i=1}^{j} \left(1 - \frac{\mathrm{k_i}\lambda_{eff}}{cu}\right) * 100}{j} \qquad (3.43)$$

where j is the number of web queue station in the model and c is the total number of server machines per web queue station.

The problem is therefore

to

minimise the service level of servers (c)

subject to

$$\mathrm{Ws\_total} \leq \alpha \qquad (3.44)$$

and

$$S_{Idle}\left(Idleserver\ period\ (\%)\right) \leq \beta \quad (3.45)$$

Where α could be the acceptable SLA waiting time and β the acceptable percentage of server idleness agreed by both the provider and the consumer in the E-Marketplace. This is done by plotting Ws_total and S_Idle server period as a function of the number of used servers (c).

3.7 SIMULATION II

The second simulation started with two server machines in each web queue station. At the end of every experiment the number of server machines was increased by one and a total of three web queue stations were used. The arrival rate was initialised to 0.1 sec and the service time set to 0.1 sec in each of the servers used. There was no bulking because the simulation started with six server machines. Each experiment was repeated ten times with a replication length of 100000 for 24hours per day. The results have been analysed and discussed in section 3.8.

3.8 RESULTS AND DISCUSION II

Both waiting time distributions and percentage of idleness based on the number of servers (c) used are provided. These are described in Table 3.3, and Figures 3.8 and 3.9 respectively. Fig 3.8 depicts the idleness percentage as a function of the used server machines. Similarly Table 3.3 and Fig. 3.8 confirmed that the percentage (%) of the Idle period increased as the number of used server machines was increased. Similarly in Fig 3.9, the waiting time reduced as the server machine increased. The result in Fig. 3.10 revealed the acceptable Aspiration Level 1 and 2 respectively where the constraints α and β are the two parameters of aspirations to be specified by the decision makers as required by equation 3.44 and 3.55..

The result reveals the acceptable Aspiration Level 1 and 2 respectively where the constraints α and β are the two levels of aspirations specified by the decision makers as required in equations 3.44 and 3.55. In this context, the decision makers may involve the cloud E-Market provider and the consumer. In Fig. 3.10 it was observed that the condition was set to determine the feasible server range to be used when the SLA waiting time ($\alpha 1$) was 0.00002693 and the server idle period ($\beta 1$) was set to 88%. The figure also showed that the Aspiration service level 1 (ASL1) ranges from 9 to-15 servers. This was considered the feasible region in which the SLA was enforced. Also, the Aspiration Service Level (ASL 2) adjusted to a new range of 12 to 17 server machines when the two conflicting measures were changed from $\alpha 1\ to\ \alpha 2$, and that of $\beta 1\ to\ \beta 2$, that is from 0.00002693 to 0.00002 (sec) and from 88% to 91.80% (see Fig. 3.11). One noticeable trend in these two figures was that the percentage of idleness increases while that of waiting time reduces. The reverse could be the case depending on the given constraints.

On the gain and loss of the model, it was observed that as the waiting time goes down, the number of server machines increases, which will inevitably increase costs. This is shown in Figure 3.12. For example, as the waiting time dropped from 0.00002693 to 0,00002, the researcher observed a drop of 0,00000693 (K1), which brings an increase of 2 additional server machines (i.e from 15 to 17).

Furthermore, as the percentage of the idle period increased, the number of server machines also went up, with corresponding upward cost. However, a drop in the percentage of idle period would reduce the servers requirement (k2) which will be at the expense of high cost of consumer's dissatisfaction, such as loss of future business and actual processing costs of complaints. Therefore, striking this balance depends on the given conflicting measures.

One major advantage of this model in Cloud E-Marketplaces is that, it allows the decision makers to predict the feasible outcome of the event based on the given

constraint (α, β). For intance, a provider with 20 server machines having the percentage of idleness of 86% in this context will not accept a consumer's request with an SLA of waiting time that is less than 0.000015 ($W_t \leq 0.000015$) because it is outside the feasible region and as such the provider cannot meet the server requirement.

As ealier mentioned, the research aim is not to determine the optimal solution but to find an acceptable range for the service level by specifying reasonable limits the provider wishes to reach on the conflicting measure of performance**.** One issue that needs to be discussed even though this did not happen in this experiment is when the two conditions cannot be satisfied simultanously. In that case, one or both must be relaxed before a feasible range can be attained.

This chapter has considered a typical E-Marketplace where consumers' requests are taken through some process. This is in contrast to previous attempts, (Xiong and Perros in [12]) that used the generalised approach of a single processing unit. The limitation of the generalised mechanism in the Cloud context is the reason for the work. This research even went further to investigate cost minimisation when the Cloud E-Marketplaces are modelled as networks of queues with feedback from the database. The specific components of Cloud considered were limited to the Dispatcher-In, database and the web sub stations queue. Cloud requests were distributed by the Dispatcher-In while the feedback information was collated by the database unit. The web queue stations served as the processing unit as in Figure 3.1.

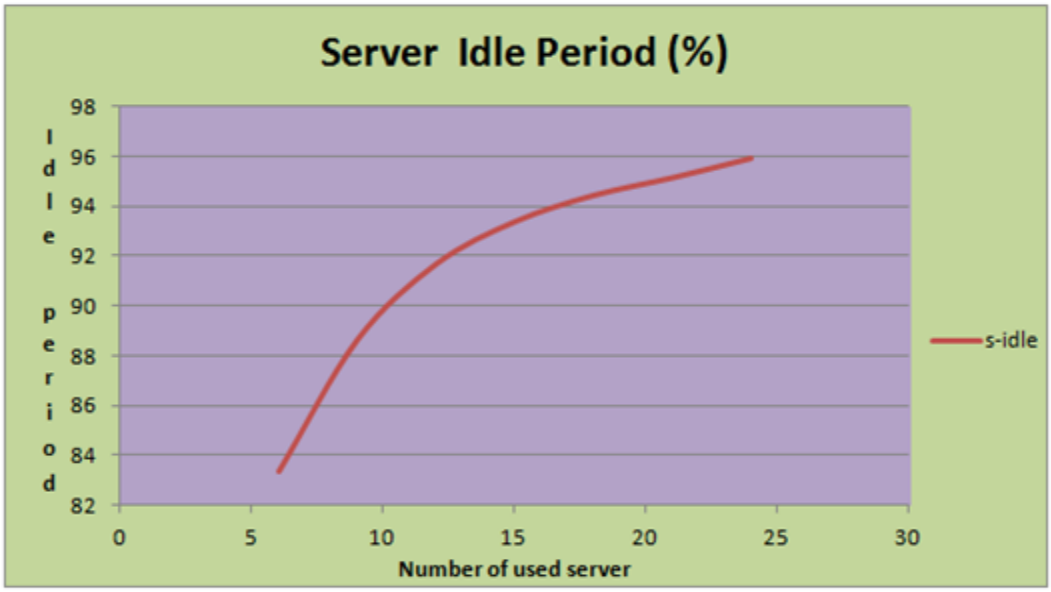

Fig. 3.8: Idle Period- Service Level

Table 3-3: Waiting Time and Idle Period Distributions

No of used server	Wt (sec)	S-idle (%)	Server Utilisation
6	0.00005459	83.31	0.1669
9	0.0000354	88.62	0.1138
12	0.00002693	91.662626	0.0833737
15	0.00002162	93.363636	0.0663636
18	0.00001838	94.422222	0.0557777
21	0.00001598	95.132323	0.0486767
24	0.00001369	95.913131	0.0408686

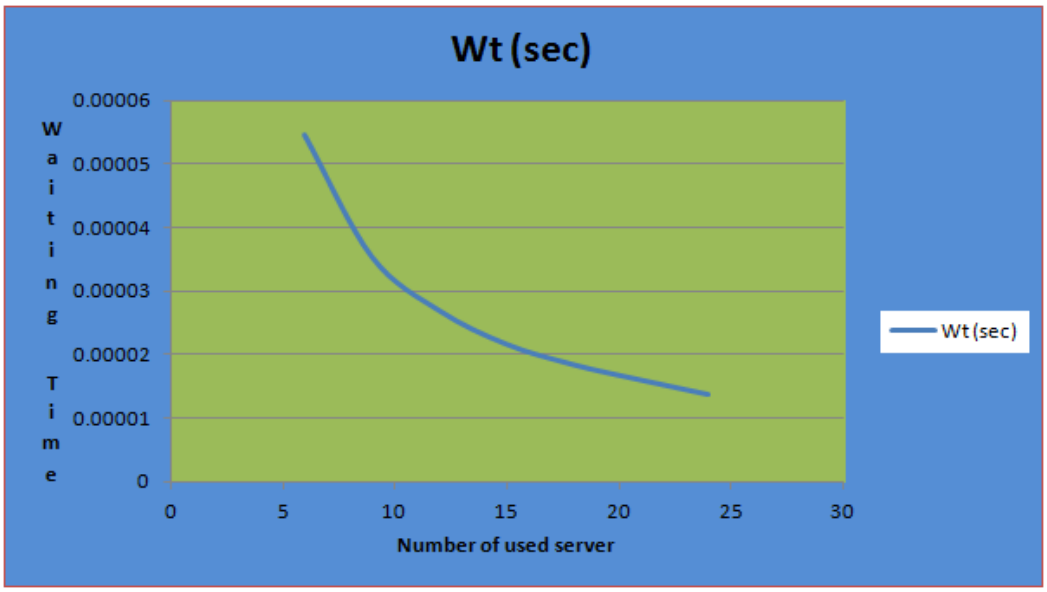

Fig. 3.9: Waiting Time of Consumers

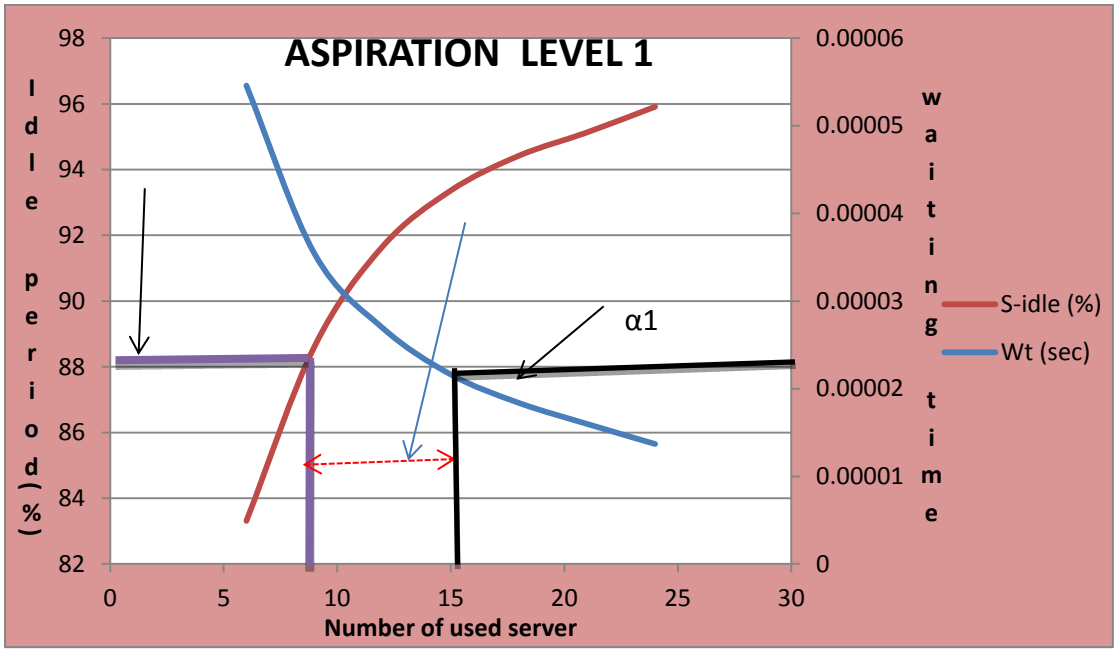

Fig. 3.10: Aspiration Level 1 (ASL 1)

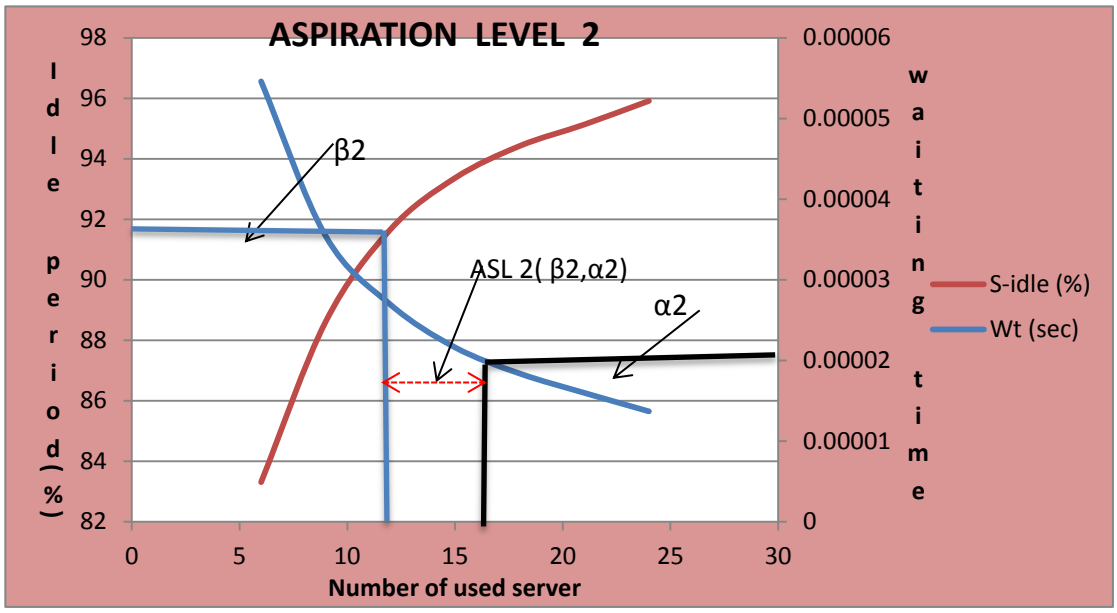

Fig. 3.11: Aspiration Level 2 (ASL 2)

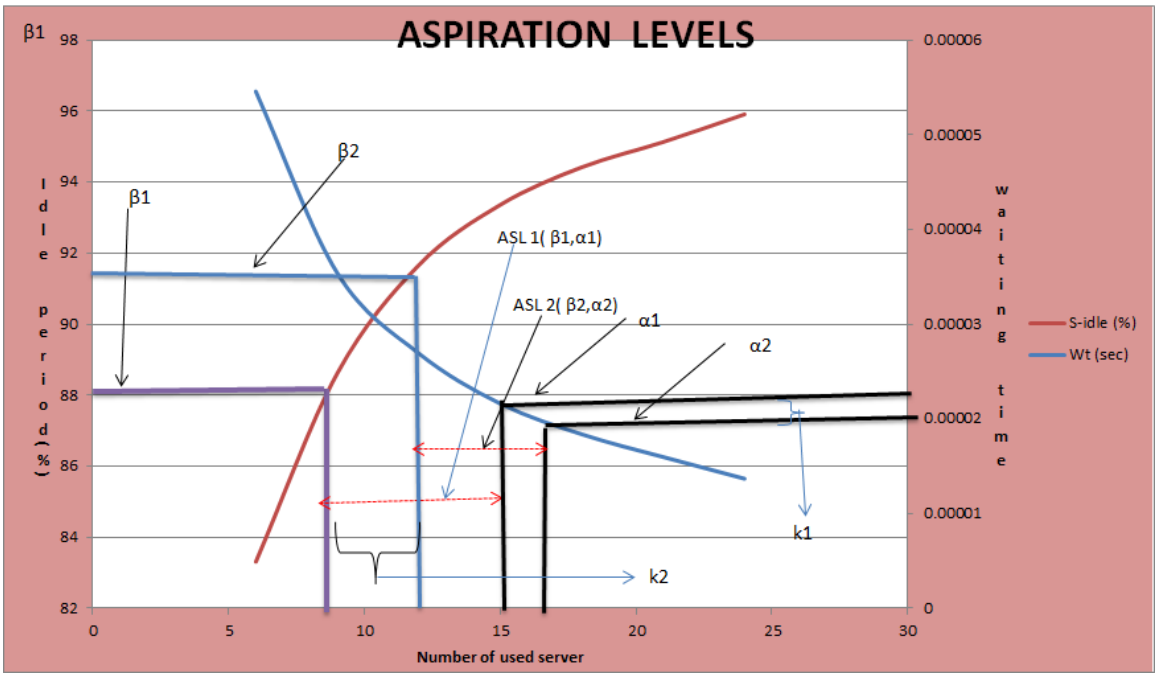

Fig. 3.12: Combined Aspiration Levels

3.9 CHAPTER SUMMARY

Dealing with the trade-off between resource management on the cost of providing good service and consumer waiting time in the context of a non-priority model was the focus of this chapter. This was achieved by re-engineering the model as networks of queues with feedback from the database to get an accurate performance measure for specified cost structure. The results from the first simulation demonstrated how the optimal number of sever machines needed to minimise cost and improve consumer waiting time could be determined. This was achieved at the point where the total cost is minimal. Because of the difficulty involved in determine consumer waiting time, the initial result was further improved through the use of the aspiration model. The focus of the investigation was to find an acceptable range for the service level by specifying reasonable limits the E-Market decision maker wishes to reach on the conflicting measures of performance. The aspiration model served as a regulatory mechanism that prescribes an acceptable range of server resources, which is an essential component of cloud resource management for better capacity planning**.** In the next section, the performance impact on consumer waiting time in Cloud E-Marketplaces in the context of two differentiated service provisioning is proposed.

CHAPTER FOUR

PERFORMANCE MODELLING OF CLOUD E-MARKETPLACE BASED ON TWO PRIORITY NON PREEMPTIVE MODEL

Cost minimisation in a non-priority environment has been discussed in chapter three. The proposed model can only be applied where the service provisioning is monotonic. As the E-Markets continue to grow and service consumers request more than one service provisioning this becomes a challenge. This chapter proposes and evaluates a study of performance impact on consumers waiting time in Cloud E-Marketplaces in the context of two differentiated service provisioning. This is achieved by MODELLING and evaluating a typical Cloud E-Marketplace with two classes of consumers under non preemptive priority discipline. The performance impact of these two classes is studied and compared with the non-priority discipline. The approach taken is both analytical and simulative.

4.1 INTRODUCTION

With the drift of consumers to Cloud E-Marketplaces for avoidable and cost effective service under different service provisioning, waiting time is of interest to every consumer and also a key source of competitive advantage for any Cloud E-Market provider. Therefore, the need for proper evaluation of the performance impact on consumers' waiting time and providers' cost is imperative, especially in the context of differentiated service offerings. In this research, a typical Cloud E-Marketplace is modeled as networks of queues under two classes. Higher priority is given to class one and lower to the other class. To achieve the objective, mathematical and simulation approaches are selected. In addition to the model described in chapter three, A Dispatcher-Out is proposed.

The dispatcher-In and Dispatcher-Out are modeled as M/M/1/Pr while that of the Web queue stations as M/M/c/Pr. For clarity purposes, the definitions of additional terms are included in Fig 4.1.

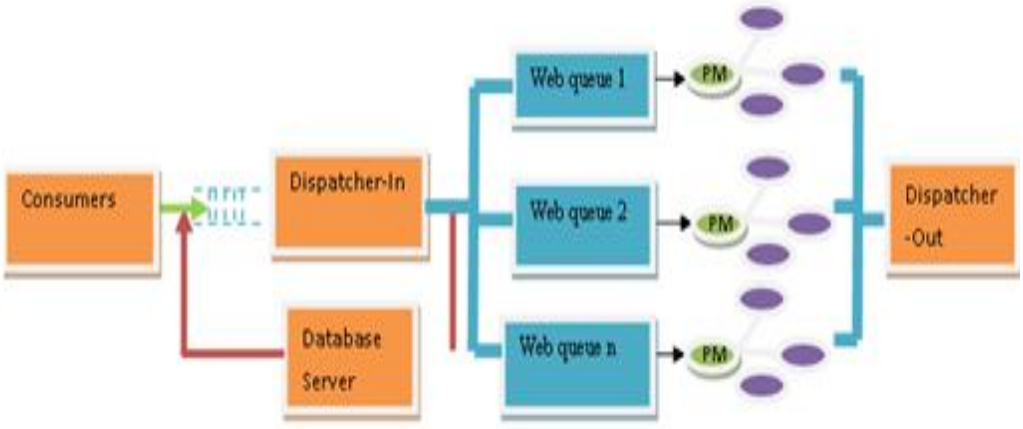

Fig 4.1 The Non Pre-emptive priority Model

ρ = Server utilisation i.e the percentage of the time the system is busy.

μ = Service rate in Dispatcher-In, and each of the web queues

P_0 = Probability of no consumer in the system

P_n = Probability of n consumers in the system

L_{1d} L_{2d} = Mean numbers of consumers in the system for each of the two priority classes in the dispatcher queue.

Lq_{1d} ,Lq_{2d} = represent the mean queue lengths of the two classes in the dispatcher queue.

Wq_{1d} Wq_{2d} = waiting time in the queue by each class in the dispatcher queue.

W_{qst} = Expected total waiting time experienced by the two classes in each service station.

$W(ave)_{qst}$= Average waiting time experienced by the two classes in the service stations

$H_{(y,\ z)}$ = The joint generating function for two priorities regardless of the one in service

The remainder of this chapter is organised as follows: Section 4.2 discusses the analytical description of the two priority models. Its three sub Sections discuss the mathematical MODELLING of the dispatcher-In queue, web station queue and the Dispatcher-Out queue. In section 4.3 the numerical validation and simulation set up are discussed and Section 4.4 focuses on results and discussion. The researcher concludes with the chapter summary in section 4.5.

4.2 THE ANALYTICAL MODEL DESCRIPTION OF TWO PRIORITY MODEL

In this chapter the model considers two priority classes as network of queues. The researcher consider the arrival entry of class 1 consumers with arrival rate λ_1 and class 2 with arrival rate λ_2. Consumers' requests are transmitted to the dispatcher-In web queue and then dispatched to any of the web service stations with equal probability distribution for virtual machine processing. Getting to any of these stations another queue is built up in any of the service stations, after which the consumer's request moves out through the Dispatcher-Out queue. This is depicted in Figure 4.1. This model considers class 1 consumers to have higher priority over class 2. The idea is that when a consumer of lower priority is on the queue and that of higher priority arrives, the higher one gets priority attention over the lower one. This model is non-preemptive, therefore whatever the priority is of a consumer in service the consumer has to complete its service before another is admitted and any priority of the same class value is based on First Come First Served (FCFS). Arrival and service time follow exponential distribution with an infinite buffer capacity from an infinite population.

Another assumption in this network, which is in line with chapter three, is that the latency of internal communication between the Dispatcher-In, the web queue service stations and Dispatcher-Out is insignificant [109]. To get the

performance measure the steps stated in [126] [45] and the law of conservation of flow [127] are followed.

4.2.1 MATHEMATICAL MODELLING OF THE DISPATCHER-IN QUEUE

Two MODELLING approaches were studied. The M/M/1/c and the M/M/1/Pr. The M/M/1/Pr is selected because it is the model that considers priority queue. Therefore, the Dispatcher-In web queue is modeled as M/M/1/Pr.

Let $\lambda = \lambda_1 + \lambda_2$

and

$p_{mnr} \equiv$ Pr{ at time t, m units of Class 1 and n unit of Class 2 consumers are in the Dispatcher-In and consumers of Class r = 1 or 2 in service}

The following difference stationary differential equations hold since $\rho = \frac{\lambda}{\mu}$ and $p_o = 1 - \rho$

$$0 = -\lambda p_0 + \mu(p_{101} + p_{012}) \quad (4.1)$$

$$0 = -(\lambda + \mu)p_{101} + \lambda_1 p_0 + \mu(p_{201} + p_{112}) \quad (4.2)$$

$$0 = -(\lambda + \mu)p_{012} + \lambda_2 p_0 + \mu(p_{111} + p_{022}) \quad (4.3)$$

$$0 = -(\lambda + \mu)p_{m01} + \lambda_1 p_{(m>1)1,0,1} + \mu(p_{m+1,0,1} + p_{m12}) \quad (4.4)$$

$$0 = -(\lambda + \mu)p_{0n2} + \lambda_2 p_{0,(n>1),2} + \mu(p_{1n1} + p_{0,n+1,2}) \quad (4.5)$$

$$0 = -(\lambda + \mu)p_{1n1} + \lambda_2 p_{1,(n>0),1} + \mu(p_{2n1} + p_{1,n+1,2}) \quad (4.6)$$

$$0 = -(\lambda + \mu)p_{m12} + \lambda_1 p_{(m>1),1,2} \quad (4.7)$$

$$0 = -(\lambda + \mu)p_{mn1} + \lambda_1 p_{m-1,n,1} + \lambda_2 p_{m,(n-1,1} + \mu(p_{m+1,n,1} + p_{m,n+1,2}) \ (m > 1, n > 0) \quad (4.8)$$

$$0 = -(\lambda + \mu)p_{mn2} + \lambda_1 p_{m-1,n,2} + \lambda_2 p_{m,(n-1,2} \quad (m > 0, n > 1) \tag{4.9}$$

The changing of service discipline has no effect on ρ_0 (Probability of idleness) therefore

$$p_0 = 1 - \rho \tag{4.10}$$

and

$$\rho_n = \sum_{m=0}^{n-1} (p_{n-m,m,1} + p_{m,n-m,2)} = (1-\rho)\rho^n \quad (\mathrm{n} > 0) \tag{4.11}$$

The busy percentage time of consumer with class r is

$\rho\lambda_r/\lambda$ and

$$\rho_1 = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p_{mn1} = \frac{\lambda_1}{\mu} \tag{4.12}$$

and

$$\rho_2 = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p_{mn2} = \frac{\lambda_2}{\mu} \tag{4.13}$$

Because of the triple subscripts, the two dimensional generating functions is applied. Therefore,

$$P_{m1}(z) = \sum_{n=0}^{\infty} z^n p_{mn1} \tag{4.14}$$

$$P_{m2}(z) = \sum_{n=0}^{\infty} z^n p_{mn2} \tag{4.15}$$

$$H_{1(y,\ z)} = \sum_{m=1}^{\infty} y^m p_{m1}(z) \ [with\ H_1(1,1) = \frac{\lambda_1}{\mu}] \tag{4.16}$$

$$H_{2(y,\ z)} = \sum_{m=0}^{\infty} y^m p_{m2}(z) \ [with\ H_2(1,1) = \frac{\lambda_2}{\mu}] \tag{4.17}$$

and

$$H_{(y,\ z)} = H_1(y,z) + H_2(y,z) + P_0 \tag{4.18}$$

$$= \sum_{m=1}^{\infty}\sum_{n=0}^{\infty} y^m z^n p_{mn1} + \sum_{m=o}^{\infty}\sum_{n=1}^{\infty} y^m z^n p_{mn2} + P_0 \tag{4.19}$$

$$= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} y^m z^n (p_{mn1} + p_{mn1+}) + \sum_{m=1}^{\infty} y^m p_{m01} + \sum_{n=1}^{\infty} z^n p_{0n2} + P_0 \tag{4.20}$$

where $H_{(y,\ z)}$ represents the generating function for the two priorities regardless of the service.

Representing $L_{1d}\ and\ L_{2d}$ as the mean number of consumers present for each of the two priorities classes of the dispatcher queue, then

$$\left.\frac{\partial H(y,z)}{\partial y}\right|_{y=z=1} = L_{1d} = Lq_{1d} + \frac{\lambda_1}{\mu} = \lambda_{1d}W_{1d} \tag{4.21}$$

$$\left.\frac{\partial H(y,z)}{\partial y}\right|_{y=z=1} = L_{2d} = Lq_{2d} + \frac{\lambda_2}{\mu} = \lambda_{2d}W_{2d} \tag{4.22}$$

where $Lq_{1d}\ and\ Lq_{2d}$ represent the mean queue length of class 1 and 2 in the dispatcher queue.

Let Eq. 4.1- 4.9 be represented as Eq. X, then multiplying Eq. X and the appropriate power of y and z and then summing it up will give

$$\left[1+\rho-\frac{\lambda_1 y}{\mu}-\frac{\lambda_2 z}{\mu}-\frac{1}{y}\right]H_{1(y,\ z)} = \frac{H_{2(y,\ z)}}{z}+\frac{\lambda_1 y P_0}{\mu}-P_{11}(z)-\frac{P_{02}(z)}{z} \tag{4.23}$$

and

$$\left[1+\rho-\frac{\lambda_1 y}{\mu}-\frac{\lambda_2 z}{\mu}-\frac{1}{y}\right]H_{2(y,\ z)} = P_{11}(z)+\frac{P_{02}(z)}{z}-\left[\rho-\frac{\lambda_2 z}{\mu}\right]p_0 \tag{4.24}$$

The values of $P_{11}(z),\ P_{02}(z)\ and\ P_0$ need to be known in order to have fully the generating functions of $H_1\ and\ H_2$ respectively. This is done by summing z^n ($n = 2,3\ldots$) times equation 1 that involves P_{0n2} .

$$P_{11}(z) = \left(1+\rho-\frac{\lambda_2 z}{\mu}-\frac{1}{z}\right)P_{02}(z)+\left(\rho-\frac{\lambda_2 z}{\mu}\right)p_0 \tag{4.25}$$

Substitute this into Eq. 23 and 24 then $H_1\ and\ H_2$ become a function of $p_0\ and\ P_{02}$, therefore

$$H_{(y,\ z)} = H_{1(y,\ z)} + H_{2(y,\ z)} + p_0 \tag{4.26}$$

$$= \frac{(1-y)\,p_0}{1-y-\rho y\left(1-z-\frac{\lambda_1 y}{\lambda}+\frac{\lambda_1}{\lambda}\right)} + \frac{\left(1+\rho-\rho z+\frac{\lambda_1 z}{u}\right)(z-y)P_{02}\,(z)}{z\left[1+\rho-\frac{\lambda_1 y}{u}-\frac{\lambda_2 z}{u}\right]\left[1-y-\rho y\left(1-z-\frac{\lambda_1 y}{\lambda}+\frac{\lambda_1 z}{\lambda}\right)\right]} \tag{4.27}$$

Since the condition that $H(1,1) = 1$ then,

$$P_{02}\,(1) = \frac{\lambda_2 p_0/u}{(1+\lambda_1/u)(1-\rho)} \tag{4.28}$$

Taking the partial derivative of H with respect to both y and z evaluate to find the means measure of effectiveness i.e $L_{1d}\ and\ L_{2d}$ where the $P_{02}\,(1)$ surfaces.

$$L_{1d} = \frac{(\lambda_1/u)(1+\rho-\lambda_1/u)}{1-\lambda_1/u} \tag{4.29}$$

$$Lq_{1d} = \frac{\rho\lambda_1/u}{1-\lambda_1/u} \tag{4.30}$$

$$Wq_{1d} = \frac{\lambda}{u(u-\lambda_1)} \tag{4.31}$$

$$L_{2d} = \frac{(\lambda_2/u)(1-\lambda_1/u+\rho\,\lambda_1/u)}{(1-\rho)\,(1-\lambda_1/u)} \tag{4.32}$$

$$Lq_{2d} = \frac{\rho\,\lambda_2/u}{(1-\rho)\,(1-\lambda_1/u)} \tag{4.33}$$

$$Wq_{2d} = \frac{\lambda}{(u-\lambda)(u-\lambda_1)} \tag{4.34}$$

4.2.2 MODELLING OF WEB QUEUE STATION

Unlike the M/M/1/Pr that considered only one station, The M/M/c/Pr that considers many service stations is used. This is because the web queue stations have n number of servers where n in the experiment is 3. The M/M/c/Pr is governed by identical exponential distributions for each priority at each of the c channels within a station. As mentioned earlier, the service rate is equal in this context.

therefore

$$\rho_{k} = \frac{\lambda_{\mathrm{k}}}{c\mu_{k}} \quad (1 \leq k \leq r) \tag{4.35}$$

and

$$\sigma_{k} = \sum_{i=1}^{k} \rho_{\mathrm{k}} \quad (\sigma_{0} \equiv \rho = \lambda/\mathrm{c}\mu) \tag{4.36}$$

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

$$W_{q}^{(i)} = \sum_{k=1}^{i-1} E[S_{k}'] + \sum_{k=1}^{i} E[\mathrm{S}_{0}] \tag{4.37}$$

S_{k} is the time required to serve n_{k} consumers of the kth priority in the line ahead of the consumer. S_{k}' is the service time of the n_{k}' consumers of priority k which arrive during $W_{q}^{(i)}$ and S_{0} is the amount of time remaining until the next server becomes available.

Therefore

$$E[\mathrm{S}_{0}] = \Pr\left(\begin{matrix}all\ servers\ are\ busy\ within\ a \\ service\ station\end{matrix}\right).\mathrm{E}[\mathrm{S}_{0}|all$$

$$server\ are\ busy\ within\ a\ service\ station$$

$$= \left(\sum_{n-c}^{\infty} P_{n}\right)\frac{1}{c\mu} = P_{0}\left(\sum_{n-c}^{\infty} \frac{c\rho^{n}}{c^{n-c}\,c!}\right)\frac{1}{c\mu} = \frac{P_{0}(c\rho)^{c}}{c!(1-p)} \tag{4.38}$$

$$= \frac{(c\rho)^{c}}{c!\,(1-\rho)(\mathrm{c}\mu)}\left[\sum_{n=0}^{c-1}\frac{(c\rho)^{n}}{n!} + \frac{(c\rho)^{c}}{c!\,(1-\rho)}\right]^{-1} \tag{4.39}$$

but

$$E[S_0] = \rho \sum_{k=1}^{r} \frac{1}{u_k} \frac{\rho_k}{\rho} = \sum_{k=1}^{r} \frac{\rho_k}{u_k} \tag{4.40}$$

$$W_q^{(i)} = \frac{E[S_0]}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{4.41}$$

$$W_q^{(i)} = \frac{\sum_{k=1}^{r} \frac{\rho_k}{u_k}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{4.42}$$

Therefore

$$W_q^{(i)} = \frac{\left[c!\,(1-\rho)(c\mu)\sum_{n=0}^{c-1}(c\rho)^{\frac{n-c}{n!}} + c\mu\right]^{-1}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{4.43}$$

The expected time taken in a service station is

$$W_{qst} = \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_{qst}^{(i)} \tag{4.44}$$

The overall average time taken in j service stations is

$$W(ave)_{qst} = \frac{\sum_{n=1}^{j} \left[\sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_{qst}^{(i)}\right]}{j} \tag{4.45}$$

4.2.3 MODELLING THE DISPATCHER-OUT

The Dispatcher–Out is similar to the Dispatcher-in, therefore the performance measures are given below

$$L_{1dout} = \frac{(\lambda_1/u)(1+\rho-\lambda_1/u)}{1-\lambda_1/u} \tag{4.46}$$

$$Lq_{1dout} = \frac{(\rho\lambda_1/u)}{1-\lambda_1/u} \tag{4.47}$$

$$Wq_{1dout} = \frac{\lambda}{u(u - \lambda_1)} \tag{4.48}$$

$$L_{2dout} = \frac{(\lambda_2/u)(1 - \lambda_1/u + \rho\,\lambda_1/u)}{(1-\rho)\,(1-\lambda_1/u)} \tag{4.49}$$

$$Lq_{2dout} = \frac{\rho\,\lambda_2/u}{(1-\rho)\,(1-\lambda_1/u)} \tag{4.50}$$

$$Wq_{2dout} = \frac{\lambda}{(u-\lambda)(u - \lambda_1)} \tag{4.51}$$

The concern in this experiment is the waiting time experienced by consumers. Therefore the total waiting time experienced in the queue by the consumer is

$$Wq_{Tot} = Wq_{1d} + Wq_{2d} + W(ave)_{qst} + Wq_{1dout} + Wq_{2dout} \tag{4.52}$$

4.3 NUMERICAL VALIDATION AND SIMULATION

The illustration of this model is shown by numerical examples and the impact on the two performance of the priority classes are analysed in the results and discussion section. The research considers two arrival processes $\lambda_1 and\ \lambda_2$ where $\lambda_1 = \lambda_2$

where $\lambda = \lambda_1 + \lambda_2 = 10,20,30, \ldots\ldots 97$ respectively and c = 2 in each of the service stations.

As earlier mentioned, the Discrete Event Simulator (DES), tool is Arena. The same values are used to ascertain the degree of variability. This simulation is run with a replication length of 1000 in 24 hours per day with base time in hours and replicated 5 times. In addition, the researcher set up a similar experiment with the same configurations but under a non-priority discipline using FCFS discipline. The comparison result is also discussed in the results and discussion section.

4.4 RESULTS AND DISCUSSION

The results of the analytical and simulation models are compared to ascertain the degree of correctness. This represents the Non-pre-emptive waiting time of consumers in the analytical and the simulation model. The result reveals that the analytical approach agrees with the simulation as shown in Figure 4.2 and Table 4.1. The graph in Figure 4.3 shows the two classes as functions of consumer waiting time and service utilisation (ρ). From the Non-preemptive graph it is observed that as $\rho \rightarrow 1$, the waiting time of class 2 consumers $\rightarrow \infty$. That is, it increases such that the deviation from the total waiting time is very close, unlike the class 1 consumers where only a small change is experienced with almost a finite limit.

When the consumer total waiting time is added, it is noticed that this class 2 consumer waiting time is very close to the total waiting time. This is shown in Figure 4.4. This implies that the class 1 consumers spent less time, at the expense of the class 2 consumers. For example, where the total waiting time is 1.1148 in Table 4.1 and Figure 4.4, the waiting time of the class 1 consumer is 0.04008 while that of class 2 is 1.0748, which is almost close to the total waiting time. This is a great danger as this may bring loss of consumers' satisfaction and also reduces the patronage of consumers to the Cloud provider.

When the non-priority discipline is implemented using the FCFS policy as shown in Table 4.2, observations reveal that the total waiting time in the simulation of Table 1 in non-preemptive priority service discipline and Table 4.2 of the non-priority service discipline are the same. The observation further reveals different waiting time distribution in each class. This implies that the total waiting time is independent of the service discipline but with different waiting time distributions.—Both this Table and Figure 4.5 show that the waiting time distributions of the two classes under the non-priority are almost the same, therefore the class 2 colour overshadow that of class 1 in the same figure.

The overall performance of the non-preemptive priority and the non-priority is shown in Figure 4.6. Here, the waiting time distributions of all the classes have a close range when the server utilisation is small. That implies that the effect may not be felt much when the consumers' arrival rate is small. As this grows, then Class 2 Non pre-emptive priority (Class 2 np) stays longer than both class 1 of non-preemptive priority (Class 1np) and the two classes of the non-priority discipline. As mentioned earlier, the distributions of class 1non priority (class 1 npr) and class 2 non priority are the same and that accounts for only seeing the colour of Class 2 (Class 2 npr) overshadowing the class 2 non priority. The deduction is that a great caution is required when introducing this policy in Cloud E-Marketplaces with two different service provisioning.

In chapter three, the Cloud E-Marketplace under the cost minimisation in non-priority environment was studied. However, this can only be applied where the service provisioning is monotonic. As the E-Marketplaces continue to grow and service consumers request more than one service provisioning this becomes a challenge. In this chapter, a non –preemptive model was designed where each model sub-unit are prioritized. This caters for two differentiated services where consumers' request are dispatched and processed in a non-preemptive prioritized way as shown in Figure 4.1. The evaluation of the performance impact on consumers' waiting time and providers' cost is the focus of this chapter. What differentiates this chapter from the previous chapter are the context and purpose. In this chapter, the context is based on a non-preemptive prioritised environment as against the non-priority in chapter three, and the purpose is the evaluation of the performance impact on consumers' waiting time. This research is closely related to [109][110] where these authors model the Cloud as a series of queues. What differentiated this work in this chapter are:

i. Each service station is modeled as M/M/c/Pr in contrast to the M/M/1 proposed elsewhere which requires a different mathematical concept.

ii. No dedicated server is given or allocated to any class.

Apart from the fact that great caution is required when introducing this policy in the Cloud E-Marketplace, its hallmark is a great improvement in consumer waiting time in both Classes compared to similar studies.

.

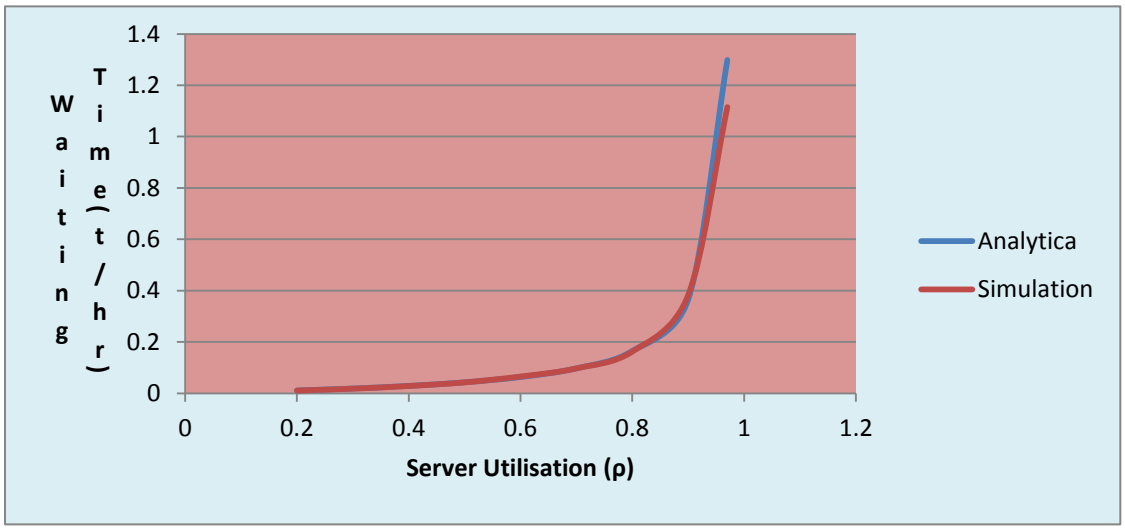

Fig. 4.1: Analytical and Simulation

Table 4-1: Result of the Simulation and Analytical

	Simulation			**Analytical**		
			Total Waiting			Total Waiting
ρ	class 1	class2	Time	class 1	class 2	Time
0.2	0.0049	0.0059	0.0108	0.005	0.0061	0.0111
0.3	0.0077	0.0104	0.0181	0.0078	0.011	0.0188
0.4	0.0109	0.0176	0.0285	0.011	0.0179	0.0289
0.5	0.0142	0.0284	0.0426	0.0146	0.0283	0.0429
0.6	0.0188	0.0466	0.0654	0.0186	0.0449	0.0636
0.7	0.0233	0.0739	0.0972	0.0231	0.0742	0.0973
0.8	0.0286	0.1341	0.1628	0.0286	0.1365	0.1651
0.9	0.0359	0.3471	0.383	0.0349	0.3312	0.3661
0.97	0.0401	1.0748	1.1148	0.0399	1.2585	1.2984

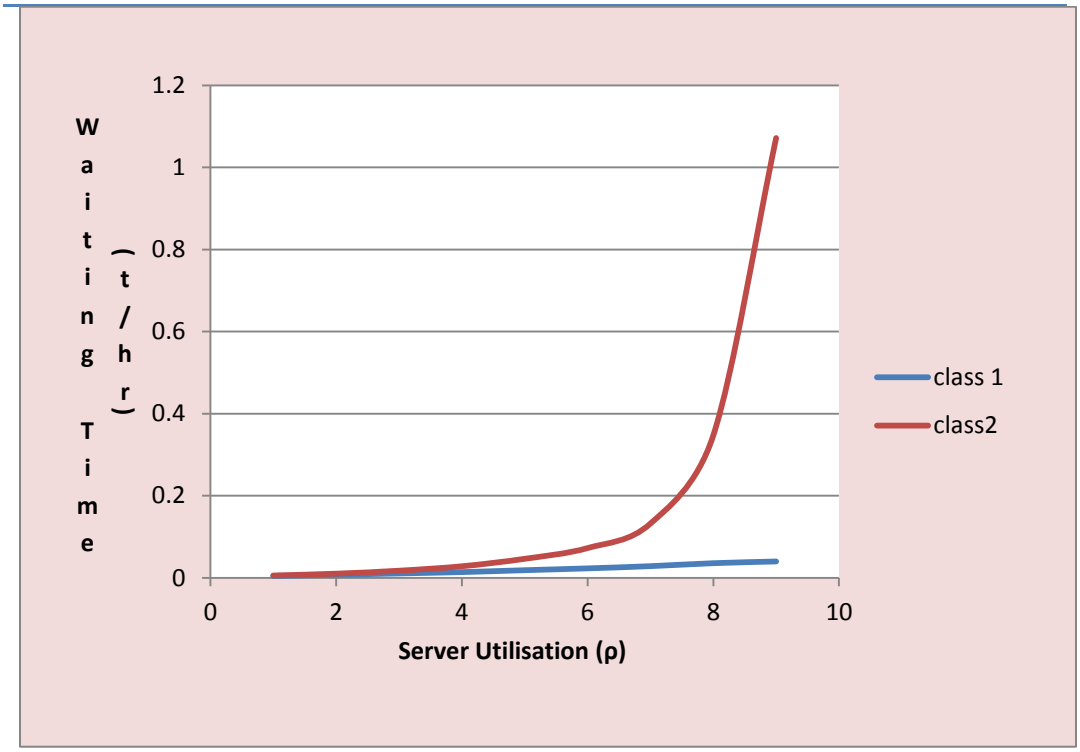

Fig. 4.2: Two Class Non Preemptive Priority

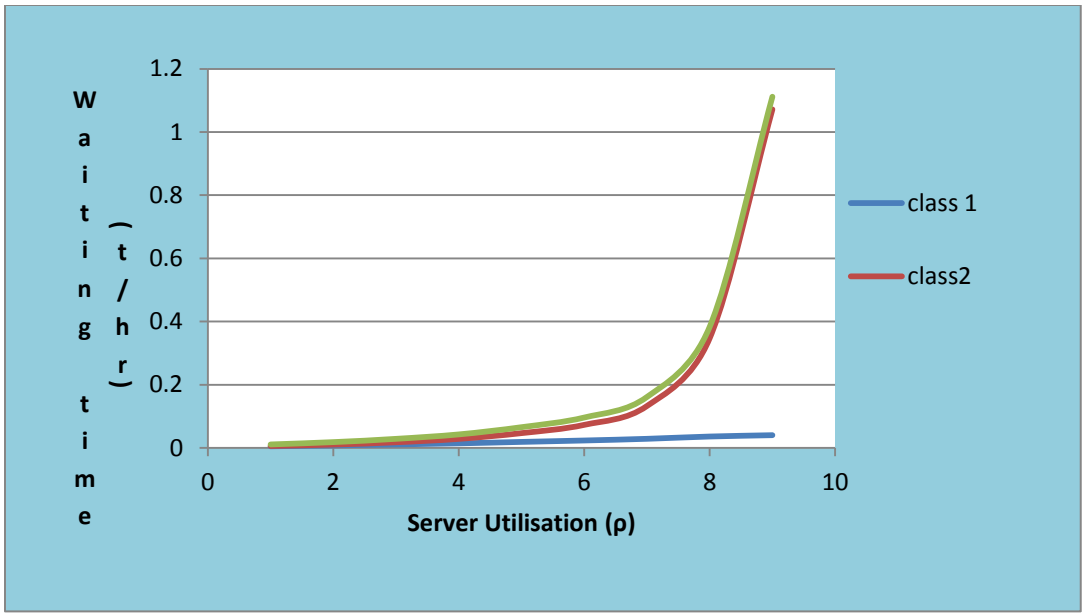

Fig. 4.3: Two Class Priority with Total Waiting Time

Table 4-2: Non Priority System (FCFS)

			Total Waiting
ρ	class 1	class 2	Time
0.2	0.0055	0.0053	0.01077649
0.3	0.0091	0.009	0.01810646
0.4	0.0143	0.0142	0.02854418
0.5	0.021	0.0215	0.04255906
0.6	0.0329	0.0324	0.06531286
0.7	0.0487	0.0484	0.09712725
0.8	0.081	0.0815	0.16256195
0.9	0.1922	0.1926	0.3847291
0.97	0.5574	0.5573	1.1147

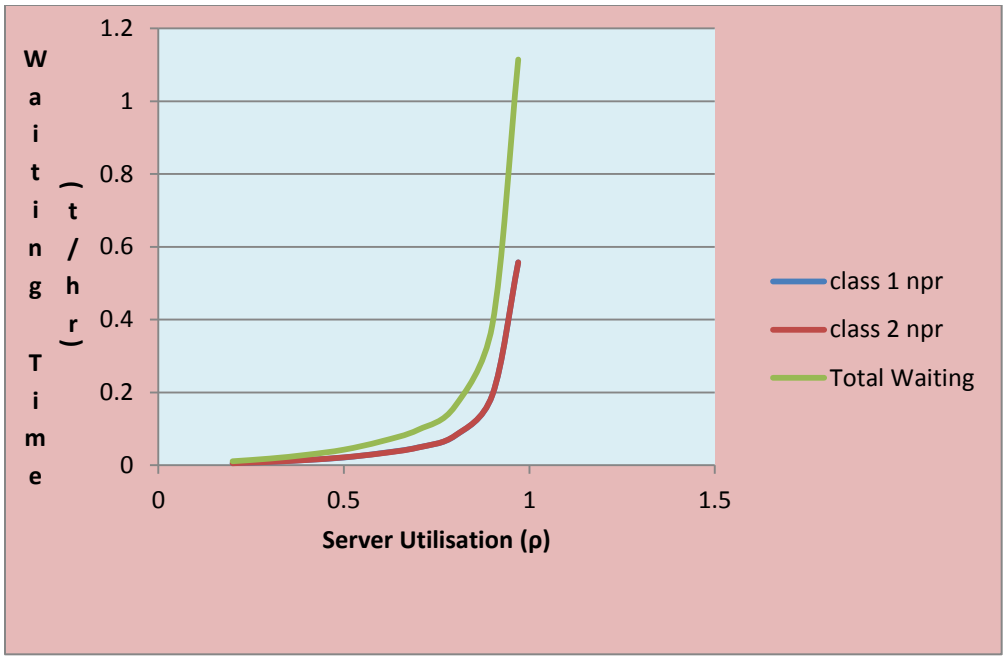

Fig. 4.4: Two Class Non Priority with Total Waiting Time

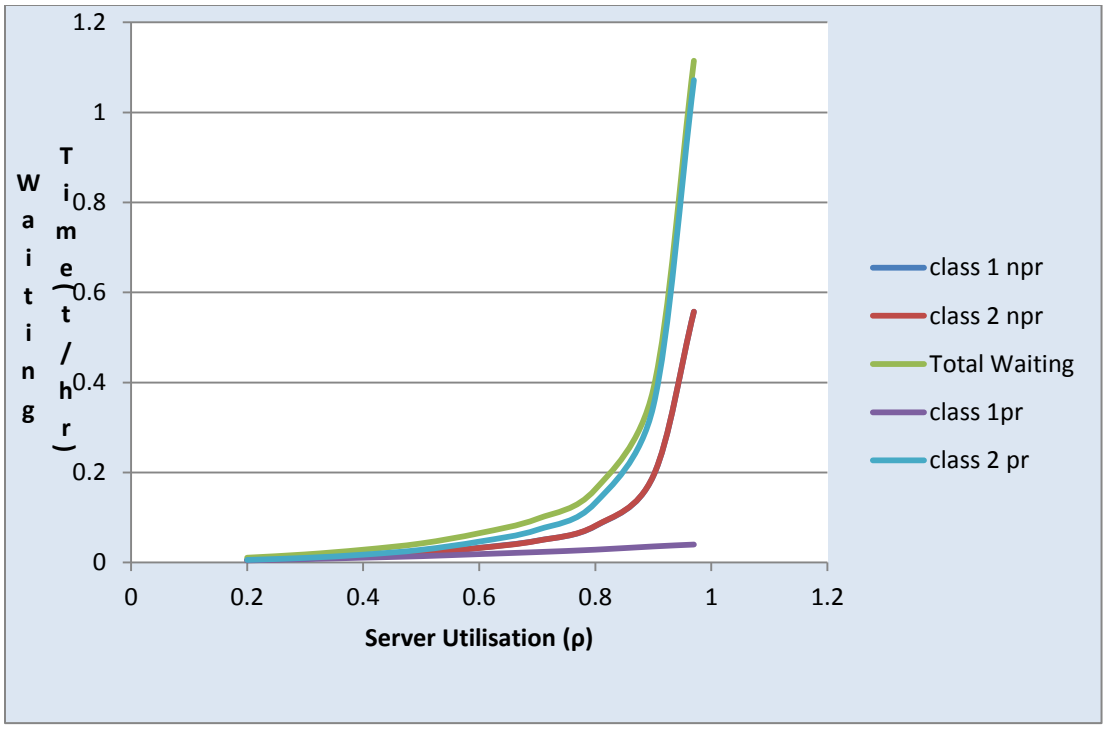

Fig. 4.5: Performance of the Non Priority /Non Preemptive Priority Classes

4.5 CHAPTER SUMMARY

In this chapter, the performance evaluation of the two priority queue system in a typical Cloud E-Marketplace was presented. The Cloud was depicted as networks of queues, and the performance impact of its two classes on consumers' waiting time was studied. The performance results revealed that the Class two consumers have a longer waiting time than the Class one consumers, but the average waiting time remains the same across the board. Also the total waiting time was independent of the service discipline. Although this will be applicable in the business world, especially where the service provisioning is differentiated based on time and cost, great caution is required because as $\rho \to 1$, the expectations of the average Class two consumer and their waiting time in queue $\to \infty$ thereby leading to lower consumers' satisfaction.

CHAPTER FIVE

PERFORMANCE MODELLING OF CLOUD E-MARKETPLACE BASED ON GENERALISED NON PREEMPTIVE MODEL FOR BALACING SERVICE LEVEL AND CONSUMERS' WAITING TIME

In chapter four, the performance impact of consumer waiting time was studied using a two class Non pre-emptive policy. With the rapid growth of Cloud E-Marketplaces, using two class priority may not be the best option as most providers and consumers offer more than two services. Therefore, there is a need for the provisioning of a heterogeneous service that caters for more than two services. This generalised idea requires complex theoretical and practical design. This work further extends existing and widely adopted theories to a generalised Non Preemptive model by using the queuing theory to formulate the mathematical model and the simulation to demonstrate a real life scenario. The performance analysis of the E-Marketplace under a generalised approach using five classes of consumers is first studied. Formulation of a good cost structure to get the optimal service level that balances consumers' waiting time and providers' cost is later studied.

5.1 INTRODUCTION

With a concomitant increase in the number of consumers in Cloud E-Marketplace, many Traditional service providers are rebranding cloud-hosted services. The idea of monotonic provisioning of service may not fit the requirements of the real business environment. Therefore, there is the need for provisioning of a heterogeneous service.

As Cloud E-Marketplaces grow, most providers are rolling out multiple service offerings to their consumers. For example, Amazon Elastic Compute Cloud (EC2) offers three different services, classified as Reserved, Spot, and On-Demand [20]. Reserved services have been pre-paid, with the assurance of no or minimal

delay, and therefore requiring higher priority. Spot services are also allocated in advance while the On-Demand has no facilities for advance payment or reservations and there is no commitment. Therefore, in this case of Amazon, using the two class non-preemptive approach may be practically unrealistic.

To resolve this, the use of pre-emptive policy has been proposed by scholars, for example Snathosh and Ravichandran, in [28]. However, there is a price to pay [128]; the policy is costly In addition, it is associated with appreciably high response time to consumers' requests especially when the requests are deadline constrained [30]. This chapter applies existing and widely adopted theories from the Pre-emptive to Non Preemptive models in the context of the Non Preemptive system. The research first evaluates the performance impact on consumers' waiting time. It then determines the optimal machine configuration that will minimise costs, and at the same time satisfy consumers' waiting time in a typical Cloud E-Marketplace in the context of a Non Preemptive system.

The remainder of this chapter is organised as follows: section 5.2 describes the generalised non pre-emptive model. In section 5.3, the non-pre-emptive cost function is formulated, while simulation and numerical validation are discussed in section 5.4. The results are laid out and extensively discussed in section 5.5. The chapter closes with a summary in section 5.6.

5.2 GENERALISED NON PREEMPTIVE MODEL

There are five classes of consumers in the Cloud E-Marketplace, based on priority, with class 1 consumers having higher priority than class 2,3,4 and 5.

When an incoming request meets a lower one on the queue, it takes

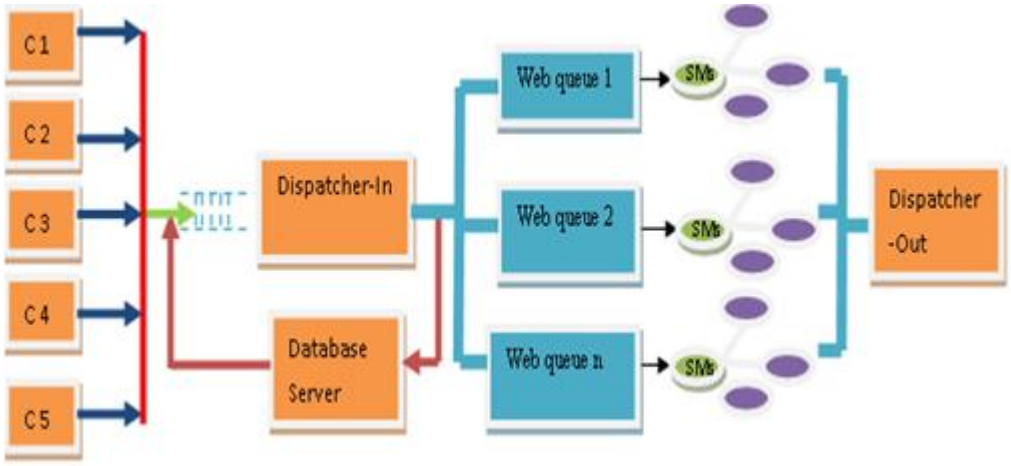

Fig. 5.1: Generalised Non Pre-emptive Priority Model

over from that request, but when a lower request is being processed that request is allowed to finish. The proposed generalised Non-Preemptive model is depicted in Fig. 5.I. When requests of the same priority are on the queue then the order is based on First Come First Serve (FCFS). Every request gets to the Dispatcher-In server as the first point and the requests are then distributed to the web queue stations for processing. The processed requests then move out through the Dispatcher-Out.

Unlike most related works [26][4][109][110], where the non-preemptive policy has been implemented at the first point of entry alone, the researcher modeled the non-preemptive model at every point in a queue. This is because at every point in a queue, there is the likelihood that a higher priority request will arrive when a lower one is still on the queue. That now enables the researcher to model the Dispatcher-In queue as M/M/1/Pr, the Web station queue as M/M/c/Pr and the Dispatcher out Queue as M/M/1/Pr.

5.2.1 MATHEMATICAL MODELLING OF DISPATCHER-IN QUEUE AS M/M/1/Pr

Let consumers with service of the k[th] priority (the smaller the number, the higher the priority) arrive before the Dispatcher-In according to a Poisson distribution with parameter λ_k (k = 1,2,….r) and that these consumers wait on a First Come First Served basis within their respective priorities. Let the service distribution for the k[th] priority be exponential with mean $1/\mu_k$. As said earlier, whatever the priority of a unit in service it has to complete its service before another item is admitted.

therefore

$$\rho_{k=} \frac{\lambda_k}{\mu_k} \quad (1 \leq k \leq r) \tag{5.1}$$

$$\sigma_k = \sum_{i=1}^{k} \rho_k \quad (\sigma_0 \equiv 0, \sigma_r \equiv \rho\,) \tag{5.2}$$

The system is stationary for $\sigma_r = \rho < 1$. Let a consumer of priority *i* arrive at time t_0 and enter service at time t_1. Its line wait is thus $T_q = t_1 - t_0$. At t_0. Consider for example n_1 consumers of priority one in line ahead of this new arrival, n_2 of priority two, n_3 of priority three, and so on. Let S_0 be the total time required to finish the job already in service and S_k be the total time required to serve n_k. During the new consumer's waiting time T_q (say) n'_k consumers of priority $k < 1$ will arrive and go to service ahead of this current arrival. If S'_k is the total service time of all n'_k, then it can be seen that

$$T_q = \sum_{k=1}^{i-1} S'_k + \sum_{k=1}^{i} S_k + S_0 \tag{5.3}$$

Taking the expected values from both sides of the foregoing, then

$$W'_k = E\left[T_q\right] = \sum_{k=1}^{i-1} E[S'_k] + \sum_{k=1}^{i} E[S_k] + E[S_0\,] \tag{5.4}$$

Since $\sigma_{i-1} < \sigma_i$ for all *i*, then ρ < 1 implies that $\sigma_{i-1} < 1\ for\ all\ i.$

To find $\mathrm{E}[\mathrm{S}_0]$,it is observed that the combined service distribution is the mixed exponential, which is formed from the law of total probability as

$$B(t) = \sum_{k=1}^{i} \frac{\lambda_k}{\lambda} (1 - e^{-\mu_k t} \tag{5.5}$$

where

$$\lambda = \sum_{k=1}^{r} \lambda_k \tag{5.6}$$

The random variable remaining time of service S_0 has the value 0 when the system is idle; hence

$$\mathrm{E}[\mathrm{S}_0] = \Pr\{\text{ system is busy}\}\ E\ [\ \mathrm{S}_0\ |busy\ system]$$

But the probability that the system is busy is

$$\lambda = \sum_{k=1}^{r} \frac{\lambda_k}{\lambda} \frac{1}{\mu_k} = \rho \tag{5.7}$$

where ρ is the server utilisation and

$$E[\ \mathrm{S}_0\ |busy\ system] = \textstyle\sum_{k=1}^{r} \lambda_k\ E[\ \mathrm{S}_0 \mid system\ busy\ wit\ k\ type\ consumer\ ]$$

$$.\ \Pr\{\ cutomer\ has\ priority\ k\} \tag{5.8}$$

$$= \textstyle\sum_{k=1}^{r} \frac{1}{\mu_k} \frac{\rho_k}{\rho} \tag{5.9}$$

therefore,

$$E[\mathrm{S}_0] = \rho \textstyle\sum_{k=1}^{r} \frac{1}{\mu_k} \frac{\rho_k}{\rho} = \sum_{k=1}^{r} \frac{\rho_k}{\rho} \tag{5.10}$$

Since n_k and the service times of individual consumers $S_k^{(n)}$, are independent,

$$E[\mathrm{S_k}] = E\left[\mathrm{n_k} S_k^{(n)}\right] = E\left[\mathrm{n_k}\right] E[S_k^{(n)}] = \frac{E[\mathrm{n_k}]}{\mu_k} \tag{5.11}$$

utilising the little's formula then gives

$$E[\mathrm{S_k}] = \frac{\lambda_k W_q^{(k)}}{\mu_k} = \rho_k W_q^{(k)} \tag{5.12}$$

Similarly,

$$E[S_k'] = \frac{E[\mathrm{n_k'}]}{\mu_k} \tag{5.13}$$

and then utilizing the uniform properties of the Poisson then

$$E[S_k'] = \frac{E[\mathrm{n_k}]}{\mu_k}\frac{\lambda_k W_q^{(i)}}{\mu_k} \tag{5.14}$$

Therefore $W_q^{(i)}$, which is the waiting time for the consumer of *i* priority is

$$W_q^{(i)} = W_q^{(i)} \sum_{k=1}^{i-1} \rho_k \quad + \sum_{k=1}^{i} \rho_k W_q^{(k)} + \quad E[\mathrm{S_0}] \tag{5.15}$$

$$W_q^{(i)} = \frac{\sum_{k=1}^{i} \rho_k W_q^{(k)} \ + E[\mathrm{S_0}]}{1-\sigma_{i-1}} \tag{5.16}$$

$$W_q^{(i)} = \frac{E[\mathrm{S_0}]}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.17}$$

Using Equation 10 finally gives

$$W_q^{(i)} = \frac{\sum_{k=1}^{r} \rho_k / \mu_k}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.18}$$

The above equation holds as long as $\sigma_i = \sum_{k=1}^{r} \rho_k < 1$

Therefore, from Little's formula

$$L_q^{(i)} = \sum_{k=1}^{r} \lambda_i W_k^{(i)} \ = \frac{\lambda_i \sum_{k=1}^{r} \rho_k / \mu_k}{(1-\sigma_{i-1})(1-\sigma_i)}$$

The total expected system size in each of the single channel is

$$L_q = \sum_{i=1}^{r} L_q^i \ = \frac{\lambda_i \ \sum_{k=1}^{r} \rho_k / \mu_k}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.19}$$

Because this model, unlike others, [12][4][16], has feedback from the database, the researcher needs to know the expected number of visits to the dispatcher ($\mathrm{Ex_{visitdisp}}$) from the database.

$$\mathrm{Ex}_{\text{visit disp}} = 1/(1 - \Lambda_{eff}) \tag{5.20}$$

The researcher derives the other performance measure where $\mathrm{Ex}_{\text{visitdisp}}$ represents the number of visit(s) to the dispatcher. That is

$$W_q^i = \frac{\sum_{k=1}^{r} \rho_k / \mu_k}{(1 - \sigma_{i-1})(1 - \sigma_i)} * \mathrm{Ex}_{\text{visitdisp}} \tag{5.21}$$

and total waiting time in the queue by their classes is

$$W_q^{Tdisp} = \sum_{i=1}^{r} W_q^i \tag{5.22}$$

5.2.2 MATHEMATICAL MODELLING OF DATABASE QUEUE AS M/M/1

Let the server utilisation, arrival and the service rate of the database be represented as ρ_2, $\mathrm{l}\lambda_{\text{eff}}$, μ_1. Therefore

$$\rho_2 = \frac{\mathrm{l}\lambda_{\text{eff}}}{\mu_1} \tag{5.23}$$

For a steady state, the expected rate of flow into a state is the same as the expected rate of flow out of that state. Therefore, the steady state probability for the database server is given as

$$(\lambda_{eff} + \mu_1)\ P_n = \lambda_{eff} P_{n-1} + \mu_1\ P_{n+1} \tag{5.24}$$

Where the probabilities of having one or more than one request in the database system is

$$\mathrm{P(dbase)}_n = \left(\frac{\lambda_{eff}}{\mu_1}\right)^n P_0 \tag{5.25}$$

$$\mathrm{P(dbase)}_n = {\rho_1}^n P_0 \tag{5.26}$$

Since the total probability = 1, then

$$\sum_{i=0}^{N} P_i = 1 = \sum_{i=0}^{N} \rho_1{}^n P_0 = 1 = \rho_1{}^n \left[\frac{1-\rho_1{}^{N+1}}{1-\rho_1} \right]^{-1} = 1 \quad (5.27)$$

and for the database server it is given as

because the queue cannot build up unbounded, where ρ_1 = 1

Using the L'Hospital rule, it follows that

$$P(dbase)_0 = \lim_{\rho_2 \to 1} \rho_2{}^n \left[\frac{-(N+1)\rho_1{}^N}{1-\rho_1} \right]^{-1} \quad (5.28)$$

$$= \left[\frac{N+1}{1} \right]^{-1} \quad (5.29)$$

Combining the situation where ρ_1 = 1 then

$$P(dbase)_0 = \begin{cases} \left[\frac{1-\rho_2{}^{N+1}}{1-\rho_2} \right] \; if \rho_1 < 1 \\ \left[\frac{N+1}{1} \right]^{-1} if \; \rho_1 = 1 \end{cases} \quad (5.30)$$

This implies that for all values of n, where n = 0,1,2,3,….,N

$$P(dbase)_n = \begin{cases} \left[\frac{1-\rho_1{}^{N+1}}{1-\rho_1} \right] \rho^n \quad if \rho_1 < 1 \\ \left[\frac{N+1}{1} \right]^{-1} if \; \rho_1 = 1 \end{cases} \quad (5.31)$$

In this chapter, ρ_1 is less than 1. Therefore, the expected number of requests in the database system will be:

$$\mathrm{E(webdbase)} = \left[\frac{1-\rho_1{}^{N+1}}{1-\rho_1} \right]^{-1} \rho \left[\frac{(1+\rho_1{}^{N+1}) - (N+1)\rho_1{}^N \; (1-\rho_1)}{[1-\rho_1]^2} \right] \quad (5.32)$$

Unlike most authors, for example see [11], The database waiting time is re-engineered as

$$\mathrm{Wq_{dbase}} = \left(\mathrm{Ws_{disp}} - \frac{1}{\mu_1} \right) * \mathrm{Ex_{visitdbase}} \quad (5.33)$$

$$\mathrm{Ws_{dbase}} = \frac{\mathrm{E(LS_{disp})}}{\mathrm{l\lambda_{eff}}'} * \mathrm{Ex_{visitdbase}} \quad (5.34)$$

Where $\mathrm{Ex_{visitdisp}}$ represents the number of visit(s) to the the database and is given as

$$\mathrm{Ex_{visit\ dbase}} = \frac{1}{1 - \mathrm{l\lambda_{eff}}'}$$

5.2.3 MATHEMATICAL MODELLING OF THE PRIORITISED WEB STATION QUEUE AS M/M/c/Pr

The M/M/1/Pr could only work for a single server prioritised system. The M/M/c/Pr is therefore proposed because it caters for more than one server machine. This is governed by identical exponential distributions for each priority at each of the c channels within a station. As already mentioned, the service rate is equal.

therefore

$$\rho_{k=} \frac{\lambda_k}{c\mu_k} \quad (1 \leq k \leq r) \quad (5.35)$$

and

$$\sigma_k = \sum_{i=1}^{k} \rho_k \quad (\sigma_0 \equiv \rho = \lambda/c\mu) \quad (5.36)$$

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

$$W_q^{(i)} = \sum_{k=1}^{i-1} E[S_k'] + \sum_{k=1}^{i} E[S_0] \quad (5.37)$$

S_k is the time required to serve n_k consumers of the kth priority in the line ahead of the consumer and S_k' is the service time of the n_k' 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;

$$E[\mathrm{S}_0] = \Pr\left(\begin{matrix}all\ servers\ are\ busy\ within\ a \\ service\ station\end{matrix}\right).\mathrm{E}[\mathrm{S}_0|all$$

$$server\ are\ busy\ within\ a\ service\ station$$

$$= \left(\sum_{n-c}^{\infty} P_n\right] \Big) \frac{1}{c\mu} = P_0 \left(\sum_{n-c}^{\infty} \frac{c\rho^n}{c^{n-c}\, c!}\right) \frac{1}{c\mu} = \frac{P_0(c\rho)^c}{c!(1-p)} \tag{5.38}$$

$$= \frac{(c\rho)^c}{c!(1-\rho)(c\mu)} \left[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]^{-1} \tag{5.39}$$

but

$$E[\mathrm{S}_0] = \rho \sum_{k=1}^{r} \frac{1}{u_k} \frac{\rho_k}{\rho} = \sum_{k=1}^{r} \frac{\rho_k}{u_k} \tag{5.40}$$

$$W_q^{(i)} = \frac{E[\mathrm{S}_0]}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.41}$$

$$W_q^{(i)} = \frac{\sum_{k=1}^{r} \frac{\rho_k}{u_k}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.42}$$

therefore

$$W_q^{(i)} = \frac{\left[c!\,(1-\rho)(c\mu)\sum_{n=0}^{c-1}(c\rho)^{\frac{n-c}{n!}} + c\mu\right]^{-1}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.43}$$

$$W_{qst}^{(i)} = \frac{E[\mathrm{S}_0]}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.44}$$

$$= \frac{\left[c!\,(1-\rho)(c\mu)\sum_{n=0}^{c-1}(c\rho)^{\frac{n-c}{n!}} + c\mu\right]^{-1}}{(1-\sigma_{i-1})(1-\sigma_i)} \tag{5.45}$$

The expected time taken in a service station is

$$W_{qst} = \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_{qst}^{(i)} \tag{5.46}$$

The overall average time taken in j service stations is

$$W(ave)_{qst} = \frac{\sum_{n=1}^{j} \left[\sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_q^{(i)} \right]}{j} \tag{5.47}$$

5.2.4 MATHEMATICAL MODELLING OF PRIORITSED DISPATCHER-OUT AS M/M/1/Pr

The Dispatcher–Out is similar to the Dispatcher-in, therefore the waiting time is derived using eq. 5.1 to 5.18 and 5.20. Consequently,

$$W_q^{Tdispout} = \sum_{i=1}^{r} W_q^i \tag{5.48}$$

Two things are of importance in this chapter; the waiting time of each class of consumer and the total waiting time experienced by consumers. The first is the performance impact of the non-preemptive model on consumer waiting time while the second is on the determination of the optimal solution.

The generalised total waiting for class *n* priority is

$$W_q^{tot^n} = W_{qdin}^i + W_{qst}^i + W_{qdout}^i \tag{5.49}$$

while the total waiting time experienced in the queue by the whole classes that is class 1….m is

$$w_q = \sum_{n=1}^{m} W_q^{tot^n} \tag{5.50}$$

5.3 FORMULATING THE NON PRE-EMPTIVE COST MODEL

A cost model is the mathematical equation that converts resource data into cost data. For example, one of the resource data elements is the waiting time of n priority class ($W_q^{itot^n}$) and another one is total waiting time (w_q). This is similar to the cost equation used in chapter three. The goal of the current cost model is to attempt to add to the body of knowledge towards achieving optimal service level and consumer satisfaction. This endeavour starts from estimating correctly the Expected Total Cost incurred by the provider (ETC). The terms are:
ETC(x) = Expected Total Cost per unit time.

EOC(x) = Expected Operating Cost of the cloud E-Market servers per unit time.

EWC(x) = Expected Waiting Cost by consumers per unit time.

K_n = Cost value of waiting on the queue by class *n*.

c = number of server machine(s) working.

In this study, the cost of \$1 is assigned as operating cost of energy and $(\mathrm{c}/10)$ is the servicing cost.

The cost function is defined as

$$\mathrm{ETC(x)} = \mathrm{EWC(x)} + \mathrm{EOC(x)} \tag{5.51}$$

$$\mathrm{ETC(x)} = \mathrm{EWC(x)} + (\mathrm{c}/10 + 1) \tag{5.52}$$

The $\mathrm{EWC(x)}$ for example class n is:

$$\mathrm{ETC(x)_n} = k_n * W_q^{itot^n} \tag{5.53}$$

Therefore the total cost of waiting for m classes of consumers is given as

$$\sum_{n=1}^{m} k_n * W_q^{itot^n} \tag{5.54}$$

That is, the cost of waiting of each class (n) multiplied by the waiting time of that class. In this experiment m is given the value 5.

Then

$$\mathrm{ETC(x)_n} = \mathrm{W_q} + \left(\frac{\mathrm{c}}{10} + 1\right) \tag{5.55}$$

from Eq.5. 51

$$\mathrm{ETC(x)} = \sum_{n=1}^{m} k_n * W_q^{itot^n} + \left(\frac{\mathrm{c}}{10} + 1\right) \tag{5.56}$$

5.4 SIMULATION AND NUMERICAL VALIDATION

This aspect of the study is divided into two. The first is the performance evaluation of the Cloud E-Marketplace in the context of Non-preemptive policy and the second is the determination of the optimal service level. This leads to the two experiments reported here. The first is the waiting time of the Non – Preemptive classes to be used for evaluation and the second is the waiting time measures when the service stations are varied to determine the optimal server solution. The study first validates the mathematical solution with the simulation to ascertain the degree of correctness.

This is done by considering five classes of consumers' arrival. These are represented by five arrival processes $\lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_1 and\ \lambda_5$ where $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5$ and $\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5$.

The research uses the wolfram Mathematical 9.0 as the mathematical tool and arena 14.5 as the simulator for the validation of results.

Then $\lambda = 100, 200, 300, .., 970$ respectively and c = 2 in each of the service stations. the experiment is then simulated using Arena Discrete event simulator version 14 with the same values to ascertain the degree of variability. This

simulation was run with replication length of 1000 in 24 hours per day with base time in hours and it was replicated 5 times. The service rate was set to 0.001 for the dispatcher-In and 0.0005 for each of the servers in the web Queue stations and the dispatcher-Out. A server with a low service rate of .0002 was used for the database server because of its randomness.

In addition, a similar experiment with the same configurations but under a non-priority discipline using FCFS discipline was set up. The illustration of this model is shown as numerical examples and the impact on the five performances of the priority classes are analysed in the results and discussion section.

The second experiment consisted of processes modelled by setting the total inter arrival time of the five classes to .33 seconds. The service time for the dispatcher-In and each of the web station servers was given the value of 1.2 seconds. The buffer capacity of the dispatcher-In and dispatcher-Out was set to 1000 to reduce balking while 400 was used as the maximum buffer capacity in each of the other servers. Because of the priorities issue, the model classes had to be prioritised based on cost. The reason behind this is that the cost of waiting for example class 1 consumers should be higher than class 2 because the pay per go of class 1 is higher than class 2. Therefore, $k_{1,}$……, K_5 are given the waiting cost values of \$5, \$4, \$3, \$2, \$1 for the purpose of this experiment but the generalised idea is that $k_1 > k_2$,……, $>K_n$. The base time unit is set to seconds respectively. Each experiment is performed with 10 replications for an average of 49949,0000 seconds. The experiment started with six server machines with each web station queue having two server machines based on the researcher's model in Fig. 1. At the end of each experiment, the server machines are increased in each web station by one and the waiting time in the system is being recorded. The record the waiting time of only the web station is recorded since the dispatcher-In, database and Dispatcher-Out servers are constant. This performance measure is the used to derive the cost function. Details of the

performance results are in Tables 5.1, 5.2 and Figures 5.2 -5-5. Detailed discussion on the results is presented in section 5.5.

5.5 RESULTS AND DISCUSSION

The analytical results and that of the simulation are compared and the results are shown in Table 5.1 and Fig.5.2 respectively. The degree of variation is hardly noticeable until when $\rho \to 1$ but the coefficient of variation is very small and less than unity, which is why in Fig 2, the analytical colour overshadowed the simulation colour at the initial stage. Table 5.2 provides the results of the waiting time distributions of the simulation under the Non Preemptive Priority and the exogenous Non Priority using the FCFS policy. The simulation results of Non preemptive are represented as C1s-C5s and the Non priority (FCFS) ones as C1-C5. The first observation reveals that the total waiting time simulation results of both disciplines are the same. This implies that the total waiting time on the queue is independent of the service discipline. However, the waiting time distributions of the Non Preemptive classes differ while those of Non Priority have equal distributions.

The performance result from the simulation revealed that at the initial stage when the server utilisation is below 0.5, the performance difference between the five classes is not fully noticed. As the server utilisation increases, that is the number of consumers increases in the Cloud E-Marketplace, the performance differentiation becomes very noticeable: class 1 priority demonstrated the minimum waiting time, followed by class two; while class five had the highest waiting time. All four classes had better performance at the expense of the fifth class. This is shown in Fig. 5.3 based on the information obtained from Table 5.2.

Both priority and non-priority models were compared using the First Come First Serve policy. Both displayed insignificant performance differentiation below 0.4 but as the $\rho \to 1$ where $\rho < 1$, then four classes out of the five classes observed

had better waiting time performances over the conventional exogenous non priority model. Though this is at the expense of the fifth class, this model will be good in an environment where the cost model is prioritised based on consumers' classes. This is depicted in In Fig. 5.4.

Table 5-1: Simulation and Analytical

P	Sim	Analytical
0.1	0.000681	0.000681
0.2	0.001502	0.001502
0.3	0.002612	0.002612
0.4	0.004006	0.004006
0.5	0.00592	0.00592
0.6	0.008751	0.0089
0.7	0.013447	0.013447
0.8	0.021933	0.023019
0.9	0.047009	0.0496

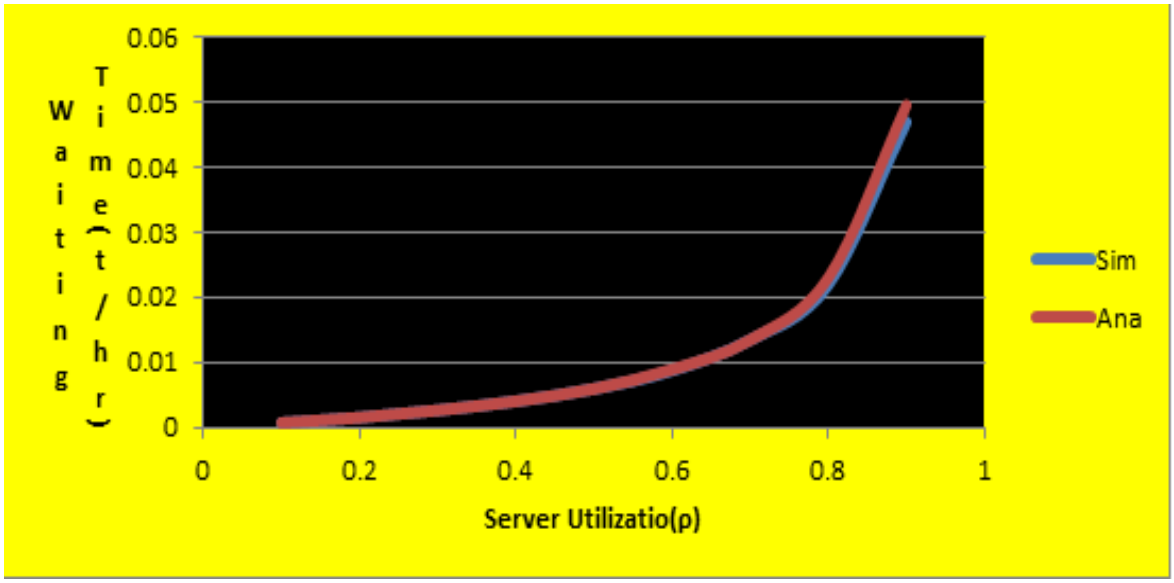

Fig. 5.1: Simulation and Analytical

Table 5-2: Detail Waiting time Results of the Non Pre-emptive Priority and the Non priority

						Total						Total
ρ	C1s	C2s	C3s	C4s	C5s	N PP	C1	C2	C3	C4	C5	NP
0.1	0.00012286	0.000136	0.0001365	0.00014099	0.00014481	0.00068116	0.00013139	0.00014131	0.00013628	0.00013645	0.00013547	0.0006809
0.2	0.00025534	0.00027293	0.00029912	0.00032435	0.00035006	0.0015018	0.00030233	0.00030177	0.00030091	0.00030183	0.00029467	0.00150151
0.3	0.0003898	0.0004408	0.00051265	0.00058382	0.00068507	0.00261214	0.00051544	0.00051535	0.00052901	0.00052627	0.00052757	0.00261364
0.4	0.00053658	0.00062279	0.00075588	0.00092363	0.00116738	0.00400626	0.00080211	0.00078963	0.00079754	0.00080607	0.00080762	0.00400297
0.5	0.00067624	0.00085014	0.00107412	0.00139387	0.00192553	0.0059199	0.0011781	0.00118155	0.00119661	0.00118224	0.00118471	0.00592321
0.6	0.00082978	0.00108347	0.00146567	0.00208695	0.00328506	0.00875093	0.00175392	0.00175024	0.00174231	0.0017576	0.0017542	0.00875827
0.7	0.00099346	0.00135409	0.00195375	0.00315464	0.00599096	0.0134469	0.00269462	0.00267701	0.0026877	0.00269428	0.00269727	0.01345088
0.8	0.00115011	0.00164642	0.00258144	0.00475185	0.01180286	0.02193268	0.00437082	0.00436584	0.00437259	0.00438619	0.00439383	0.02188927
0.9	0.00131086	0.00200109	0.0034308	0.00749416	0.03277201	0.04700892	0.00937154	0.00942055	0.00938388	0.00940722	0.00946238	0.04704557

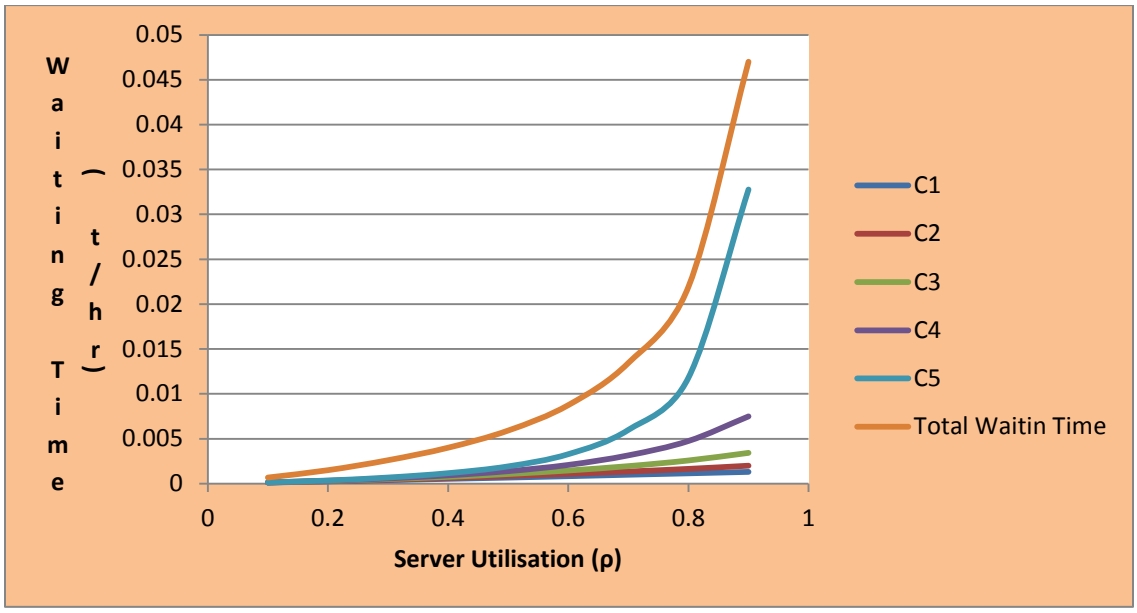

Fig. 5.2: Performance of the Five Classes Waiting with the Total Waiting Time

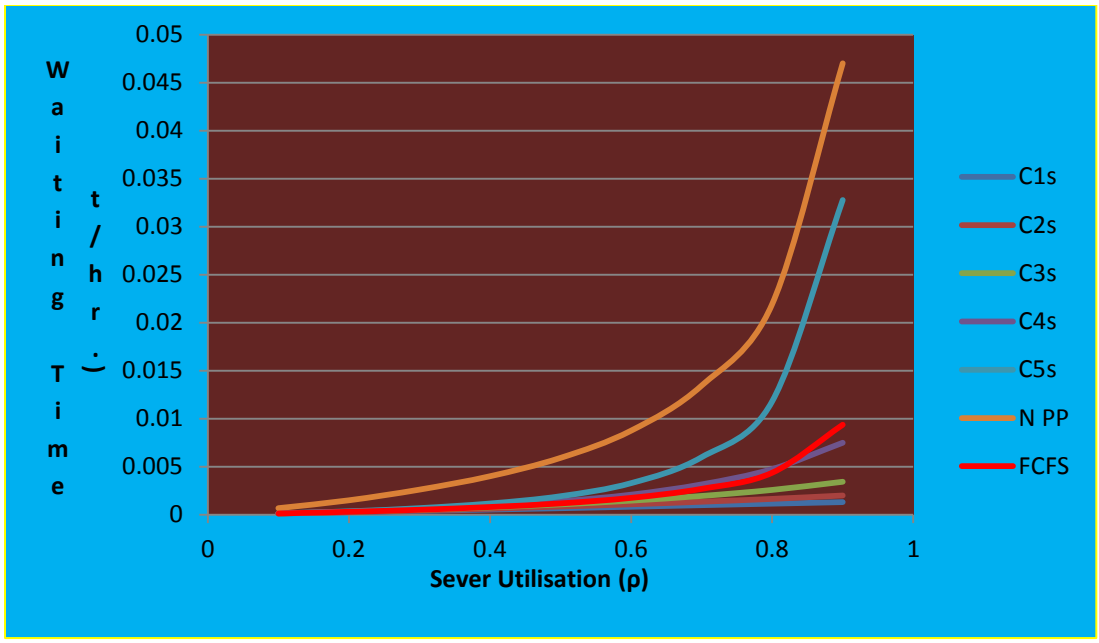

Fig. 5.3: Performance of Non Pre-emptive Priority and Non Priority

The results of the second experiment are shown in Tables 5.3, 5.4 and Fig. 5.5 and 5.6 respectively. In Fig. 5.5, in the second experiment, a total of 90,056 consumers from all five classes arrive with an average of 18011.2 from each class and are processed (Tot Web In/Out) by web station1 and 2 servers. The database server automatically generates 4524 requests for collecting statistical data, which is the re-engineering aspect of this model and in line with the model of Figure 5.1. This is shown in Figure 5.5. The waiting time of each of the classes in the system and the total server machines used is depicted in Table 5.3. The ETC(x) is based on Eq. 5. 56. From Table 5.3 it is observed that as the number of server machine increases the waiting time reduces. One major problem being addressed is the determination of the service level agreement because the researcher is of the opinion that the SLA of an individual class should differ since they have different service offerings. This research did not cover the formulation of SLA. The focus is limited to the determination of optimal service level that the provider can provide when the SLA is breached. The generalised algorithm for the SLA is left as future work.

In the final experiment, the results of the expected waiting cost, the expected operational cost and the total cost of each experiment were obtained. The optimal level is achieved at the point where the expected total cost has the minimum value and this is at the point where 6 server machines are used. These are shown in Table 5.4 and Fig. 5.6. Any additional server(s) will bring additional cost which will minimise profit.

In the previous chapter the focus was on the performance evaluation of Cloud E-Marketplaces in a non-preemptive environment using two service offerings. In this chapter, the focus is on two things: the first is the performance evaluation of Cloud E-Marketplaces in a non-preemptive environment using two or more than two service offerings (generalised) and the second is the determination of optimal service level that satisfies provider cost and consumer waiting time in non-preemptive E-Marketplace. Most work done in the literature for example [129][26][25][130] is related to networking. Even the ones that are closely related to this work, for example[109] [110], are done by prioritising only at the point of entry and not generalised. This research is differentiated from the previous chapter and the literature based on (1) different theoretical concept (2) different simulation concept and (3) the focus which is on determination of the optimal service level that balances provider cost and consumer waiting time in the context of Cloud E-Marketplaces. All these, to the best of the researcher's knowledge have not appeared in the literature in the context of Cloud E-Marketplaces.

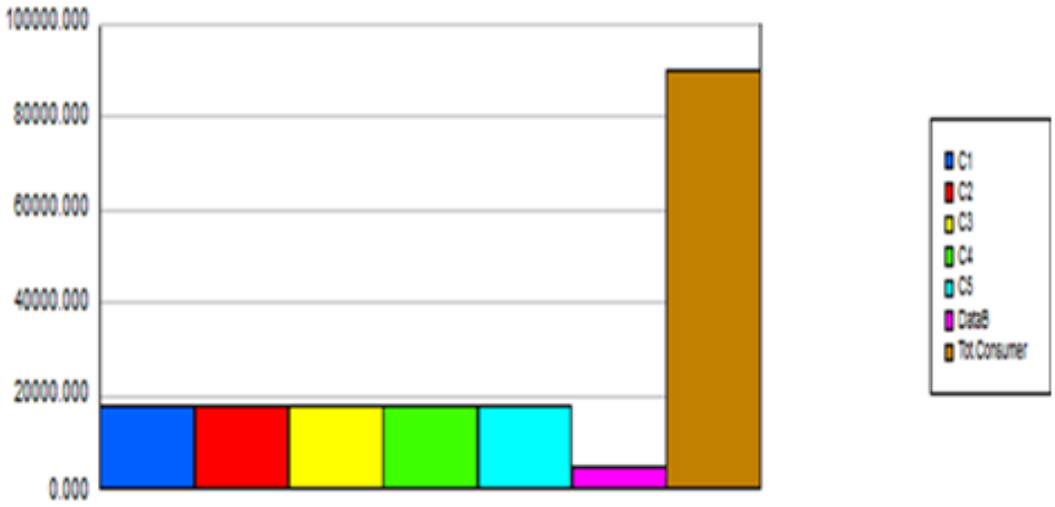

Fig. 5.4: Total Consumers Processed and Randomly Generated Requests

Table 5-3: Service Level (SM) and the Waiting Time of Five Non Pre-emptive Classes

SM	C1	C2	C3	C4	C5
4	0.00104013	0.00162461	0.00293347	0.00700013	0.03552028
6	0.00025746	0.00034349	0.00045909	0.0006603	0.00105907
8	0.00001609	0.00001873	0.00003548	0.00002621	0.00020861
10	0.00001509	0.0000177	0.00001944	0.00002521	0.00002861
12	0.00000334	0.00000423	0.00000398	0.00000456	0.00000542
14	0.0000008	0.00000081	0.00000082	0.0000009	0.00000093

Table 5-4: Total Cost

SM	EWC	EOC	ETC
4	0.31061314	1.4	1.710613
6	0.02031525	1.6	1.620315
8	0.00204844	1.8	1.802048
10	0.00081385	2	2.000814
12	0.00016775	2.2	2.200168
14	0.00003373	2.4	2.400034

Fig. 5.5: Total Cost - Service Level

5.6 CHAPTER SUMMARY

This chapter extended the existing and widely adopted theories to a generalised Non Preemptive model. Two issues were the focus of the chapter: the first was the performance impact of the Cloud E-Marketplace on consumer waiting Time in the context of different service offerings; the second was the determination of

optimal service level in the same context. On the performance impact, as the server utilisation tends toward large numbers, the results revealed better waiting time performances over the conventional exogenous non priority model in four out of the five classes observed in this experiment. Though this was at the expense of the fifth class, this model will be of benefit in an environment where the cost model is prioritised based on the class of consumer.

On the optimal solution issue, this was achieved at the point where the expected cost has the minimal value. Although, the result of the experiment showed the same optimal point for both non preemptive priority and the non-priority model but where consumers' costs are prioritized the Non preemptive priority model will be more efficient and profitable apart from the optimal server machines which have minimised consumer waiting time.

CHAPTER SIX

PERFORMANCE MODELLING OF THE CLOUD E-MARKETPLACE USING DYNAMIC CONTROL MODEL

One major issue that has been of a great interest in Cloud E-Marketplaces is the effective management of resources [131] [132]. To satisfy consumers, some E-Market providers adopt the resource overprovisioning model[133]. However, the researcher knows from existing knowledge that resource over-provisioning could lead to a largely sub optimal utilisation of the hosting environment thereby leading to unnecessarily wasteful costs [134][135]. Apart from this, the report of Industrial Development Corporation (IDC) 2012 on power consumption identified three challenges being faced by Cloud E-Market providers, namely: Skyrocketing power consumption and electricity bills, serious environmental impact, and unexpected power outages. With respect to server consolidation, the report in [136] revealed that about $45 billion was spent on server management and administrative costs in 2012 (see Fig. 6.1). Therefore, effective resource management of these Server Machines (SMs) in terms of cost minimisation is imperative. This chapter addresses this through the use of a Dynamic Control System (DCS) that changes the number of server machines dynamically based on consumers' waiting time. A comparison with the DCS model is made with the fixed servers model based on cost. Two issues are addressed in this chapter: the first is to determine which of the two models provides an optimal solution in terms of server management and the second is the profitability concept of the experiment in the model that produces an optimal solution. The results proved the DCS to be cost effective with a higher Cost-Benefit Ratio (CBR) over the fixed servers.

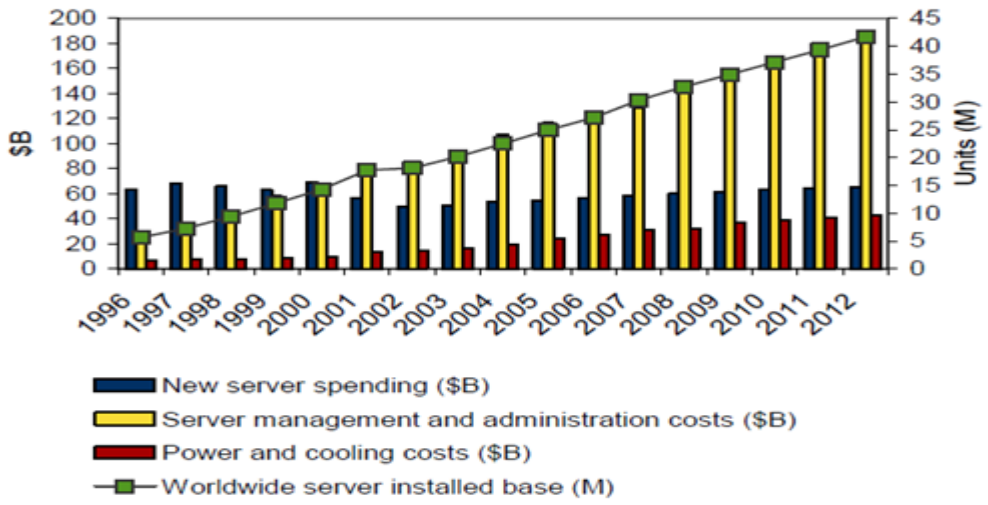

Fig. 6.1: Worldwide IT Spending on Server, Power and cooling and Management/Administration, 1996-2012

6.1 INTRODUCTION

In the previous chapters (Chapter 3, 4 and 5), the Cloud E-Marketplaces was modelled after both the Non priority and Non pre-emptive priority systems. These previous chapters focused on the use of the optimisation approach to study the descriptive model. By descriptive model this mean a model that describes what the real world should be. The Descriptive model has also been referred to severally as Optimal Design and Control of queues, and the Economic or Statics model [47] [137]. It determines the optimal system parameters, for example, the optimal number of server machines needed to satisfy consumers. One major issue that was not considered fully is that of server optimisation in terms of effective management based on good policy.

In this chapter, the focus is on the use of a prescriptive model which is an optimal control mechanism that prescribes what the real situation will be as against the previous descriptive model [47] [137]. The intention is to find the optimal situation at which to aim. The approach to achieving this is to develop a model that will dynamically change based on the input and the history (hysteresis) of the incoming consumers. The research idea is to use the appropriate policy that will control the system economically by optimising cost.

Three policies are studied: the N [138] , T [139] and D [140]. The N policy is selected because it is analytically easier to deal with [137]. By N policy this means a policy that enables the servers to be turned on when the system is busy and turned off when it is idle [141].

The remainder of this chapter is organised as follows: Section 6.2 discusses the architecture of the Dynamic Control System (DCS). The required measures of effectiveness are derived in section 6.3 and the profitability of the system is discussed in section 6.4 while the experimental set with the results and discussion are covered in sections 6.5 and 6.6 respectively. The chapter is summarised in section 6.7.

6.2 ARCHITECTURE OF THE DYNAMIC CONTROL SYSTEM (DCS)

The architecture consists of a fixed number of servers for operation (M1, M2, M3) and then provides a limited set of backups as the reservoir (Reservoir Room) with upper limit (S). When the waiting time of the service consumer reaches a point for example N, then the Dynamic Control System (DCS) transfers the incoming consumers from the dispatcher-In to the reserve centre. As soon as the waiting time reduces to N-1, it goes back to the main centre. The architecture is shown in Fig. 6.2. This experiment is carried out under two conditions. These are:

- When the total number of used servers are fixed
- When some are fixed with variable service stations as shown in Fig.6.2 (M1..M3 are fixed and M4..M5 varies)

As said earlier, two things the research aims to address in this chapter are: firstly to determine which of the two models provides an optimal solution in terms of server management and secondly to determine the profitability concept of the model that produces an optimal solution based on the experiment conducted.

To achieve this aim, the following steps are followed:

i. Determine the Measure of effectiveness

ii. Subject the two condition to the same test by

Max

The Server Utilisation

Subject to

Waiting time < predetermined Target

iii. Determine the optimal number of server machines that satisfied the maximisation problem in ii under the two conditions.

iv. Compare and contrast the result obtained using the Cost-Benefit Analysis (DCS).

v. Find the profitability of using this model over an ordinary fixed model based on the experiment conducted using the Cost-Benefit Ratio (CBR).

The analytical solution of this work is based on the use of queuing theory as the proof of concept. To get the mathematical concept behind this work various literature was searched [142] [143] [138][144][145][146][147] [148] [149] and [150]. The result of the researcher's search enables us to base the mathematical concept on the work of Moder in [143]. Moder compares very well with the researcher's DCS except that the DCS context is in the Cloud E-Marketplace. DCS is unique in that it incorporates concepts such as CBA, CBR and profitability. Secondly, it allows hysteresis in the shifting mechanism. The idea is that the number of servers is a random variable since it is a function of the queue length. In [143], the author determines the measure of effectiveness through the following steps (see the full equation in Appendix A):

- Determine the list of admissible States
- Determine the steady state equations
- Solve each of the steady state equations
- Calculate the measure of effectiveness.

The following parameters are used in this chapter

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 $\boldsymbol{\delta} < S < \infty$

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

$\gamma = \rho/\delta$

6.3 MEASURES OF EFFECTIVENESS

Measures of effectiveness are stochastic random variables that one might like to know about the queuing system. These include the consumer's waiting time, length of the queue, idle time. This chapter uses some of these measures as the fundamental parameters needed to derive the profitability and the CBR. These measures of effectiveness are derived from the mathematical proof in Appendix A. the researcher uses the idle time ($idle_{time}$), Interrupt time ($Inter_{time}$) the length of the queue L_q and the waiting time w_q to achieve the researcher's aim. These are

$$idle_{time} = \sum_{s=0}^{\delta-1}(\delta - s)P_{0\,s} = P_{0\,0}\sum_{s=0}^{\alpha-1}\frac{(\delta-s)\rho^s}{s'} \qquad (6.1)$$

$$Inter_{time} = \sum_{s=\delta}^{S-1} P_{N-s} = \gamma^{N-1}(1-\gamma)P_{o\,o}/(1-\gamma^{N-v}) \sum_{s=\delta}^{S-1} \left({}^{\rho^s}/_{S'} \right) \qquad (6.2)$$

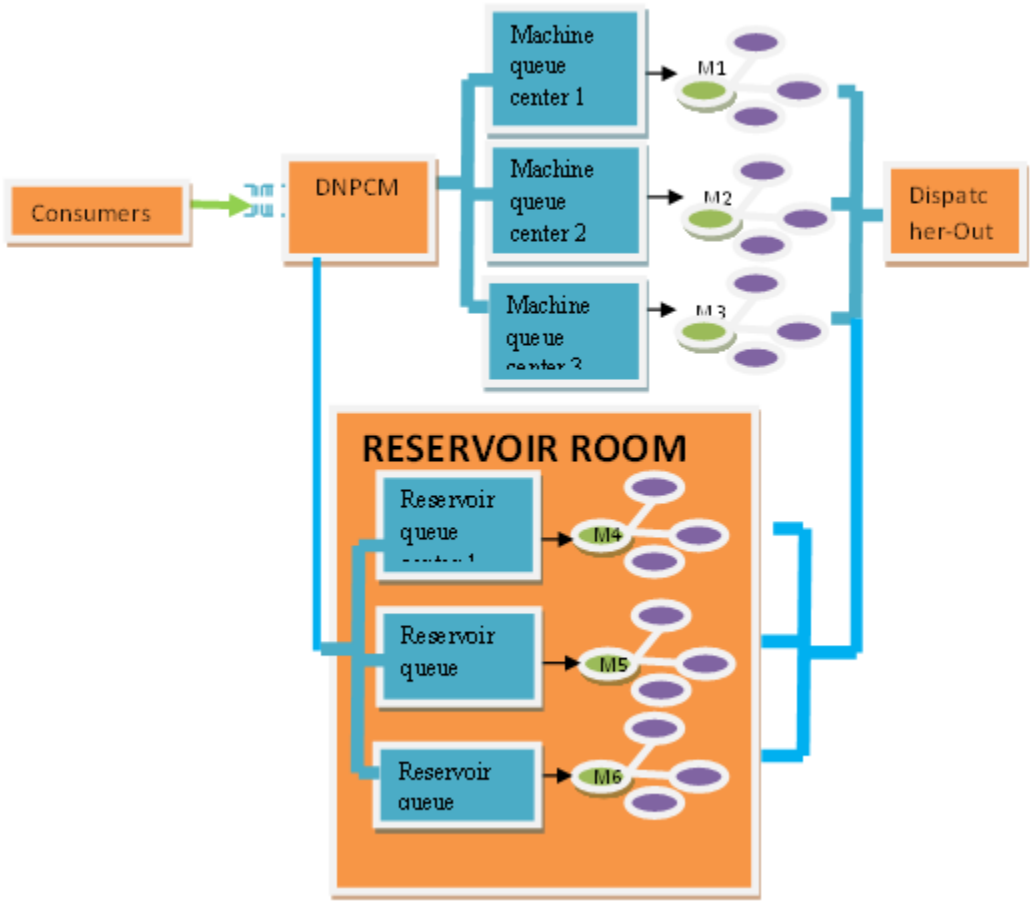

Fig. 6-.2: Architectural diagram of the Dynamic Control System (DCS)

$$L_q = \frac{P^{\delta}}{1-\gamma^{N-v}}\left[\frac{\rho^{\delta}[\gamma-(v+1)\gamma^{v+1}+v\gamma^{v+2}](1-\gamma^{v})}{\delta'(1-\gamma)^2}\right.$$

$$+\frac{\rho^{\delta}}{\delta'}\left\{\frac{\gamma^{N}(N\gamma-\gamma-N)+\gamma^{v+1}(v+1-v\gamma)}{(1-\gamma)^2}\right.$$

$$\left.-\frac{1}{2}\gamma^{N}[N(N-1)-v(v+1)]\right\}$$

$$+\frac{1}{2}\,\gamma^{N-1}(1-\gamma)[N(N-1)-v(v-1)]\left\{\sum_{s=\delta}^{S-1}\frac{\rho^{s}}{s'}\frac{\rho^{\delta}}{\delta'}\right\}$$

$$+\frac{\rho^{S}\gamma^{N-1}(1-\gamma)}{2S'(S-\rho)^3S^{N-2}}\{S^{N-1}(S-\rho)^2\big[\big((N-1)\big)\big((N-2)\big)$$

$$-v(v-1)\big]-2S^{v}\rho^{N-v}[\rho(N-2)-S(N-1)]$$

$$-2S^{N-v-1}\rho^{v+v}[Sv-(v-1)\rho]$$

$$+2(S^{N-v}$$

$$\left.-\rho^{N-v})S^{v}\big[S\big(S(N-1)-\rho(N-2)\big)\big]\}\right] \tag{6.3}$$

$$w_q = \frac{L_q}{\lambda} \tag{6.4}$$

then the researcher

$$\begin{aligned} &Max \\ &1-idle_{time}(server\ \ Utilization) \\ &subject\ to \\ &W_q < k \end{aligned} \tag{6.5}$$

where *k* is the predetermined waiting time target.

6.4 PROFITABILITY OF THE DCS

One approach to determine the profitability of this model is to use the waiting cost difference between the fixed and the variable operation and see if it is higher than that of the waiting cost when additional server (s) is/are added [151][152] . If this is true, then it is profitable. To do this, if the number of

consumers under the M/M/n (Fixed) = E(L). Then if a fixed number of servers is in operation and the expected number of consumers on the queue is $E(L_q)$ under the variable one if another server is added when n> N, then system becomes profitable to use the second sever *iff*

$$X_n\left(\mathrm{E(L)} - E(L_q)\right) > X_{n+1}\ Prob.(n > N) \tag{6.6}$$

where $X_n$and X_{n+1} are the associated cost for the difference and that of the probability that the number of consumer is greater than N respectively.

If another server added again when n>L with associated cost X_3 then it becomes profitable to use the third server when

$$X_1\left(\mathrm{E(L)} - E(L_q)\right) > X_2\ Prob.(n > L) \tag{6.7}$$

$$= X_1\left(\mathrm{E(L)} - E(L_q)\right) > X_2\ Prob.(N < n > L) + X_3\ Prob.(n > L \tag{6.8}$$

In this study, the idea is slightly different from these because the cost in a typical E-Marketplace is not likely to be based on waiting cost alone. In addition, the benefit accrued to this is also not likely to be based on one parameter alone. In the researcher's proposed solution the DCS becomes profitable if the opportunity cost for adding (an) additional server(s) is less than the benefit incurred when an additional number of servers is/are added. The Cost-Benefit Analysis (CBA) is used as the researcher's solution approach [153].

In the model, the opportunity cost is based on the following parameters:

i. The average waiting time cost (W_{qave}) incurred for staying longer in the DCS model than the fixed one
ii. Interrupt Time cost ($Inter_{time\ ave}$) that occurs when additional servers are added.

The benefits are based on the following parameters:

i. Gain or benefit accrual in unused server(s) machines

ii. Gain or benefit accrue in Energy consumed and the

iii. Utilisation gain or benefits

Associated costs (X_1, X_2, X_3, X_4, X_5) are given to the parameters based on the level of importance. For example the cost allocated to waiting time may not be the same with that of energy used or unused.

Therefore, the

$$Total\ _{cost} = X_1 W_{qave} + X_2 Inter_{time\ ave} \tag{6.9}$$

$$= X_1 [w_{qDCS} - w_{qfixed}] + X_2 \left(\left(\sum_{i=1}^{r} inter_{time} \right) / r \right) \tag{6.10}$$

$$\begin{aligned} Total_{benefit} = {} & X_3 \left[serverused_{fixed} - serverused_{DCS} \right] \\ & + X_4 \left[Energyused_{fixed} - Energyused_{DCS} \right] \\ & + X_3 \left[Utilisation_{DCS} - Utilisation_{fixed} \right] \end{aligned} \tag{6.11}$$

The DCS model becomes

- Marginal iff $Total\ _{cost} = Total_{benefit}$
- Profitable iff $Total\ _{cost} < Total_{benefit}$ and
- Non profitable iff $Total\ _{cost} > Total_{benefit}$

The Cost-Benefit ratio (CBR) is given as

$$CBR = \frac{Total_{benefit}}{Total_{cost}} \tag{6.12}$$

Given *n* number of experiments conducted and the CBR for each experiment is E_1, E_2 and E_3, then the optimal profit occurs in the experiment where CBR has the highest value.

6.5 EXPERIMENTAL SETUP

This experiment is conducted with the assumption that the arrival process is a random process and the service time is exponential. Also, the cost of using each server is the same. To achieve this, the experiment is set up under two conditions:

i. When the total number of used servers is fixed.
ii. Under the Dynamic Control system (DCS) in which some are fixed and some are varied as shown in Fig. 6.2 (M1, M2, M3 as fixed and M4..M5 and varies).

Under each of the above conditions, three experiments were conducted. In all the experiments, 5 server machines were used for both conditions. The service time was homogenous and set to 0.5. In each of the experiments, the fixed and the variable experiment were subjected to the same waiting time condition (W<=K= 0.0006 sec). The arrival is a random process and the service rate is an exponential process. Under the first condition, all used servers are fixed (5 servers) and under the second condition, two are fixed while the others are varied (3) and the shift point N is set to 5.

Arena simulator is used as the simulator. Under the two conditions, the arrival time $\left({}^{1}/_{\lambda}\right) = 0.4, 0.3, 0.2$ respectively. The associated costs ($) X_1, X_2, X_3, X_4, and X_5, were set to 4, 4, 2, 2, and 4 respectively. The value of 4 watts was given as the energy consumed when the server is in operation. Each experiment was run for an average of 47 hours and replicated 10 times in the interests of accuracy. The results are in Fig. 6.3-6.6 respectively and explanations are given in the results and discussion section.

6.6 RESULTS AND DISCUSSION

In each of the experiments conducted, a total of 30,000 requests were processed when the arrival time ($1/\lambda$) were set to 0.4, 0.3 and 0.2 as shown in Figures 6.3 and 6.4. These two figures show the number of consumers that were processed by each machine under DCS (Fig. 6.3) and the fixed classical model (Fig. 6.4). Under the DCS, It was observed that as the arrival time reduces, the number of servers increases in DCS. In the first experiment, when ($1/\lambda$) =0.4 the maximum number out of the five manned servers that was used was two as against five that were used in the classical fixed method. A gain of three server machines was recorded, which was used to generate the researcher's server benefit based on equation 6.11. Another gain of two servers was recorded in the experiment when ($1/\lambda$) = 0.3 while one was recorded when ($1/\lambda$) =0.2. The reason for the gain in server machines used under the DCS is explained using Fig. 6.5 and 6.6. These figures represent the server utilisation with the number of machines used. In these two figures, in each of the experiments conducted, the instantaneous utilisation of the DCS is higher than the classical fixed model. For example, when the service time was 0.4, the instantaneous utilisation of M1 under DCS was about 0.739 as against the classical fixed one which was about 0.304. Also, when the service time was 0.2 in the second experiment, the instantaneous server utilisation was about 0.990 by server M1 in DCS as against 0.603 in the classical fixed model. That the server was able to be used to its maximum potential was as result of the Dynamic Control mechanism. This accounted for the increase usage of the server machines recorded The DCS has a better performance in terms of server utilisation and optimal server provisioning as against the classical fixed model that over- satisfies the delay requirement.

On the profitability concept of this model, the waiting time, idle time, interrupt time and the number of used servers were recorded. Each server used is allotted 2 watts as shown in Table 6.1. The results obtained from the simulation are shown in Fig 6.1. As the service time increases in each experiment so do the

waiting time and the server utilisation. Unlike the waiting time, the researcher recorded zero for interrupt time on the first experiment under the DCS and in all the fixed server experiments because only two fixed machines worked under the DCS in the first experiment unlike the fixed server model where all the machines were in operation without interrupt.

The output of the experiment recorded in Table 6.1 was used to formulate the Cost-Benefit Table (see Table 6.2) using equations 6.10, 6.11and 6.12. In this Table (Table 6.2) the total cost increases while the benefit decreases as the service time increases as shown in experiments 1, 2, and 3. The reason is due to the high gain of server recorded in experiment 1 when only 2 servers were in operation as against the 5 used in the fixed operation. This also accounted for the high CBR recorded in experiment 1.

One noticeable thing which requires questioning in Table 6.1 is the low server utilisation in experiment 2 compared to experiment 1. The reason is that when $1/\lambda$ =0.3, though the server utilisation in the fixed machine under DCS (See fig. 6.5) were higher than that of experiment 1, but the low server utilisation of the third machine in experiment 2 (M3 indicated in red in Fig. 6.5) accounted for the average drop in server utilisation of experiment 2's computation. Even with this drop, a gain of 2 server machines was recorded.

All three experiments were observed to be profitable due to the Cost-Benefit Ratios that are greater than unity (183.6514: 30.54533: 8.881209) in all the output of all the experiments. The optimal profitability was recorded in experiment 1 with the value of 183.6514. The reason for this is the high gain of the unused servers and the server utilisation in the first experiment. For example, two 2 servers were used under the DCS as against five servers in the fixed rate.

Effective management of Cloud resources is imperative for profit maximisation. This chapter has further looked into how to effectively manage Cloud E-Marketplaces. This is done by designing a DCS model as shown in Figure 6. This model is a prescriptive model, unlike all the researcher's previous models in the previous chapters (2-5) which are descriptive. This allows web applications to be guided by a control mechanism that determines the period of changing from one system to the other. The work of Moder compares very well with the DCS except that the DCS context is Cloud E-Marketplaces. DCS is unique in that it incorporates concepts such as CBA, CBR and profitability. Based on the experiments conducted, the results obtained prove to be cost effective when compared to the fixed server system. For example, 2 server machines were used under the DCS as against 5 under the fixed system. The design of DCS to determine the profitability and the CBA is the core focus of this chapter which, to the best of the researcher's knowledge has not appeared in the literature in the context of Cloud E-Marketplaces. The management of the Cloud E-Marketplace depends on how the available resources like the server machine can be effectively managed.

The Dynamic Control System will have a role to play in server management as a mechanism for maximising instantaneous server utilisation, thereby minimizing the number of servers used.

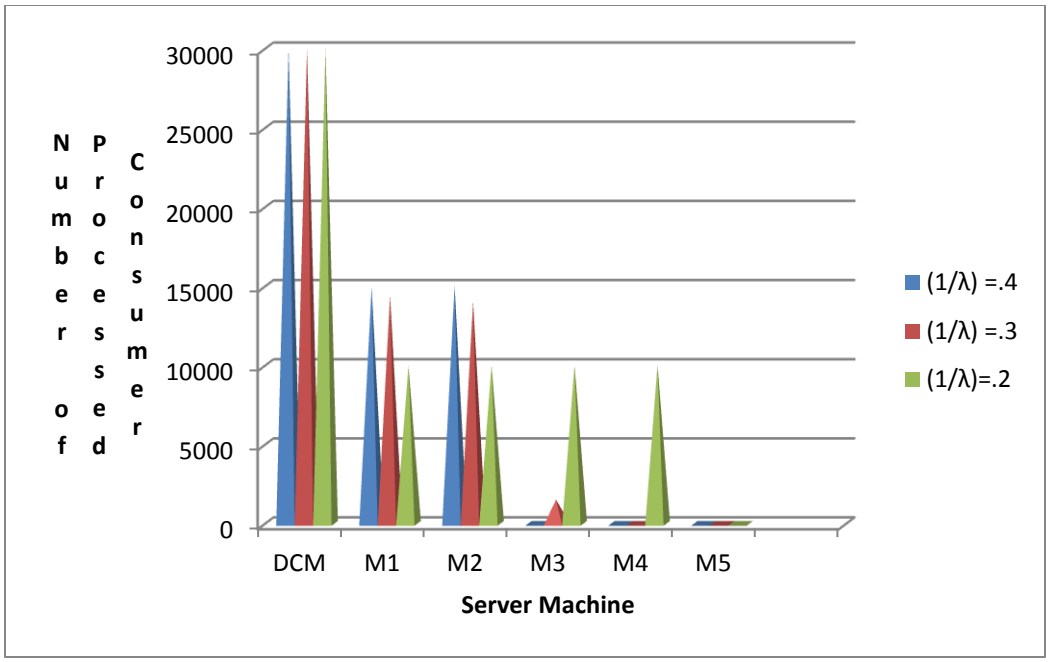

Fig. 6.3: Number of processed consumers under DCS

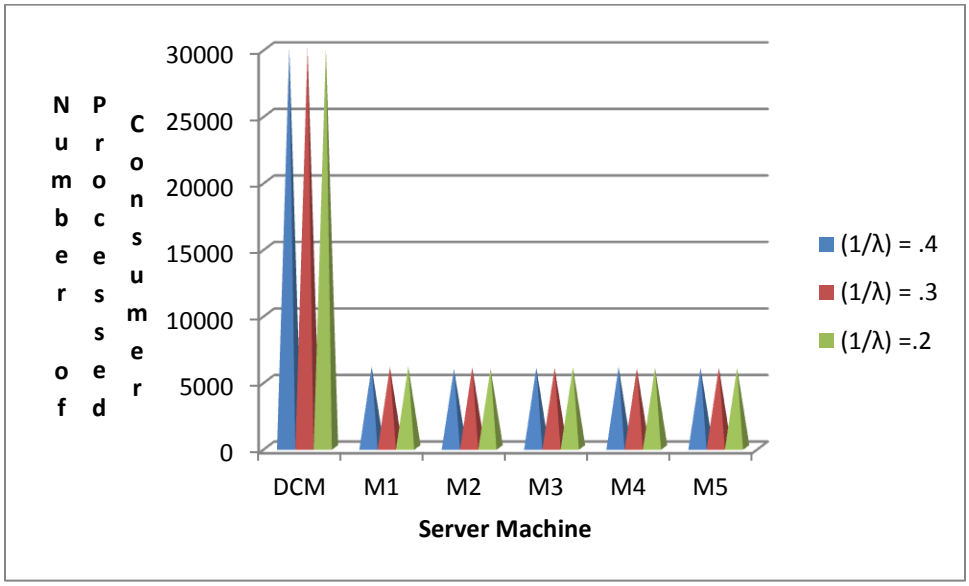

Fig. 6.4: Number of processed consumers under the classical fixed model

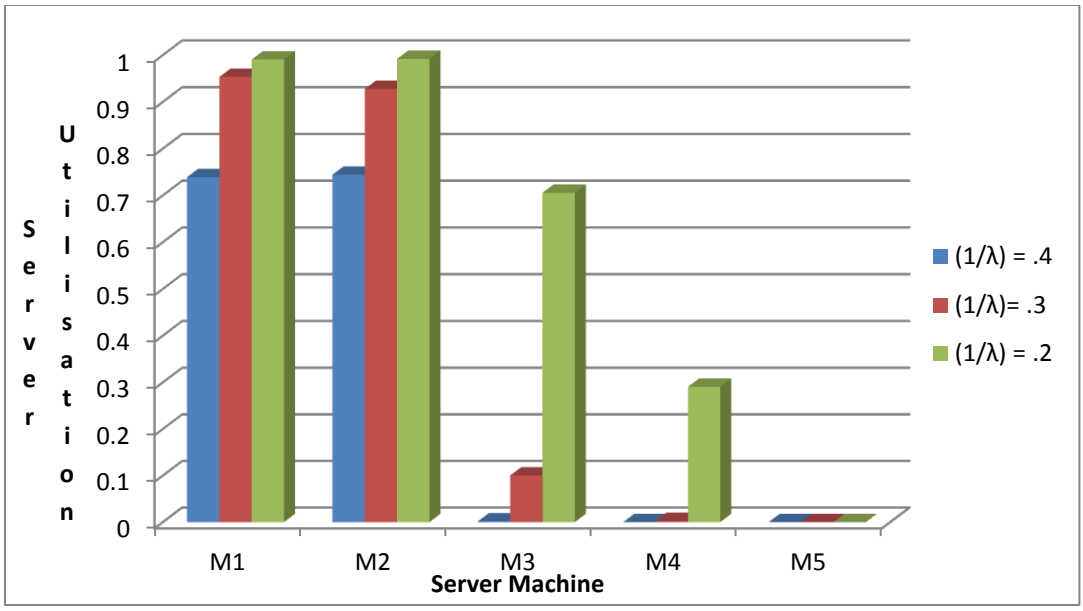

Fig. 6.5: Server Utilisation under DCS

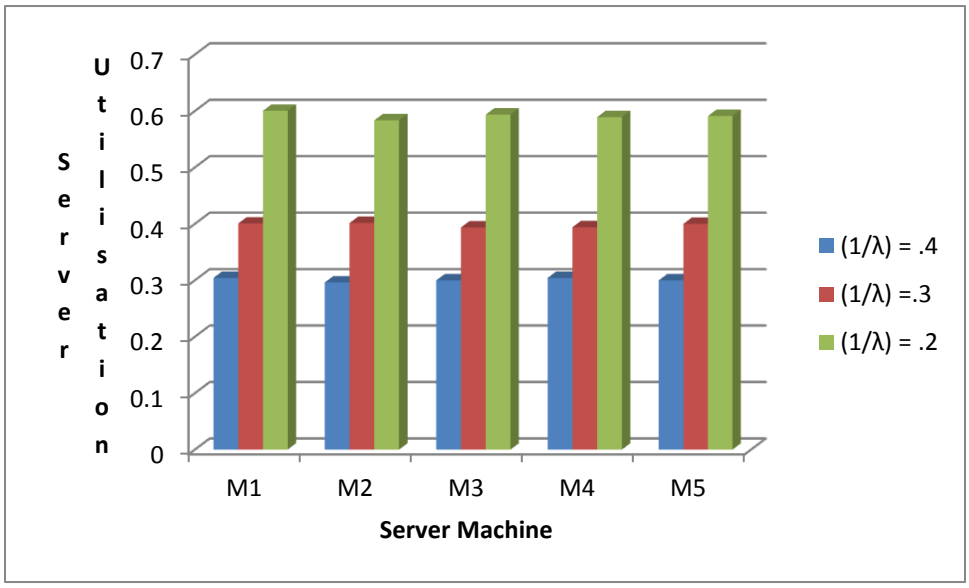

Fig. 6.6: Server Utilisation under the classical fixed model

Table 6-1: Detail results obtained based on the used parameters under DSC and the classical fixed methods

Parameter	DCS			Fixed		
	Experiment 1	Experiment 2	Experiment 3	Experiment 1	Experiment 2	Experiment 3
Waiting time (sec)	0.000547	0.001315	0.002698	0.000303	0.000357	0.000544
Ave. Server Utilisation	0.742531	0.661123	0.992867	0.3009	0.3979	0.5915
Interrupt Time (min)	0.00	.0.000004	0.0000072	0.0	0.0	0.0
Server used	2	3	4	5	5	5
Power consumed (Watts)	4	6	8	10	10	10

Table 6-2: Cost-Benefit Analysis of DCS and classical methods

Cost -Benefit			
	Experiment 1	Experiment 2	Experiment 3
Total Cost ($)	0.058625	0.230899	0.518563
Total Benefit ($)	10.76652	7.052892	4.605468
Cost-Benefit Ratio ($)	183.6514	30.54533	8.881209

6.7 CHAPTER SUMMARY

In this chapter, The N-policy has been investigated in Cloud E-Marketplaces using a Dynamic Control System (DCS). The author has addressed two things in this chapter: the first is to determine which of the two models provide an optimal solution in terms of server management and the second is the profitability concept of the experiment in the model that produces an optimal solution.

This was done by manning some servers and varying some. On the first one, a comparative study of the DCS model with the classical M/M/s was studied. The researcher's results reveal a better performance in terms of server utilisation and optimal server provisioning as against the classical model that over- satisfies the delay requirement. On the second one, a systematic Cost-Benefit Analysis (CBA) was developed to determine profitability. Profitability is a function of the cost and the other parameters used. Since the CBA indicated that the benefit accrued is greater than the cost, that is, CBA is greater than unity in all the experiments conducted, this then implies that the DCS is profitable. The Cost-Benefit Ratio (CBR) proved to be the experiment that produced the optimal profit. This model will hopefully guard against server under-provisioning and overprovisioning in Cloud E-Marketplaces. It will be a good prescriptive mechanism for effective management of the Cloud E-Marketplace.

CHAPTER SEVEN

CONCLUSION AND FUTURE WORK

7.1 SUMMARY

As consumers drift to the Cloud E-Marketplace shopping for services with different service offerings, three things are important in these markets; the Cloud implementations, performance and the optimal service level that will minimise cost and consumer waiting time. A rich body of new knowledge exists with regards to Cloud implementations but the performance aspect and the trade-off issues have hardly been discussed. This thesis focused on two issues: that of performance impact of provider offerings on consumers' waiting time, and the balancing of trade-offs between consumer waiting time and provider costs. Evidence exists in the context of the Cloud E-Marketplace, that research efforts are abundant on the two issues using the Pre-emptive model. However, it is well known that in practice, preemption and migration of virtual machines are costly and the switching time plays a significant role. This work explored existing and widely adopted theories related to Non Priority and Non Pre-emptive priority systems. The research methodology largely relied on the queuing theory and its corresponding simulation to fulfill the goal and objectives of the thesis.

The study applied the Non priority queue model to an environment where the service offering is the same for the consumers, and that of the Non pre-emptive model to that in which the service offerings are different. The issue of how optimal service level with better consumer waiting time can be achieved without the breach of SLA in E-Marketplaces in the context of the non-priority environment was investigated. The result from the first simulation was achieved at the point where the total cost is minimal. Due to the difficulty involved in quantifying consumers' waiting time, the use of the Aspiration model was adopted. The experimental results reveal an acceptable range based on the

conflicting measures of consumer waiting time and provider cost as supplied by the stakeholders.

The issue of better performance that improves consumers' and providers' satisfaction in the context of differentiated service provisioning was studied. The consideration was on the E-Marketplaces that have only two service offerings. The summary of the results shows that Class two consumers experience a longer waiting time than Class one consumers but the average waiting time remains the same in both priorities. In addition, the total waiting time is independent of the service discipline. While this result is recommended to the business world as particularly applicable where service provisioning is differentiated based on time and cost, a note of caution is sounded about the effect the implementation could have on promoting consumer dissatisfaction.

The preliminary work was based on cost minimisation in a non-priority environment and performance evaluation using two different service offerings in non-preemptive environment. Due to an increase in the number of consumers demanding different service offerings, the issue of how the non-preemptive model can work in a multi service provisioning environment was studied. The study consists of first addressing the issue of how the two different service offerings earlier considered could be generalised and the performance investigated. The second aspect is the determination of optimal service level in the same context. On performance impact, as the server utilisation tends towards large number, better waiting time performances over the conventional exogenous non priority model was observed. This inference was based on four out of the five classes observed in an experiment. The optimal solution was achieved at the point where the expected cost has the minimal value. Although, the same optimal point is recorded for both Non preemptive priority and the non-priority model; but where consumers' costs are prioritized, then the Non preemptive priority model appeared to be more efficient and profitable, apart

from the optimal number of server machines which has minimised consumers' waiting time.

Effective resource management was imperative to avoid the issues of resource over-provisioning and under-provisioning. A Dynamic control mechanism was set up to monitor the incoming requests. A comparative study of this model with the classical M/M/s was carried out. The results revealed a better performance in terms of server utilisation and optimal server provisioning in contrast to the classical model that over satisfied the delay requirement. The basic goal in this context is to have prescriptive measures in place to avoid service over-provisioning and under-provisioning in time dependent Cloud E-Marketplaces while at the same time satisfying consumers' requests.

In order to establish the thesis that minimisation of server machine cost and consumer waiting time in the context of non-priority and non-pre-emptive priority policy is imperative, the study successfully addressed the followings:

iii. Reviewed extensively the existing body of knowledge on the performance of E-Marketplaces;
iv. Saw the need to re-engineer the existing Cloud E-Marketplace architecture as networks of queues with parallel web stations with a feedback scheduler (Dispatcher-In) in the context of non-priority and non-pre-emptive policy without dedicating any web-station to any class to achieve optimal service level;
v. Evaluated the Non Priority First Come First Serve, FCFS service discipline and the Non-Preemptive model in order to see the performance impact on consumers' waiting time and Providers' cost;
vi. Formulated a cost structure that balances the server machine (Service Level) and consumers' waiting time in both the non-priority and non-preemptive models.

vii. Formulated a dynamic waiting time optimisation control mechanism that further addressed the issues of service over- and under-provisioning.

7.2 CONTRIBUTIONS

This study successfully examined the trends in the Cloud E-Marketplace. The contributions are:

The evaluative study of non-priority queue in series against the generalised approach that uses a single point of entry as proposed by others in the literature. This was used to determine the optimal service level and consumer waiting time.

The exhaustive evaluation of a novel non-preemptive architectural model of the Cloud E-Marketplace with each of service stations modeled as M/M/c/Pr against the M/M/1 proposed in the literature. This model was unique in that it:

i. Explored a different mathematical and simulation concept and also;
ii. Resolved the challenge of dedicating or allocating servers to a particular consumer class thereby reducing consumer waiting time.
iii. Investigated E-Marketplaces under the non-priority and also the two service non pre-emptive and the generalised models.
iv. Introduced the novel concept of profitability and Cost Benefit Ratio by using Dynamic Control Model (DCM) over the Fixed Server Model (FSM).

7.3 LIMITATIONS OF THE RESEARCH

The scope of this research covers the issue of the performance impact of provider offerings on consumers' waiting time and also the trade off balance of consumer-provider in terms of waiting time and costs. However, certain issues go beyond the scope of this research. For example, developing a threshold control mechanism that will control a higher priority class (class 2) so that when it reaches a certain threshold the control switches back to lower class (Class 1)

was not discussed. In addition, the issue of how the unutilised server idle period under the Dynamic Control System could be optimised from another domain for optimal profit maximisation was also not discussed.

Furthermore, the application of this work has been limited to solving E-health issues [154] , How this can be extended to other domains like the military was beyond the scope of the thesis.

7.4 FUTURE WORK

While this thesis could be used as the building block or the foundation to formulate other policies on how to optimise waiting time and costs in Cloud E-Marketplaces the performance issue goes beyond waiting time and cost. Other performance related issues are: SLA compliance, and the Network and Application challenges. Therefore, making Performance Monitoring as Service (PMaaS) would be a great solution to these challenges. It must be stated that traditional server monitoring is quite different from performance monitoring in the Cloud. This is because traditional performance monitoring focuses on specific components rather than having a holistic view of the cloud environment [155]. This could be done through a systematic performance monitoring framework. This framework could be broken down into sub-frameworks to address each of the performance related challenges.

The first of the sub-frameworks could address the issue of SLA, which is the reciprocal agreement between the provider of an IT service and the consumer of that service about the level of service, or QoS, to be delivered [156]. This sub framework could be created to contain some components like the monitoring, processing, data and reporting components. These components can then be integrated through some interface units. These components should consider the dynamic nature of the cloud E-Markets where resource usage changes. It should

also consider the issue of different service offerings provided by different service providers when determining the various QoS requirements. Part of the components should also introduce a third party mechanism that will monitor, checkmate the abuse and non- compliance to the agreed QoS. Another critical area under the SLA is how the consumer-provider information that is used by the components could be loosely coupled such that only necessary information is provided without the detailed information revealed to the receiving end for security sake because of the involvement of a third party. This is because tightly coupled data may not be able to take advantage of many performance-enhancing features of E-Market clouds, such as placing database processing in a series of elastic instances or using a database as a service in the host E-Market cloud [157].

The issue of Network monitoring is also important because slow networks mean slow systems and also poor performance. The other sub-framework should address the issue of network monitoring in cloud E-Marketplaces. This will look into how various fault tolerance mechanisms could be put in place to cater for some network problems like network failure, and attenuation that sometimes has a great effect on the QoS of consumers. In addition, issues like VPN and bandwidth monitoring should be addressed. The researcher's idea of proposing PMaaS is to lift the burden of IT infrastructure from the consumer and at the same time reduce costs and allow service consumers to concentrate on their core business, thereby allowing performance related issues to be handled by cloud E-Market providers.

Apart from the issue of creating performance monitoring as a service, the Modelling structure (M/M/1/c/FCFS, M/M/c/FCFS, M/M/1/c/Pr, M/M/c/Pr) used by the researcher in this thesis could be changed or remodelled to reflect new arrival, service, discipline and population patterns in the cloud E-Marketplaces.

This will require a different mathematical queuing theory with different simulation and real life concepts. Different research questions could be generated. Systematic analysis and evaluations could then be carried out to know the level of correctness of such a model and the degree of improvement over the existing ones. For example, the arrival and service patterns of consumers could be remodelled as markovian and General (M/G/c/Pr) for non-preemptive priority operation. This may be General (G/G/c/FCFS) for non-priority FCFS operation depending on the existing service strategy that will strike a good balance between the QoS or parameters under consideration.

APPENDIX A

A.1 DETERMINATION THE LIST OF ADMISSIBLE STATE

The admissible state are written below

$$0 \leq s \leq \delta - 1 \qquad then\ n = 0 \tag{A.1}$$

$$s = \delta \qquad then\ 0 \leq n \leq N - 1 \tag{A.2}$$

$$\delta + 1 \leq s \leq S - 1 \qquad then\ v \leq n \leq N - 1 \tag{A.3}$$

$$s = S \qquad then\ v \leq n \tag{A.4}$$

A.2 DERIVATION OF THE STEADY STATE EQUATION

Knowing that the mean transition rate of inflow of one subset is the same as the mean transition rate out of that subset, then the steady state equations are given as

$$\lambda P_{0,s} = \mu(s+1)P_{0,s+1} \qquad (n = 0, 0 \leq s \leq \delta - 1) \tag{A.5}$$

$$\lambda P_{n,\delta} = \delta\mu P_{n+1,\delta} \qquad (0 \leq n \leq v - 1, s = \delta) \tag{A.6}$$

$$\lambda P_{n,\delta} = \delta\mu P_{n+1,\delta} + \lambda P_{N-1,,\delta} \qquad (v \leq n \leq N - 2, s = \delta,) \tag{A.7}$$

$$\lambda P_{n,s} + s\mu P_{v,s} = s\mu P_{n+1,s} + \lambda P_{N-1,,s} \tag{A.8}$$

$$s\mu P_{v,s} = \lambda P_{N-1,s-1} \qquad (\delta + 1 \leq s \leq S \tag{A.9}$$

$$\lambda P_{n,s} + S\mu P_{v,s} = S\mu P_{n+1,s} \qquad (v \leq n \leq N - 2, \delta + s = S) \tag{A.10}$$

$$\lambda P_{n,s} = S\mu P_{n+1,s} \qquad (N - 1 \leq n) \tag{A.11}$$

A.3 DERIVATION OF THE STEADY STATE PROBABILITY

A general approach is used to determine the steady state probability.

Writing

$$P_{0,s+1} = \left[{}^{\rho}/_{S+1}\right]P_{0,s}$$

and let s=0, s=1, e.t.c can be expressed as a function of P_{00} as shown in equation A.12. then

$\sum_1^k P_{n,s} = 1$ where k is the total number of the admissible state.

The summary to the solutions of equation A.5 to A.11 is given below

$$P_{0,s} = \left({}^{\rho^s}/_{s'}\right)P_{0\,0} \qquad (n = 0, 0 \le s \le \delta) \qquad (A.12)$$

$$P_{n,\delta} = \gamma^n \left({}^{\rho^\delta}/_{\delta'}\right)P_{0\,0} \qquad (0 \le n \le v, s = \delta) \qquad (A.13)$$

$$P_{n\,\delta} = \left(\frac{\rho^\delta}{\delta'}\right)\left[(\gamma^n - \gamma^N)/(1 - \gamma^{N-v})\right]P_{0\,0} \qquad (v \le n \le N-1, s = \delta) \quad (A.14)$$

$$P_{n,s} = P_{v\,s} \qquad (v \le n \le N-1, \quad \delta+1 \le s \le S-1) \quad (A.15)$$

$$P_{n\,s} = \left(\frac{\rho^\delta}{s'}\right)\left[\left(\frac{\gamma^{N-1} - \gamma^N}{1 - \gamma^{N-v}}\right)\right]P_{0\,0} \qquad (v \le n \le N-1, \ \delta+1 \le s \le S-1) \qquad A.16)$$

$$P_{n\,S} = \rho^S \left[\left(\frac{(S^{n-v+1} - \rho^{n-v+1})(\gamma^{N-1} - \gamma^N)}{S^{n-v}S'\,(S-\rho)(1 - \gamma^{N-v})}\right)\right]P_{0\,0} \qquad (v \le n \le N-1, s = S) \qquad (A.17)$$

$$P_{n\,S} = \rho^{S+N+n+1}\left[\left(\frac{(S^{N-v}-\rho^{N-v})(\gamma^{N-1}-\gamma^{N})}{S^{n-v}S'\,(S-\rho)(1-\gamma^{N-v})}\right)\right]P_{0\,0} \qquad (n \geq N-1, s = S) \qquad (A.18)$$

because the sum of all probability = 1 (the admissible state) then

$$P_{0\,0} + \sum_{s=1}^{s=\delta} P_{0\,s} + \sum_{n=1}^{n=v} P_{n\,\delta} + \sum_{n=v+1}^{N-1} P_{n\,\delta} + \sum_{n=v}^{N-1}\sum_{s=\delta+1}^{S-1} P_{n\,s} + \sum_{n=v}^{N-1} P_{n\,S} + \sum_{n=N}^{n=\infty} P_{0\,S} = 1 \quad (A.19)$$

evaluating $P_{0\,0}$ gives

$$P_{0\,0} = \left[\sum_{s=1}^{s=\delta} P_{0\,s} + \sum_{n=1}^{n=v} P_{n\,\delta} + \sum_{n=v+1}^{N-1} P_{n\,\delta} + \sum_{n=v}^{N-1}\sum_{s=\delta+1}^{S-1} P_{n\,s} + \sum_{n=v}^{N-1} P_{n\,S} + \sum_{n=N}^{n=\infty} P_{0\,S}\right]^{-1} \qquad (A.20)$$

$$idle_{time} = \sum_{s=0}^{\delta-1}(\delta - s)P_{0\,s} = P_{0\,0}\sum_{s=0}^{\alpha-1}\frac{(\delta-s)\rho^{s}}{s'} \qquad (A.21)$$

$$Inter_{time} = \sum_{s=\delta}^{S-1} P_{N-s} = \gamma^{N-1}(1-\gamma)P_{o\,o}/(1-\gamma^{N-v})\sum_{s=\delta}^{S-1}\left(\rho^{s}/_{s'}\right) \quad A.22)$$

$$L_q = \frac{P^{\delta}}{1-\gamma^{N-v}}\Bigg[\frac{\rho^{\delta}[\gamma-(v+1)\gamma^{v+1}+v\gamma^{v+2}](1-\gamma^{v})}{\delta'(1-\gamma)^2}$$
$$+\frac{\rho^{\delta}}{\delta'}\Bigg\{\frac{\gamma^{N}(N\gamma-\gamma-N)+\gamma^{v+1}(v+1-v\gamma)}{(1-\gamma)^2}$$
$$-\frac{1}{2}\gamma^{N}[N(N-1)-v(v+1)]\Bigg\}$$
$$+\frac{1}{2}\,\gamma^{N-1}(1-\gamma)[N(N-1)-v(v-1)]\left\{\sum_{s=\delta}^{S-1}\frac{\rho^{s}}{s'}\frac{\rho^{\delta}}{\delta'}\right\}$$
$$+\frac{\rho^{S}\gamma^{N-1}(1-\gamma)}{2S'(S-\rho)^3S^{N-2}}\{S^{N-1}(S-\rho)^2\big[\big((N-1)\big)\big((N-2)\big)$$
$$-v(v-1)\big]-2S^{v}\rho^{N-v}[\rho(N-2)-S(N-1)]$$
$$-2S^{N-v-1}\rho^{v+v}[Sv-(v-1)\rho]$$
$$+2(S^{N-v}$$
$$-\rho^{N-v})S^{v}\big[S\big(S(N-1)$$
$$-\rho(N-2)\big)\big]\}\Bigg] \qquad (A.23)$$

$$w_q = \frac{L_q}{\lambda} \qquad (A.24)$$

REFERENCES

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

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

[3] 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.

[4] 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.

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

[6] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Futur. Gener. Comput. Syst.*, vol. 25, no. 6, pp. 599–616, Jun. 2009.

[7] H. El-Gohary, "E-Marketing-A literature Review from a Small Businesses perspective," *Int. J. Bus. Soc. Sci.*, pp. 214–244, 2010.

[8] W. K.Chong, "Performances of B2B e-Marketplace for SMEs: The Research Methods and Survey Results." [Online]. Available: http://www.ibimapublishing.com/journals/CIBIMA/volume9/v9n22.pdf. [Accessed: 15-Jun-2015].

[9] K. Laudon and J. Laudon, *Management Information Systems: Managing the Digital Firm*. Prentice-Hall International, New Jersey, 2002.

[10] D. Chaffey, *E-Business and E-Commerce Management*. Financial Times-Prentice Hall, London, 2004.

[11] B. Sculley, W. William, and A. Woods, *The Killer Application in the Busines-to-Business Internet Revolution*. Harper Paperbacks, 2001.

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

[13] H. Khazaei, J. Misic, and V. B. Misic, “Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 5, pp. 936–943, May 2012.

[14] 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].

[15] A. N. Toosi, R. N. Calheiros, R. K. Thulasiram, and R. Buyya, “Resource Provisioning Policies to Increase IaaS Provider’s Profit in a Federated Cloud Environment,” in *2011 IEEE International Conference on High Performance Computing and Communications*, 2011, pp. 279–287.

[16] L. Wu, S. K. Garg, and R. Buyya, “SLA-Based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments,” *2011 11th IEEE/ACM Int. Symp. Clust. Cloud Grid Comput.*, pp. 195–204, May 2011.

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

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

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

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

[21] K. Kant and M. M. Srinivasan, *Introduction to computer system performance evaluation*. McGraw-Hill, 1992.

[22] C. Carlos, L. S. Kim, T. Helen, and W. H. Ronald, “Management Accounting: Information for Managing and Creating Value:,” *9780077116903: Amazon.com: Books*, 2013. [Online]. Available: http://www.amazon.com/Management-Accounting-Information-Managing-Creating/dp/0077116909. [Accessed: 25-Jul-2014].

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] H. Goudarzi and M. Pedram, “Multi-dimensional SLA-Based Resource Allocation for Multi-tier Cloud Computing Systems,” in *2011 IEEE 4th International Conference on Cloud Computing*, 2011, pp. 324–331.

[33] R. Ghosh, K. S. Trivedi, V. K. Naik, and D. S. Kim, “End-to-End Performability Analysis for Infrastructure-as-a-Service Cloud: An Interacting Stochastic Models Approach,” in *2010 IEEE 16th Pacific Rim International Symposium on Dependable Computing*, 2010, pp. 125–132.

[34] E. Pakbaznia and M. Pedram, “Minimizing data center cooling and server power costs,” in *in Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design - ISLPED ’09*, 2009, pp. 145–150.

[35] H. Khazaei, S. Member, and J. Miˇ, “A Fine-Grained Performance Model of Cloud Computing Centers,” vol. X, pp. 1–11, 2012.

[36] M. R. Ahmadi and D. Maleki, “Performance evaluation of server virtualization in data center applications,” in *2010 5th International Symposium on Telecommunications*, 2010, pp. 638–644.

[37] H. Khazaei, “Performance Modeling of Cloud Computing Centers,” *Ph.D Thesis, Univ. Manitoba Winnipeg Manitoba, Canada*, no. January, 2013.

[38] S. Begum, “Review of Load Balancing in Cloud Computing,” vol. 10, no. 1, pp. 343–353, 2013.

[39] F. Mustafa and T. L. McCluskey, “Dynamic Web Service Composition,” in *2009 International Conference on Computer Engineering and Technology*, 2009, pp. 463–467.

[40] Akingbesote A.O, M. O. Adigun, J. Oladosu, and E. Jembere, “The Trade-off between consumer’s satisfaction and resource service level by e-market providers in e-market places,” *in International Conference on Information Technology (ICIT)*, 2013. [Online]. Available: http://connection.ebscohost.com/c/articles/95511261/modeling-cloud-e-marketplaces-cost-minimization-using-queuing-model.

[41] D. F. Garcia, J. Garcia, J. Entrialgo, M. Garcia, P. Valledor, R. Garcia, and A. M. Campos, “A QoS Control Mechanism to Provide Service Differentiation and Overload Protection to Internet Scalable Servers,” *IEEE Trans. Serv. Comput.*, vol. 2, no. 1, pp. 3–16, Jan. 2009.

[42] Y. Nezih, L. Alexandru, E. Dick, and O. . Simon, “C Meter :A Framework for Performance Analysis of Computing Clouds 9th IEEE/ACM International Symposium on Cluster Computing and the Grid,” in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2009.

[43] H. Chen and S. Li, “A Queueing-based Model for Performance Management on Cloud,” in *on Advanced Information Management and Service (IMS), Seoul*, 2010, pp. 83–88.

[44] B. Li, J. Li, J. Huai, T. Wo, Q. Li, and L. Zhong, “EnaCloud: An Energy-Saving Application Live Placement Approach for Cloud Computing Environments,”

in *2009 IEEE International Conference on Cloud Computing*, 2009, pp. 17–24.

[45] P. B. M, P. S. K. P, and P. P. G. V, “Performance factors of cloud computing data centers using M / G / m / m + r queuing systems,” vol. 2, no. 9, pp. 6–10, 2012.

[46] H. Khazaei, “Modelling of Cloud Computing Centers Using M / G / m Queues,” 2011.

[47] D. Gross and F. John, *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc. New York, NY, USA, 1985.

[48] 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.

[49] 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].

[50] J. P. Bailey and Y. Bakos, “An Exploratory Study of the Emerging Role of Electronic Intermediaries,” *Int. J. Electron. Commer.*, vol. 1, no. 3, pp. 7–20, 1997.

[51] J. G. Norm Archer, “Managing in the context of the new electronic marketplace,” in *Proceedings 1st World Congress on the Management of Electronic Commerce*, 2000.

[52] N. Russ, “E-marketplaces: New Challenges for Enterprise Policy, Competition and Standardisation.,” Workshop Report, Brussels Apri 2001.

[53] J. Il Kim, C. M. Lee, and K. H. Ahn, “Dongdaemun, a traditional market place wearing a modern suit: The importance of the social fabric in physical redevelopments,” *Habitat Int.*, vol. 28, no. 1, pp. 143–161, 2004.

[54] M. Grieger, “Electronic marketplaces: A literature review and a call for supply chain management research,” *Eur. J. Oper. Res.*, vol. 144, no. 2, pp. 280–294, 2003.

[55] D. R. Henderson, “Electronic marketing in principle and practice.,” *merican J. Agric. Econ.*, vol. A 66 (5), pp. 848–853, 1984.

[56] J. H. McCoy and M. E. Sarhan, *Livestock and Meat Marketing*. Avi Pub Co, 1988.

[57] 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.

[58] “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].

[59] T. J. Strander and M. J. Shaw, “Characteristics of electronic Markets,” *Decis. Support Syst. North-holl.*, vol. 21, no. 3, pp. 185–198, 1997.

[60] J. Y. Bakos, “Reducing search costs : Implications for electronic marketplaces,” *Manage. Sci.*, vol. 43, p. 17, 1997.

[61] K. Berryman, L. Harrington, D. Layton-Rodin, and V. Rerolle, “Electronic Commerce: Three Emerging Strategies,” *McKinsey Q.*, no. 1, p. 152, Jan. 1998.

[62] R. Stockdale, “Benefits and barriers of electronic marketplace participation: an SME perspective,” *J. Enterp. Inf. Manag.*, 2004.

[63] M. Norris and S. West, *eBusiness Essentials*. Chichester, UK: John Wiley & Sons, Ltd, 2001.

[64] R. Nathan, “E-Marketplaces: new challenges for enterprise policy, competition and standardisation,” Brussel, pp. 1–28, 2001.

[65] A. Kambil and E. Van Heck, *Making markets: How Firms Can Design and Profit from Online Auctions and Exchanges*. Harvard Business Press, 2002.

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

[67] W. K. Chong and M. Shafaghi, “Performances of B2B e-Marketplace for SMEs : The Research Methods and Survey Results,” *Commun. IBIMA*, vol. 9, no. 22, pp. 185–192, 2009.

[68] Q. Z. Sheng, X. Qiao, A. V. Vasilakos, C. Szabo, S. Bourne, and X. Xu, “Web services composition: A decade’s overview,” *Inf. Sci. (Ny).*, vol. 280, pp. 218–238, Oct. 2014.

[69] L. Zeng, B. Benatallah, A. H. H. Ngu, M. Dumas, J. Kalagnanam, and H. Chang, “QoS-aware middleware for Web services composition,” *IEEE Trans. Softw. Eng.*, vol. 30, no. 5, pp. 311–327, 2004.

[70] Z. Shen and J. Su, “Web service discovery based on behavior signatures,” *2005 IEEE Int. Conf. Serv. Comput.*, vol. 1, 2005.

[71] E. Al-Masri and Q. H. Mahmoud, “QoS-based discovery and ranking of Web services,” *Proc. - Int. Conf. Comput. Commun. Networks, ICCCN*, pp. 529–534, 2007.

[72] M. Makhlughian, S. M. Hashemi, and Y. Rastegari, “W Eb Service Selection Based on R Anking of,” *Int. J. Web Serv. Comput.*, vol. 3, no. 1, pp. 1–14, 2012.

[73] D. Palanikkumar, “An Algorithmic Evaluation of Optimal Service Selection using BCO,” *Eur. J. Sci. Res.*, vol. 68, no. 4, pp. 591–605, 2012.

[74] X. Fan and X. Fang, “On optimal decision for QoS-aware composite service selection,” *Inf. Technol. J.*, vol. 9, no. 6, pp. 1207–1211, 2010.

[75] L. Chen, Y. Feng, J. Wu, and Z. Zheng, “An Enhanced QoS Prediction Approach for Service Selection,” *2011 IEEE Int. Conf. Serv. Comput.*, no. 3, pp. 727–728, 2011.

[76] N. Sasikaladevi and L. Arockiam, “Genetic Approach for Service Selection problem in Composite Web Service,” *Int. J. Comput. Appl.*, vol. 44, no. 4, pp. 22–29, 2012.

[77] D. Palanikkumar and G. Kousalya, “An Evolutionary Algorithmic Approach based Optimal Web Service Selection for Composition with Quality of Service Department of CSE , Anna University of Technology , Coimbatore-47 , Coimbatore Department of CSE , Sri Krishna College of Engineering and Tech,” *J. Comput. Sci.*, vol. 8, no. 4, pp. 573–578, 2012.

[78] R. Ben lakhal and W. Chainbi, “A Multi-Criteria Approach for Web Service Discovery,” *Procedia Comput. Sci.*, vol. 10, pp. 609–616, Jan. 2012.

[79] M. Sathya and M. Swarnamugi, “Evaluation of QoS Based Web- Service Selection Techniques for Service Composition,” *Int. J. Softw. Eng.*, vol. 1, no. 5, pp. 73–90, 2010.

[80] N. H. Priya, “Optimal Selection and Composition of Web Services – a Survey,” *Int. J. Comput. Appl.*, vol. 49, no. 2, pp. 1–5, 2012.

[81] D. B. Claro, P. Albers, and J. Hao, “Web services composition,” *Semant. Web Serv. Process. Appl.*, pp. 195–225, 2006.

[82] N. Milanovic and M. Malek, “Current solutions for Web service composition,” *IEEE Internet Comput.*, vol. 8, no. 6, pp. 51–59, 2004.

[83] C.-F. Lin, R.-K. Sheu, Y.-S. Chang, and S.-M. Yuan, “A relaxable service selection algorithm for QoS-based web service composition,” *Inf. Softw. Technol.*, vol. 53, no. 12, pp. 1370–1381, 2011.

[84] N. Li, “Modified Particle Swarm Optimization and Its Application in Multimodal Function,” in *2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE) December 16-18, Changchun, China*, 2011, pp. 375–378.

[85] A. O. Akingbesote, M. O. Adigun, J. B. Oladosu, and E. Jembere, “A Quality of Service Aware Multi-Level Strategy for Selection of Optimal Web Service,” in *in 5th IEEE ICAST International Conference on Adaptive Science and Technology, Pretorial, South Africa*, 2013, pp. 25–27.

[86] J. Y. J. Yang, M. P. Papazoglou, and W.-J. Van Den Heuvel, “Tackling the challenges of service composition in e-marketplaces,” *Proc. Twelfth Int. Work. Res. Issues Data Eng. Eng. E-commerce/e-bus. Syst. RIDE-2EC*, 2002.

[87] 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.

[88] 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.

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

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

[91] 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.

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

[93] 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.

[94] 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.

[95] 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.

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

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

[98] A. Joshua and F. Ogwueleka, “Cloud Computing with Related Enabling Technologies,” *Int. J. Cloud Comput. Serv. Sci.*, vol. 2, no. 1, pp. 40–49, 2013.

[99] P. Mell and T. Grance, “The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology,” *Nist Spec. Publ.*, vol. 145, p. 7, 2011.

[100] N. Khanghahi and R. Ravanmehr, “Cloud Computing Performance Evaluation : Issues and Challenges,” *… J. Cloud Comput. …*, vol. 3, no. 5, pp. 29–41, 2013.

[101] S. M. Habib, S. Hauke, S. Ries, and M. Mühlhäuser, “Trust as a facilitator in cloud computing: a survey,” *J. Cloud Comput. Adv. Syst. Appl.*, vol. 1, no. 1, p. 19, 2012.

[102] H. Alhakami, H. Aldabbas, and T. Alwada, “C Omparison B Etween C Loud and G Rid C Omputing : R Eview Paper,” vol. 2, no. 4, pp. 1–21, 2012.

[103] S. De Chaves, R. Uriarte, and C. Westphall, “Toward an architecture for monitoring private clouds,” *IEEE Commun. Mag.*, vol. 49, no. 12, pp. 130–137, Dec. 2011.

[104] “Available.” [Online]. Available: http://mobiledevices.about.com/od/additionnalresources/a/Cloud-Computing-Is-It-Really-All-That-Beneficial.htm,http://www.rickscloud.com/how-performance-issues-impact-cloud-adoption.

[105] “Seedy.” [Online]. Available: http://www.comsoc.org/files/Publications/Magazines/ni/cfp/Cfp.

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

[107] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, and T. D. Nguyen, “Reducing electricity cost through virtual machine placement in high performance computing clouds,” *Proc. 2011 Int. Conf. High Perform. Comput. Networking, Storage Anal. - SC ’11*, p. 22, 2011.

[108] P. Appandairajan, N. Zafar Ali Khan, and M. Madiajagan, “ERP on Cloud: Implementation strategies and challenges,” in *Proceedings of 2012 International Conference on Cloud Computing Technologies, Applications and Management, ICCCTAM*, 2012, pp. 56–59.

[109] 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.

[110] F. Kamoun, “Performance Analysis of Two Priority Queuing Systems in Tandem,” vol. 2012, no. November, pp. 509–518, 2012.

[111] S. Nagendram, “Efficient Resource Scheduling in Data Centers using MRIS,” *(IJCSE)(IJCSE)*, vol. 2, no. 5, pp. 764–769, 2011.

[112] F. Wang, N. Helian, and G. Akanmu, “User-priority guided Min-Min scheduling algorithm for load balancing in cloud computing,” *2013 Natl. Conf. Parallel Comput. Technol.*, pp. 1–8, Feb. 2013.

[113] 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.

[114] J. Zhao, C. Wu, and Z. Li, “Cost Minimization in Multiple IaaS Clouds : A Double Auction Approach,” *arXiv Prepr. arXiv1308.0841*, 2013.

[115] H. Xu and B. Li, “Cost efficient datacenter selection for cloud services,” *2012 1st IEEE Int. Conf. Commun. China*, pp. 51–56, Aug. 2012.

[116] Y. C. Lee, C. Wang, A. Y. Zomaya, and B. B. Zhou, “Profit-driven service request scheduling in clouds,” in *CCGrid 2010 - 10th IEEE/ACM International Conference on Cluster, Cloud, and Grid Computing*, 2010, pp. 15–24.

[117] B. D. Bunday, *An Introduction to Queueing Theory*. Arnold, 1996.

[118] Sundarapandian, *Probability, Statistics and Queuing Theory*. PHI Learning Pvt. Ltd., 2009.

[119] Y. Chiang and Y. Ouyang, "Profit Optimization in SLA-Aware Cloud Services with a Finite Capacity Queuing Model," vol. 2014, 2014.

[120] A. Taha, "Operations Research an Introduction: Hamdy A. Taha: 9788120322356: Amazon.com: Books." [Online]. Available: http://www.amazon.com/Operations-Research-Introduction-Hamdy-Taha/dp/8120322355/ref=sr_1_4?s=books&ie=UTF8&qid=1410515660&sr=1-4. [Accessed: 12-Sep-2014].

[121] H. Kobayashi and B. L. Mark, *System Modeling and Analysis: Foundations of System Performance Evaluation*. Pearson Education India, 2009.

[122] H. R. Frank, *Microeconomics and Behavior (Mcgraw-Hill/Irwin Series in Economics): 9780078021695: Economics Books @ Amazon.com*. McGraw-Hill Irwin, Inc, 2006.

[123] J. Hirshleifer, A. Glazer, and D. Hirshleifer, *Price Theory and Applications: Decisions, Markets, and Information*, vol. 0. Cambridge University Press, 2005.

[124] A. Rosenfeld and S. Kraus, "Modeling Agents Based on Aspiration Adaptation Theory," *JAAMAS : http://www.icons.umd.edu/*.

[125] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," vol. 47, no. 2, pp. 263–292, 2007.

[126] D. Gross and C. M. Harris, *Fundamentals of queueing theory (2nd ed.).* John Wiley & Sons, Inc. New York, NY, USA, 1985.

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

[128] 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.

[129] 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.

[130] M. A. Salehi, B. Javadi, and R. Buyya, “Preemption-aware Admission Control in a Virtualized Grid Federation,” *2012 IEEE 26th Int. Conf. Adv. Inf. Netw. Appl.*, pp. 854–861, Mar. 2012.

[131] A. O. Akingbesote, M. O. Adigun, J. Oladosu, E. Jembere, I. Kaseeram, S. Africa, and S. Africa, “Modeling the Cloud e-Marketplaces for Cost Minimization Using Queuing Model,” *Aust. J. Basic Appl. Sci.*, vol. 8, no. 4, pp. 59–67, 2014.

[132] S. . Sunilkumar and K. S. Gopal, “Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey,” *J. Netw. Comput. Appl.*, vol. 41, no. 2, p. 2014.

[133] C. Madhumathi and G. Ganapathy, “An Academic Cloud Framework for Adapting e-Learning in Universities,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. Vol. 2, Is, 2013.

[134] S. Ferretti, V. Ghini, F. Panzieri, M. Pellegrini, and E. Turrini, “QoS-Aware Clouds,” in *2010 IEEE 3rd International Conference on Cloud Computing*, 2010, pp. 321–328.

[135] R. Pal and P. Hui, “Economic models for cloud service markets: Pricing and Capacity planning,” *Theor. Comput. Sci.*, vol. 496, pp. 113–124, Jul. 2013.

[136] M. Bailey, “The Economics of Virtualization : Moving Toward an Application-Based Cost Model,” 2009.

[137] L. Tadj and G. Choudhury, “Optimal design and control of queues,” *Top*, vol. 13, no. 2, pp. 359–412, 2005.

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

[139] Daniel P. Heyman, “The T-Policy for the M/G/1 Queue,” *Management Science: INFORMS*, 1977. [Online]. Available: http://pubsonline.informs.org/doi/pdf/10.1287/mnsc.23.7.775. [Accessed: 10-Jun-2015].

[140] K. R. Balachandran, “Control Policies for a Single Server System,” *Manage. Sci.*, vol. 19, no. 9, pp. 1013–1018, 1973.

[141] H.Takagi, “Queueing analysis :A foundation of Performance Evaluation Volume 1: Vacation and priority systems,” *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 9, no. 2, 1991.

[142] 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.

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

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

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

[146] 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.

[147] M. M. Systems, “Queuing system with variable server number,” no. 4, pp. 63–65, 2007.

[148] 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.

[149] 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.

[150] 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].

[151] S. S. Xu, “Queue-dependent servers,” *J. Eng. Math.*, vol. 7, no. 2, pp. 123–126, Apr. 1973.

[152] R. L. Garg and P. Singh, “Queue-dependent servers queueing system,” *Microelectron. Reliab.*, vol. 33, no. 15, pp. 2289–2295, Dec. 1993.

[153] T. A. Weber, *Price Theory in Economics*. 2008.

[154] 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.

[155] Infosys, “Performance Monitoring in Cloud,” *White Pap. www.infosys.com*, 2012.

[156] E. Stahl, P. K. Isom, and T. R. Stockslager, “Performance Implications of Cloud,” 2012. [Online]. Available: http://www.redbooks.ibm.com/redpapers/pdfs/redp4875.p df.

[157] “3 rules for getting top enterprise cloud performance | InfoWorld.” [Online]. Available: http://www.infoworld.com/article/2612755/cloud-computing/3-rules-for-getting-top-enterprise-cloud-performance.html. [Accessed: 06-Nov-2015].