Resilience Analysis of the IMS based Networks

Kamyod, Chayapol

DOI (link to publication from Publisher):
10.5278/VBN.PHD.ENGSCI.00116

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

RESILIENCE ANALYSIS OF THE IMS BASED NETWORKS

BY
CHAYAPOL KAMYOD

DISSERTATION SUBMITTED 2016

AALBORG UNIVERSITY
DENMARK

Resilience Analysis of the IMS based Networks

by

Chayapol Kamyod

AALBORG UNIVERSITY
DENMARK

Dissertation submitted

Dissertation submitted: June 2016

PhD supervisor: Associate Professor Neeli Rashmi Prasad
Aalborg University, Denmark

Assistant PhD supervisor: Assistant Professor Rasmus Hjorth Nielsen
Aalborg University, Denmark

PhD committee: Professor Knud Erik Skouby (chairman)
Aalborg University, Denmark

Professor Thipparaju Rama Rao
SRM University, India

Director Sudhir Dixit
Skydoot Inc., Woodside, USA

PhD Series: Faculty of Engineering and Science, Aalborg University

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-721-8

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2016

Curriculum Vitae

Chayapol Kamyod

I received my bachelor's degree in telecommunication engineering and master's degree in laser technology and photonics from Suranaree University of Technology, Nakhon Ratchasima, Thailand. I also earned my master's degree in engineering from the City Collage of New York, New York, United States. In addition to network quality of services and resilience, my research interests are reliability and security of the next generation networks and internet of things.

Abstract

This thesis introduced modern reliability or resilience evaluation framework. The reliability is one of the quality properties, therefore, this thesis discusses limitation and reliability related factors of various quality terms that affect overall reliability of contemporary service and system. The study also focuses on end-to-end availability and reliability evaluation method and models to enhance the overall reliability of the principal IP Multimedia Subsystem(IMS)based communication scenarios: intra-domain and inter-network communications. The thesis distinguishes various essential network parameters and quantitatively estimate end-to-end reliability characteristics through this parameters by using the proposed hybrid models. The projected model for each simplex and redundancy was compared with current progressive models and valid through the numerical analysis and network simulation. The results demonstrate that the proposed models and evaluation methods can provide better reliability properties of the system. Moreover, a selective resilience parameters and modern resilience evaluation framework are presented. The projected methodology can properly incorporate each subjective and objective parameters into the analysis.

Dansk Resumé

Denne afhandling indfører moderne pålidelighed eller ramme for robusthedsevaluering. Pålidelighed er en kvalitetsegenskab, og derfor diskuterer denne afhandling begrænsnings- og pålidelighedsrelaterede faktorer for forskellige kvalitetsudtryk, som påvirker den samlede pålidelighed af moderne service og systemer. Undersøgelsen fokuserer også på end-to-end tilgængelighed og evalueringsmetode for pålidelighed og modeller til at øge den samlede pålidelighed af de vigtigste IMS-baserede kommunikation scenarier: intra-domæne og internetværkskommunikation. Afhandlingen skelner mellem forskellige væsentlige end-to-end netværksparametre og estimerer kvantitativt end-to-end pålidelighedsegenskaber gennem disse parametre ved at bruge de foreslåede hybridmodeller. Den foreslåede model for både simplex og redundans bliver sammenlignet med de nuværende state-of-the-art modeller og valideret gennem numerisk analyse og netværkssimulering. Resultaterne viser, at de foreslåede modeller og evalueringsmetoder kan tilvejebringe bedre pålidelighedsegenskaber af systemet. Desuden præsenteres selektive modstandskraft parametre og moderne modstandskraft evalueringsnetværk. Den foreslåede metode kan på en ordentlig måde inkorporere både subjektive og objektive parametre i bedømmelsen.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Neeli R. Prasad for guiding me, for her patience, motivation, and immense knowledge. I am also grateful to Prof. Ramjee Prasad for providing valuable guidance, encouragement and helping me during the time of research and writing this thesis. Without their guidance and support, I would not have completed my Ph.D.

Besides my supervisor, I would like to thank my co-supervisor, Dr. Rasmus H. Nielsen, for his understanding, and patience. His immense knowledge and guidance helped me during all the time of research and writing of this thesis. I owe him my eternal gratitude.

I would also like to thank Ms. Susanne Nørrevang and Ms. Inga Hauge for their continuous support during my stay at Aalborg University.

I would like to thank my colleagues at MFU and CTIF for exchange of knowledge and skills which helped enrich my experience. A very special thanks to Mr. Prateek Mathur for his constant suggestions and support.

My sincere thanks also go to Dr. Albena D. Mihovska and Dr. Punnarumol Temdee who gave a useful comments for this thesis.

Last but not the least, I would like to thank my family: my parents and my sister and my girlfriend for supporting me spiritually throughout writing this thesis and my life in general.

Table of Contents

List of Figures

List of Tables

List of Acronyms

3GPP 3rd generation partnership project

3GPP2 3rd generation partnership project 2

A(t) instantaneous availability

A availability

A_i**(t)** interval availability

AS application server

AS application services

ATM asynchronous transfer mode

BGCF border gateway control function

CoS class of service

CSCF call session control functions

CTMC continuous-time Markov chain

DES discrete event simulation

DiffServ Differentiated services

DiffServ differentiated services

DTMC discrete time markov chain

E-CSCF emergency-call state control function

ETSI european telecommunications standards institute

FMC fixed-mobile convergence

GGSN GSM GPRS serving node

GoS grad of service

GPRS general packet radio service

GUI graphic user interface

HLR home location register

HSS home subscriber server

HTTP hypertext transfer protocol

I-CSCF interrogating-call state control function

IEEE institute of electrical and electronics engineers

IETF internet engineering task force

IETF internet engineering task force

IMS IP multimedia subsystem

IntServ integrated service

IP internet protocal

IPLR IP packet loss ratio

ISDN integrated services for digital network

ISP internet service provider

ISUP ISDN User Part

ITU international telecommunication union

LMP link management protocol

MDT mean down time

MG media gateway

MGCF media gateway control function

MOS mean opinion score

MRF media resource function

MRFC media resource function controller

MRFP media resource function processor

MRM markov reward model

MRM markov reward models

MT mobile terminal

MTBF mean time between failure

MTBI mean time between interruptions

MTTF mean time to failure

MTTR mean time to recovery

MUT mean up time

NGN next-generation networks

NS network simulator

PCRF policy and charging rules function

P-CSCF proxy-call state control function

PDF policy decision function

PDP policy decision point

PEP policy enforcement points

PSTN public switched telephone network

QNM queueing network model

QoD quality of delivery

QoE quality of experience

QoP quality of presentation

QoR quality of resilience

QoS quality of service

QPN queuing petri nets

R(t) instantaneous reliability

RAN radio access network

RBD reliability block diagram

RIP routing information protocol

RSVP resource reservation protocol

RTCP real time control protocol

RTP real-time transport protocol

S-CSCF serving-call state control function

SDP session description protocol

SGW signaling gateway

SGW signaling gateway

SIP session initiation protocol

SIP session initiation protocol

SIP-P SIP proxies

SIP-R SIP Registrars

SIP-UA SIP user agents

SLM service level agreement

SPN stochastic petri net

TCP transmission control protocol

TE terminal equipment

UE user end device

UGF universal generating function

URI universal resource identifier

URL universal resource locator

UTRAN/GERAN universal terrestrial GSM GPRS EDGE radio access networks

VoIP voice over IP

VPN virtual private network

OPNET Optimized Network Engineering Tools

Chapter 1

Introduction

The Quality of Service (QoS) is one of the key factors that is used to provide a certain quality level of computer networks and services and guarantee customer satisfaction and experience. Reliability or resilience have also become another important feature. A valid quality and reliability evaluation method can, therefore, improve and maintain both service and network quality and reliability of the Next-Generation Network (NGN) and services.

1.1 Background and Motivation

The development of the network equipment and infrastructure will try to improve speed, bandwidth, and support various types of data transmission and online services. The mixture of data, voice and multimedia will become the future trend of the telecommunication services [1]. The NGN aims to provide and support huge amount of bandwidth and high data transmission rate. New value added services and applications can profit from such network. With the growth and wide use of the internet technology, a combination of different types of data and services is moving toward IP or packet-based platform. Therefore, not only the convergence of data, the solution of convergence of different access technologies such as mobile and fixed network infrastructure have been recently researched and developed for the NGN [2]. The IP multimedia subsystem (IMS) architectural framework is NGN architecture and is created to provide and support the convergence of data and networks (wired and wireless) on an IP-based platform [3]. The IMS has been developed by the 3rd Generation Partnership Project (3GPP) and the 3rd Generation Partnership Project 2 (3GPP2). The Session Initiation Protocol (SIP), developed by the Internet Engineering Task Force (IETF), was employed via 3GPP Release 6 as the standardized IMS signaling protocol [4]. IMS is not only supported real-time

multimedia services but also provide end-to-end QoS guarantee [5, 6]. However, there are many challenges implementing the IMS technology on NGN such as interoperability between traditional communication and various versions of packet-based technologies, guaranteeing end-to-end QoS, and providing suitable network management issues [7, 8, 9]. Therefore, not only QoS but also the reliability of new services and networks is undoubtedly needed for development toward NGN. The managing of network parameters via regular QoS term will not be sufficient for future convergence networks and services [10]. This thesis describes the study of the new concept of the quality term called Quality of Resilience (QoR) , reliability, availability, and its corresponding network parameters [10, 11]. Moreover, the reliability characteristic of the converged wired and wireless for NGN will be represented by reflecting on the IMS for further consideration of its performance via QoS and QoR . This Ph.D. study focuses on the investigation of the network parameters that will influence the overall reliability and performance of the NGN in term of end-to-end QoS and QoR.

1.2 Quality of Service and Quality of Resilience

The term QoR is an extended idea of QoS [10, 12], therefore, the understanding of the term QoS is needed in order to understand QoR. The International Telecommunication Union (ITU) recommendation E. 800 established a definition of QoS as "The collective effect of service performances that determine the degree of satisfaction of a user of the service." Various QoS terms have been defined to serve three main communication architectures: Asynchronous Transfer Mode (ATM) within ITU and the Integrated service (IntServ) and Differentiated services (DiffServ) architectures defined by the Internet Engineering Task Force (IETF). Nevertheless, DiffServ is most likely desired to implement on the IP-based network especially on Virtual Private Network (VPN) or Voice over IP (VoIP) services by Internet Service Providers (ISP) [13]. The measurement of service quality based on recovery methods has gained more interests [14, 15]. The QoR can be defined as the quantitative way to estimate the survivability and maintainability of the network. It is used to compare different types of recovery schemes implemented on the network by consideration of network downtimes over long operation time [12, 16]. The QoR is useful for network operator when making a decision on proper recovery schemes for different services with a reasonable price provided to a client than the traditional QoS method [14, 16].

The information of network parameters (factors) is the key information for quantifying QoS as well as QoR. To illustrate, the main QoS parameters of the communication network are availability, bandwidth, delay, jitter, and loss. The

purpose of QoR is to provide a detailed characterization of the probability of having available service [16]. The QoR adjusts QoS into short-term and long-term characteristics by focusing on survivability or service availability [10, 12]. This adjustment can be done by classification of some network factors into the short-term called availability parameters and the long-term called QoR parameters [12, 16]. For that reason, QoR parameters can only be measured with long-term quality factors. The short-term parameters are also important due to it can measure the short period (Δt) of user satisfaction. Then the satisfied or unsatisfied results can represent two service states: available and not available respectively. The results are called down and available period and can be represented by binary values: 0 and 1 respectively. Then, the summation of all variation of downtime periods can be calculated as downtime distribution and can be used to evaluate long-term QoR parameters [10, 12].

Therefore, the detailed study of the QoR parameters is further provided. The QoR will be useful for a network operator in consideration of resilience as a quality additive factors. These factors are very important for many delicate or financial services or medical applications such as eHealth [17].

1.3 Quality of Resilience and Challenges

This thesis focuses on finding new methods to evaluate end-to-end reliability or resilience of the IMS-based network to serve and improve overall reliability and resilience of the network and system. However, the reliability of the systems depends on many factors such as hardware, software, environment and operating regulations as mention in the Chapter 2. This section highlights some challenges and issues in system resilience evaluation.

Evaluation of system reliability: the exact reliability value of each system component is difficult to predict and the prediction value is, even more, difficult when considering the reliability of the whole system.

Network parameters: reliability estimation involves several network parameters. These parameters need to be carefully chosen to enhance the accuracy and decrease fault of an evaluation. Therefore, the key parameters, which are essentially related to reliability, are needed to be reviewed and carefully selected for an evaluation.

Reliability model: reliability quantity, in theory, is a success probability or probability of no failure (Reliability=1-Probability of Failure). It involves mathematics and stochastic parameters. Besides, there are no standard stochastic evaluation models and methods which will work for all types of systems. Therefore, suitable stochastic models and methods are needed to be established for future communication network or system, as well as, a new reliability evaluation framework.

Validation, optimization and cost: there is always an uncertainty in most of the evaluation processes. Therefore, simulation of the evaluation models is needed for validation and prove of the purpose concept model. Furthermore, in reality, it is difficult to find a real root cause and prevent every failure from happening. Therefore, there are various costs and outcomes after each failure. Accordingly, an effective preventive, maintenance, or resource optimization plan are needed to improve the reliability of the system.

Therefore, based on the above challenges, this thesis aims to address the following research objectives mentioned in Section 1.4.

1.4 Thesis Objectives

Based on the challenges mentioned in Section 1.3, this thesis focuses on end-to-end system reliability estimation method and models for improving the overall reliability of the IMS-based network: inter- and intra-network communications. To achieve this, an evaluation framework needs to be defined together with the key network parameters that can represent overall system reliability characteristics for the evaluation. In particular, the thesis makes a comprehensive study on three main important challenges mentioned above: end-to-end reliability or resilience evaluation framework and models, important end-to-end network parameters and their effect on the network reliability and performance, optimization, and validation of the evaluation methods. The main objectives of the thesis are summarized as follows:

- To design an IMS-based reference network architectures that can be adopted to evaluate end-to-end system reliability at various communication scenarios.
- To carry out a comprehensive study on the impact of network parameters to the reliability of the system and to identify the key end-to-end network parameters that can represent reliability characteristics for an overall system reliability evaluation.
- To develop end-to-end reliability evaluation framework and models to estimate the end-to-end reliability behaviors of the IMS-based network for both intra-domain and inter-domain communications.
- To optimize the key network parameters or factors and cost to improve end-to-end system reliability and to validate the proposed evaluation methods and models.

1.5 Contributions of the Thesis

The contributions of this thesis addressing the challenges through the research objectives asserted before are summarized below, the corresponding publication(s) relating to the contributions are stated in brackets.

- The end-to-end availability and reliability model are developed through a combination of stochastic models and Reliability Block Diagrams (RBD). The contribution is the proposed stochastic models: five-state Continuous-time Markov Chain model and Markov Reward Models (MRMs), and validation of the proposed models through a detailed numerical analysis of the proposed models and the state-of-the-art models. Moreover, the redundancy effect models are also formed to evaluate the overall system availability and reliability characteristics of intra-domain and inter-domain IMS communication scenarios (Chapter 3, and Chapter 4, publication: *b,c,e*).

- The key network parameters are examined and arranged for representing the end-to-end reliability of the IMS-based system. Moreover, the resilience of different reference IMS-based network topologies had been simulated and compared by using the well-known network simulator: OPNET. The simulation scenarios include IMS-based communication within similar registered home IMS domains, across registered home IMS domains, and communication across multiple IMS domains to describe end-to-end reliability behaviors of diverse IMS-based communications. Besides, the reliability impact when adding redundancy and communication traffics are exhibited (Chapter 2, and Chapter 5, publication: *d,f*).

- A new reliability evaluation frameworks are established where various state-of-the-art quality terms are investigated toward user satisfaction which is the most important quality objectives for modern networks and applications. Besides, new selective resilience parameter algorithm and the modern reliability evaluation method are introduced by using the modern Bayesian inference.The proposed algorithm can present practical reliability analysis and can implement as a defensive failure plan. Besides, the estimation approach can incorporate both subjective and objective parameters into the system reliability evaluation (Chapter 6, publication: *a*).

- The parallel redundancy effects of different IMS-based network communication scenarios have been analyzed and simulated for analyzing system availability and reliability characteristics with different redundancy conditions. Moreover, optimization of different parallel redundancy conditions has been simulated and compared to exhibit opportunity to design

high reliability and availability system for the Next Generation Networks (Chapter 4, publication: *b*).

1.5.1 Publications

Journals

a) C. Kamyod, R. H. Nielsen, N. Prasad and R. Prasad,"A Novel Estimation Frame-work for Quality of Resilience," *Wireless Personal Communications*, (accepted), 2016

b) C. Kamyod, R. H. Nielsen, N. Prasad and R. Prasad,"End-to-End Reliability and Optimization of Intra and Inter-domain IMS-based Communication Networks,"*IET Communications*, (under review), 2016

Conferences

c) C. Kamyod, R.H. Nielsen, N.R. Prasad, and R. Prasad, "End-to-End Availability Analysis of IMS-Based Networks: Simplex and Redundant System", *In Wireless Communications and Networking Conference (WCNC)*, (pp. 1103-1108). IEEE. April 2013.

d) C. Kamyod, R.H. Nielsen, N.R. Prasad, and R. Prasad, "IMS intra-and inter domain end-to-end resilience analysis", *In Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE)*, 2013 3rd International Conference on (pp. 1-5). IEEE. June 2013.

e) C. Kamyod, R.H. Nielsen, N.R. Prasad, R. Prasad, "Resilience in IMS: End-to-end reliability analysis via Markov Reward Models", *In Wireless Personal Multimedia Communications (WPMC)*, 2012 15th International Symposium on (pp. 564-568). IEEE. September 2012.

f) C. Kamyod, R.H. Nielsen, N.R. Prasad, and R. Prasad, "Resilience of the IMS system: The resilience effect of inter-domain communications", *In Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE)*, 2014 4th International Conference on (pp. 1-4). IEEE. May 2014.

1.6 Thesis Outline

The rest of the thesis is structured as follows:

An appearance of the thesis is shown in Fig. 1.1,with the aforestated publications and their respective chapters. The current chapter (chapter 1) present

a basic overview of QoS and QoR and its relationship along with contemporary challenges in estimating end-to-end reliability for contemporary online services and networks, and also the original contributions of this thesis.

Chapter 2 provides an introduction to reliability and performance analysis of the IMS, together, with the communication signaling and its communication architecture. The chapter also presents an opportunity to provide end-to-end QoS and QoR with the core IMS elements. Moreover, the chapter draws out the key network parameters which related to system reliability in short and long term operating periods.

Chapter 3 contributes the IMS reliability and availability models. The state-of-the-art models have been studied and compared with the proposed models for both the advantages and disadvantages of representing reliability characteristics of a unique, redundant component and the whole system. The analytical results were validated by the simulation results and were shown how the proposed models have more advantages than the existing models. The proposed models aim to be employed for the later chapters. Moreover, the chapter introduces end-to-end availability and reliability analysis of various IMS-based communication scenarios: intra-domain, and inter-domain communication scenarios.

Chapter 4 extends from the previous chapter and explores the parallel redundancy effects of different IMS-based communication scenarios. Moreover, the chapter represents different reliability effects through a number of parallel redundancy conditions. Besides, an optimization of the number of redundancies, and end-to-end system reliability is determined an opportunity to further design high availability and reliability system.

Chapter 5 implement the analytical models into the effective network simulator, i.e., OPNET for simulating and comparing resilience characteristics of the various IMS-based communication situations: intra-domain communication, inter-domain communication, and multiple-domain communication. Moreover, the reliability effects when increasing communication traffics and adding redundancy are also simulated. The simulation results confirm the system reliability features through the essential network parameters.

Chapter 6 gives a comprehensive study on how various quality terms are related to each other and the relationship toward user satisfaction. Moreover, the chapter describes a modern reliability measurement framework. The framework aims to assess both quantitative and subjective reliability parameters. Therefore, user satisfaction can be measured along with other important network parameters. Moreover, the chapter outlines the selective resilience parameter algorithm. The algorithm can be applied to a regular system.

Chapter 7 concludes the thesis by revisiting the objectives of the study and summarizing the main findings of the study. The chapter also highlights the possible future work in relation to this study.

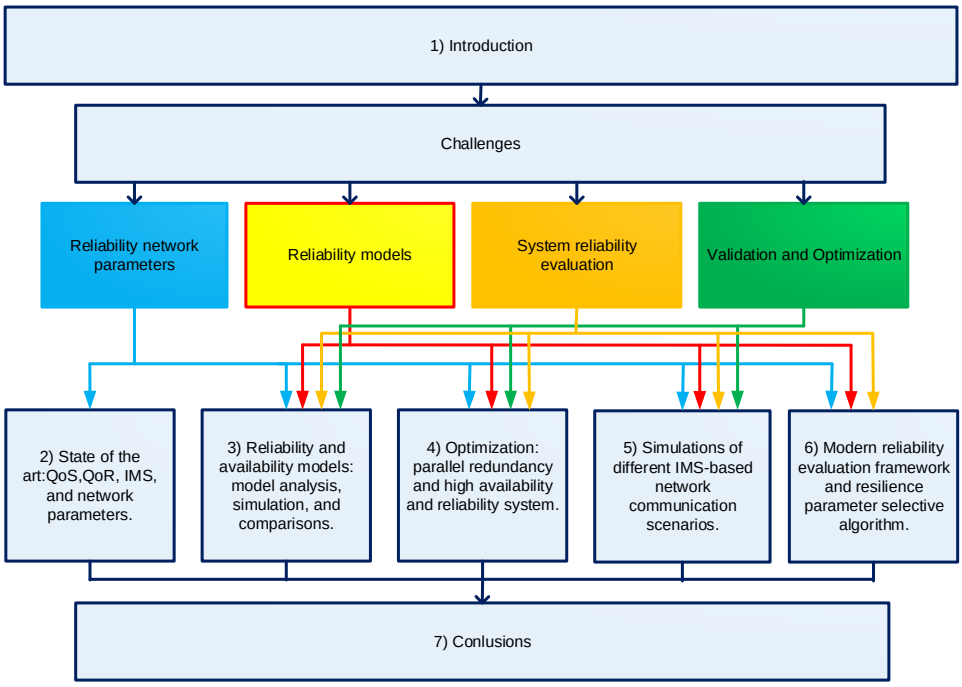


Figure 1.1: Block diagram showing inter-relations among different chapters.

1.7 References

[1] Harley R Myler. Network and media convergence: Issues, challenges and trends. In *Information and Communications Technology, 2007. ICICT 2007. ITI 5th International Conference on*. IEEE, 2007.

[2] Djamal-Eddine Meddour, Usman Javaid, Nicolas Bihannic, Tinku Rasheed, and Raouf Boutaba. Completing the convergence puzzle: a survey and a roadmap. *Wireless Communications, IEEE*, 16(3):86–96, 2009.

[3] Gonzalo Camarillo and Miguel-Angel Garcia-Martin. *The 3G IP multimedia subsystem (IMS): merging the Internet and the cellular worlds*. John Wiley & Sons, 2007.

[4] 3GPP LTE. Ip multimedia call control protocol based on session initiation protocol and session description protocol stage 3. *ETSI TS 124.229*, V12.10.0:854, 2015.

[5] 3GPP GSM. Generic access network (gan); mobile gan interface layer 3 specification. *ETSI TS 144.318*, Version 12.0.0, Release 12:252, 2014.

[6] TS ETSI. 185 001 v1. 1.1 (2005). *Technical Specification, Next Generation Network (NGN).*

[7] Mustafa Shakir. Challenging issues in ngn implementation and regulation. In *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, pages 1–4. IEEE, 2010.

[8] Shunsuke Uemura, Norihiro Fukumoto, Hideaki Yamada, and Hajime Nakamura. Qos/qoe measurement system implemented on cellular phone for ngn. In *Consumer Communications and Networking Conference, 2008. CCNC 2008. 5th IEEE*, pages 117–121. IEEE, 2008.

[9] Samir Chatterjee, Amitava Dutta, and Vinay B Chandhok. Introduction network convergence: Issues, trends and future. *Information Systems Frontiers*, 6(3):183–188, 2004.

[10] Piotr Cholda, János Tapolcai, Tibor Cinkler, Krzysztof Wajda, and Andrzej Jajszczyk. Quality of resilience as a network reliability characterization tool. *Network, IEEE*, 23(2):11–19, 2009.

[11] V Chandrakhumar, Óscar González de Dios, Juan Fernández Palacios, R Gruenzinger, Jordi Perelló Muntan, Salvatore Spadaro, IE Svinnset, E Zouganeli, P Cholda, Andrzej Jajszczyk, et al. The nobel2 approach to resilience in future transport networks. 2008.

[12] János Tapolcai, Piotr Cholda, Tibor Cinkler, Krzysztof Wajda, Andrzej Jajszczyk, and Dominique Verchere. Joint quantification of resilience and quality of service. In *Communications, 2006. ICC'06. IEEE International Conference on*, volume 2, pages 477–482. IEEE, 2006.

[13] Bruce Davie. Deployment experience with differentiated services. In *Proceedings of the ACM SIGCOMM workshop on Revisiting IP QoS: What have we learned, why do we care?*, pages 131–136. ACM, 2003.

[14] Piotr Chołda, Andrzej Jajszczyk, Bjarne E Helvik, and Anders Mykkeltveit. Service differentiation based on recovery methods. In *Proc. 2nd EuroNGI Workshop on Traffic Engineering, Protection and Restoration for NGI*, pages 21–22, 2005.

[15] Hamza Drid, Bernard Cousin, Miklos Molnar, and Samer Lahoud. A survey of survivability in multi-domain optical networks. *Computer Communications*, 33(8):1005–1012, 2010.

[16] János Tapolcai, Piotr Cholda, Tibor Cinkler, Krzysztof Wajda, Andrzej Jajszczyk, Achim Autenrieth, Stefan Bodamer, Didier Colle, Giuseppe Ferraris, Håkon Lonsethagen, et al. Quality of resilience (qor): Nobel approach to the multi-service resilience characterization. In *Broadband*

Networks, 2005. BroadNets 2005. 2nd International Conference on, pages 1328–1337. IEEE, 2005.

[17] Gunther Eysenbach and CONSORT-EHEALTH Group. Consort-ehealth: improving and standardizing evaluation reports of web-based and mobile health interventions. *Journal of medical Internet research*, 13(4), 2011.

Chapter 2

State of the Art: QoS and QoR

2.1 Introduction

To evaluate the total reliability characteristic of the system, the basic understanding about quality of service (QoS) is needed. This due to reliability is one of the quality portions. Therefore, improving overall system reliability will result in improving the overall system quality. This chapter presents reliability terms among various network quality terms, and also a chance to produce end-to-end QoS and quality of resilience (QoR) across the IP multimedia subsystem (IMS) communication architecture.

2.2 Session Initiation Protocol (SIP)

The Session Initiation Protocol (SIP) is an Internet application-layer protocol which initially developed by internet engineering task force (IETF) in 1999 [1]. It is created to provide signaling specifications and is used to set up calling sessions such as audio or video conferences, peer to peer communications. The SIP is a message or text-based client-server signaling protocol and works with another application layer protocol called Session Description Protocol (SDP). It is used to initiate, maintain, modify and terminate multimedia communications in the network. SIP is also selected to be a signaling protocol for the IMS. Therefore, its architecture and protocol need to be studied. The SIP composes of three main components which are SIP User Equipment or Agents (SIP-UA), SIP proxies (SIP-P) and SIP Registrars (SIP-R). The SIP-UA will function as the caller or service requester, the SIP-P function as relays of SIP

messages, and the SIP-R function as SIP directories. The relationship between SIP components can be represented by figure 2.1.

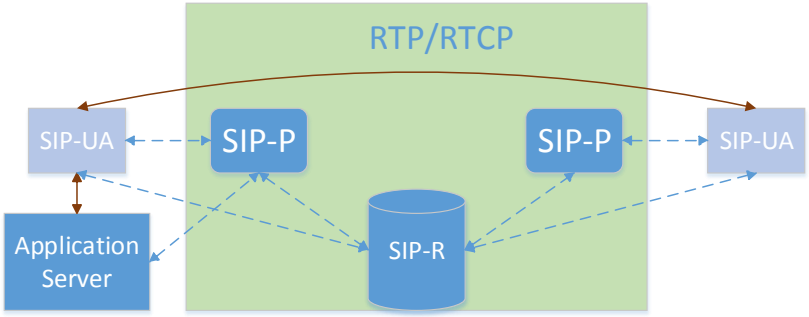


Figure 2.1: Basic SIP component system

From 2.1, the dashed line and dark line represent the signaling and data transmission path respectively. The signaling is needed to communicate between components such as communication between SIP-P and SIP-UA, between SIP-UA and SIP-R, or between each SIP-P. Therefore, increasing the component numbers will increase the need of signaling and complexity of the SIP network. The actual data or traffic flow will use a different path and Real-Time Transport Protocol (RTP) and Real Time Control Protocol (RTCP) for transmission. Many SIP proxies (within similar or different SIP communication domains) that relay SIP or SDP call setup messages may be used during communication setup between SIP User Agents. The SIP-UA is identified by a Universal Resource Identifier (URI) which is similar to the Universal Resource Locator (URL) standard.

2.3 The IP Multimedia Subsystem

IMS is an open industry standard for telecommunication carriers or communication of voice and multimedia over packet based IP network [2]. The concept of IMS is to combine fixed and mobile network services together called Fixed-Mobile convergence (FMC). It was first proposed by the 3rd generation partnership project (3GPP). The SIP is employed in the IMS as the signaling protocol for managing real-time multimedia sessions.

2.3.1 IMS architecture

The IMS framework composes of many layers such as Transport Layer, Service or Application Layer, IMS Layer and the User Layer. The IMS layer will act as

a control layer or center between Service Layer and Transport Layer as shown by figure 2.2. The Access layer of different access technologies will be attached in the Transport Layer. The user end device (UE) will be located at the User Layer. The signaling and service management are needed to connect between users across different access technologies at the Transport layer through the IMS layer [2]. Then the IMS will allow users to access services layer without the limitation of different access technologies of the users. The simplified IMS architecture can be represented per figure 2.2. To understand how the IMS is

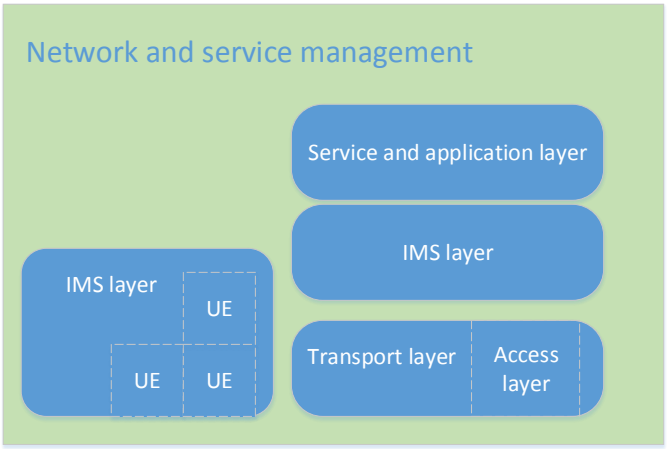


Figure 2.2: Simplified IMS Framework

operated, the main functional modules or components of the IMS architecture need to be studied. The precise detail of the IMS functional modules will be given and the main IMS architecture component and it can be represented by figure 2.3.

Call Session Control Functions (CSCF)

The call session control functions (CSCF) compose with four elements: Proxy-Call State Control Function (P-CSCF), Serving-Call State Control Function (S-CSCF), Interrogating-Call State Control Function (I-CSCF) and Emergency-Call State Control Function (E-CSCF). Their main functions and operations are explained below.

- P-CSCF: the P-CSCF will act as the first contact point between UE and the control functions. It is the SIP proxy that provides all subscriber access tasks to the multimedia services. Therefore, the requested or terminated of the signaling will be sent to or from UE by P-CSCF. The main tasks of P-CSCF are providing SIP compression that take place between

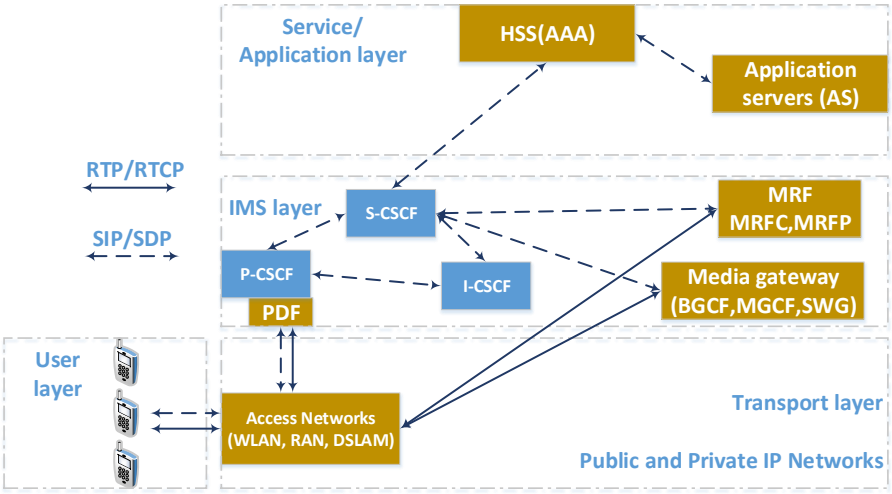


Figure 2.3: The main IMS architecture components [2]

UE and P-CSCF to reduce the packet size due to large information in the signaling header. The IPSec security association to provide secured SIP signaling, interaction with Policy and Charging Rules Function (PCRF) and emergency detection.

- S-CSCF: This SIP proxy is the central node of the signaling. It handles registration processes and controls the session or call state when the service is requested through Service Layer and also responsible for making routing decisions and maintaining session states and storing the service profiles.

- I-CSCF: It is the SIP proxy that locates and registers roaming users and also acts as the contact point between or within network operators. I-CSCF performs three unique tasks. First, obtaining the name of the next hop from the Home Subscriber Server(HSS). Second, assigning an S-CSCF based on received capabilities from the HSS. Third, routing incoming requests to an assigned S-CSCF or the application server. The P-CSCF, S-CSCF and I-CSCF perform functions during registration and session establishment. The corresponding functions can be achieved via both P-CSCF and S-CSCF such as release sessions due to user activities for example hanging session is detected via S-CSCF or user lost signal is detected via P-CSCF side.

- E-CSCF: It is the SIP proxy that responsible to handle IMS emergency requests such as police or ambulance requested services. It will select

proper emergency services that will be delivered based on calling or user location.

User Equipments (UE)

It is the end user system that provides necessary SIP signaling protocol and services related to media codec to support various access technologies.

Home Subscriber Server

It is the main IMS database server that consists of Home Location Register (HLR) functions for mobile users and all IMS user profiles. It provides authentication, authorization and management functions of the subscriber. The HSS will provide all information requested by CSCF before the CSCF create SIP connections for making a call setup.

Policy Decision Function (PDF)

It is a policy based management server that provides logical policy decisions based on QoS and Security received via P-CSCF. This function will gather data from the HSS database and use the RFC 2748 common policy service protocol to control policy enforcement points (PEP).

Application Services (AS)

It is a collection of multimedia application servers that provide and support various multimedia applications.

Media Server

The media server in IMS is called Media Resource Function (MRF) and responsible for controlling and processing by mixing, announcements, analysis and transcoding of the media streams. The MRF provides a source of media in a home network. It is divided into a signaling plane node called the Media Resource Function Controller (MRFC), and a media plane node called the Media Resource Function Processor (MRFP). The MRFC acts as SIP user agent and contains SIP interface towards the S-CSCF. The MRFP will response to all media-related functions.

Media Gateway (MG)

media gateway (MG) will convert data (voice or multimedia) from Public Switched Telephone Network (PSTN) or time division multiplexing base system into an internet protocal (IP) base system or vice versa. Three main components can be represented as the Media Gateway (MG): Media Gateway

Control Function (MGCF), Border Gateway Control Function (BGCF), and Signaling Gateway (SGW).

- The MGCF communicates with the CSCF and controls the connections for media channels and performs protocol conversion between ISDN User Part (ISUP) and the IMS call control protocols. Also, it helps exchanging voice over IP (VoIP) and multimedia packets with PSTN.
- The SGW use to map IMS SIP-based call control messages into PSTN number seven signaling protocol and messages.
- The BGCF will function as session border controllers and will control or signaling messages to another IMS domain. Therefore, the main function of media gateway is to perform a bearer inter-working between real-time transport protocol or IP network and the bearer of PSTN or legacy networks.

2.3.2 IMS architecture reference points

From the IMS architecture of the , there are many functional components and the connection between each component will depend on some specific protocols. Therefore, reference points were defined to reduce the complexity and to define the reference connection between components along with its supported protocol. There are two main protocols out of many different protocols that are used in the IMS layers (from the IMS core to services layer) which are SIP or SDP, and DIAMETER. The SIP and SDP will be used for signaling. The DIAMETER protocol will provide both security and signaling function. So the DIAMETER is normally used to handle the signaling to or from the HSS. The IMS reference points can be classified into six groups: G, M, C, D, S and U. The G group represents the group of GSM or GPRS serving node (GGSN). The M group represents the group related to the IMS layer. The C and D group will represent the access around the HSS component from the IMS layer when subscriber information is retrieved from the HSS. The S group concern with access to the HSS that are initiated by functional modules or specialized servers within the application or service layer. The U group represents the user when access to the IMS functional module by using standardized browsers and this will be supported by the hypertext transfer protocol (HTTP) or transmission control protocol (TCP) protocols. The details of the IMS reference points and interface description is shown in figure 2.4 and Table 2.1 below:

2.4 Important Network Parameters for QoR

As discussed before that some traditional QoS parameters that are correlated to survivability or service availability can be classified as short-term and long-term

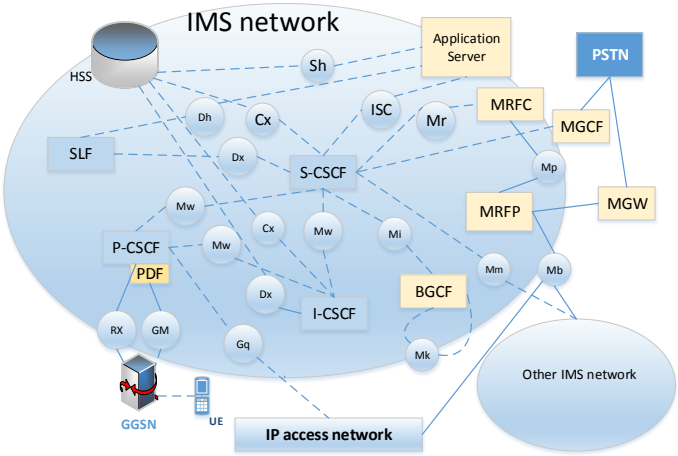


Figure 2.4: The IMS architecture components with reference points

QoR parameters for evaluation [3, 4]. The QoR parameters and its meaning corresponded to the reliability of the network can be classified as follows [3, 5, 6].

2.4.1 Long-term quality measurement

The parameters in this group can be used to measure the overall service quality for the whole service operating period.

- Mean Up Time, (MUT): the MUT can be defined as the period that service can be properly provided or perform a required function to a user [7]. It is difficult to measure this value sometimes due to it concern with working time of all equipment in consideration paths. For example along the original and ending (OD) path which have many network pieces of equipment along the path would consider having higher failure risk or less MUT. This due to the longer path than the OD path that equipped with less amount of network types of equipment. Therefore, it is difficult to calculate directly MUT in a complex network. Therefore, the value of MUT will be approximated by the Mean time to recovery (MTTR) and Mean time between failure (MTBF) [3, 7].

- Mean Time to Recovery, (MTTR): the MTTR refers to the duration of the time between the failure state and the consecutive working state (recovered after failure). The recovery time can be calculated in a practical situation by the summation of time spending on recovery mechanisms [3]. There are four main processes of the mechanism that are fault detection,

Table 2.1: The list of IMS architecture reference points and description

Interface name	Involved entities	Description	Protocal
Gm	UE,P-CSCF	The reference point is used to exchange messages between UE and CSCF.	SIP
Mw	P-CSCF,I-CSCF,S-CSCF	The reference point is used to exchange messages betwenn CSCF.	SIP
ISC	S-CSCF, I-CSCF, AS	The reference point is used to exchange messages between CSCF and AS	SIP
Cx	I-CSCF, S-CSCF, HSS	The reference point is used to communicate between I-CSCF/S-CSCF and HSS.	Diameter
Dx	I-CSCF, S-CSCF, SLF	The reference point is used by,I-CSCF/S-CSCF to find a correct HSS in a multi-HSS environment.	Diameter
Sh	IM-SSF, HSS	The reference point is used to exchange information between IM-SSF and HSS	MAP
Dh	SIP AS. OSA, SCF, IM-SSF,HSS	The reference point is used by AS to find a correct HSS in a HSS multi-HSS environment	Diameter
Mm	I-CSCF, S-CSCF,external IP network	The reference point will be used for exchanging messages between IMS and external IP networks	Not specified
Mg	MGCF ->I-CSCF	MGCF converts ISUP signaling to SIP signaling and forwards SIP signaling to I-CSCF	SIP
Mi	S-CSCF ->BGCF	The reference point is used to exchange messages between S-CSCF and BGCF	SIP
Mj	BGCF ->MGCF	The reference point is used to exchange messages between BGCF and MGCF in the same IMS network	SIP
Mk	BGCF ->BGCF	The reference point is used to exchange messages between BGCFs in different IMS networks	SIP
Mr	S-CSCF, MRFC	The reference point is used to exchange messages between S-CSCF and MRFC	SIP
Mp	MGCF, MRFP	The reference point is used to exchange messages between MRFC and MRFP	H.248
Mn	MGCF, IMS-MGW	The reference point allows control of user-plane resources	H.248
Ut	UE,AS(SIP AS OSA SCS, IM-SSF)	The reference point enables UE to manage information related to his services	HTTP
Go	PDF, GGSN	The reference point allows operators to control QoS in a user plane and exchange charging correlation information between IMS and GPRS network	COPS
Gq	P-CSCF, PDF	The reference point is used to exchange policy decisions-related information between P-CSCF and PDF	Diameter
Ro	AS, MRFC, S-CSCF, OCS	The reference point is used by AS/MRFC/S-CSCF for online charging towards OCS. Note: there might exist an interworking function between the S-CSCF and OCS.	Diameter
Rf	P-CSCF, S-CSCF, I-CSCF, BGCF, MGCF, AS, MRFC, CDF	The reference point is used by IMS entities for offline charging towards CDF.	Diameter
Rx	P-CSCF, AS, Charging Rule Function	The reference point allows dynamic charging-related service Function information to be exchanged between Charging Rules Function (CRF) and IMS entities. This information is used by the CRF for the selection and completion of charging rules.	Diameter

fault localization, fault notification and recovery. The fault detection time is the time that all errors or degradation of the signals can be detected. The fault localization time is the time that use to locate where the

error occurred by using Link Management Protocol (LMP). The fault notification and recovery time will be the time that uses to send the alarm signal and the time that use to recover the signal respectively. These times also depend on the network conditions and operator decision [3].

- Mean Time Between Failure, MTBF or Mean Time Between Interruptions, (MTBI): this term refers to the time duration between two failures or two decreased quality events.

- availability (A). This parameter has a dominant effect on QoS and QoR. This due to network unavailability even for a short period will cause unpredictable network performance perceived by users. In telecommunication practice or the context of network design, the availability is called steady state availability. It can be expressed as the probability of the system or network in working conditions or so-called upstate for some periods of time, t [7]. Moreover, the so-called instant availability is also given and defined as the probability of the system in the upstate at a given instant of time. Moreover, the availability can be considered in term of service availability and resource availability. The service availability is the probability that user find transport service working with some desired quality. The resource availability refers to the probability of the physical path that is in working condition after a failure. The availability also represents the ability of a given connection path that can be recovered after a failure. Therefore, availability can somehow represent resilience to failure of a system. However, there is a distinction between reliability and availability. The reliability is more about task completion or the probability of the system working with no interruption for some period to finish the task. Availability is related to reliability due its definition defined by the ITU-T recommendation E.800 as "The ability of an item to be in a state to perform a required function at a given instant of time or at any instant of time within a given time interval, assuming that the external resources, if required, are provided". Consequently, the reliability refers to free failure operation during an interval of operating time but availability refers to the ability or ready to work condition during that period. There are three different availability aspects and can be mathematically presented as follow [8].

- Instantaneous Availability, (A(t)): the A(t) or a single point availability represents the probability of a system or component function properly at an instant time, t. If not including repair or replacement periods, the A(t) is considered to be equal to instantaneous reliability (R(t)), R(t). There are two instantaneous availability conditions:
  a) There is no failure, and the system or component are functioning since the beginning until the instant time, t. Therefore, for the period of $(0, t]$

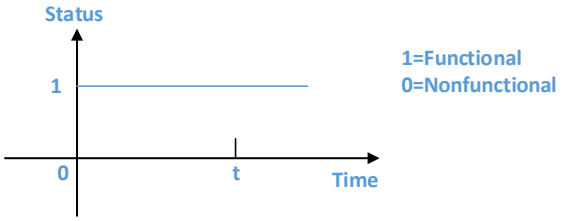


Figure 2.5: (a) Graphical representation of instantaneous availability

with no failure the functional probability is equal to $R(t)$.
b) The failure occurred at the time, x and $0 < x < t$, and with failure density equal to $m(x)$. The functional probability is given by:
The functional probability is equal to $\int_0^t R(t-x)m(x)dx$.

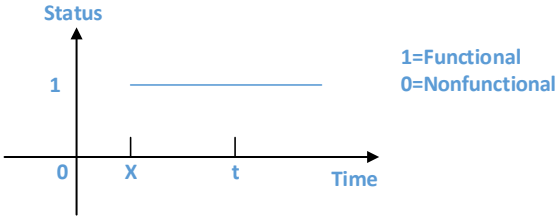


Figure 2.6: (b) Graphical representation of instantaneous availability

Therefore, the A(t) is given by

$$A(t) = R(t) + \int_0^t R(t-x)m(x)dx \tag{2.1}$$

- Steady state availability or limiting availability, A: the steady state availability can be defined as the instant availability when taking its limit into infinity. Therefore, $\lim_{t\to\infty} A(t) = A$. However, when taking the limit of reliability into infinity, its value will be equal to zero: $\lim_{t\to\infty} R(t) = 0$. The availability can be given by equation 2.2

$$A = \frac{MUT}{(MUT + MDT)} \tag{2.2}$$

According to the equation 2.2, A is related to MUT, If there is no preventive maintenance time, the system is considered to have MUT= mean time to failure (MTTF)= MTTR . Therefore, A can be rewritten as equation 2.3

$$A = \frac{MTTF}{(MTTF + MTTR)} = \frac{MTTF}{MTBF} = \frac{MTBF - MTTR}{MTBF} \tag{2.3}$$

The MTTF refers to the time start counting from the upstate (previously down) until the first failure occurred. The MTTR is the time duration start counting from the first failure occurred until the system is recovered or in upstate. Therefore, $MTTF + MTTR$ is equal to $MTBF$. So the value of "A" can be calculated from failure and repair point of view. The graphical representation of MTTF, MTTR, and MTBF can be represented by using figure 2.7 below.

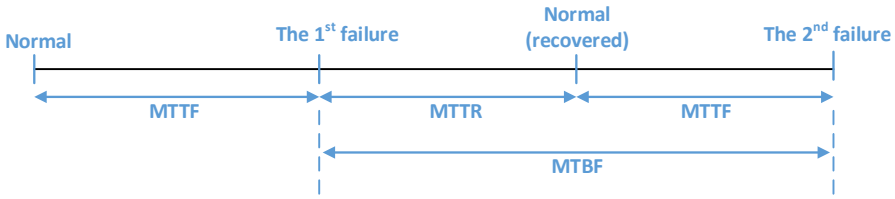


Figure 2.7: The time line and events represent the relationship of MTTF, MTTR, and MTBF

The steady state availability can also be given by

$$A = \frac{\frac{1}{\lambda}}{(\frac{1}{\lambda} + \frac{1}{\mu})} = \frac{\frac{1}{\lambda}}{\frac{\mu+\lambda}{\lambda\mu}} = \frac{\mu}{\mu+\lambda} = \frac{MTTF}{MTTF + MTTR} \tag{2.4}$$

where λ is the failure rate.
μ is the repair rate.
$\frac{1}{\lambda} = MTTF$, is equal to Mean Time to Failure.
$\frac{1}{\mu} = MTTR$, is equal to Mean Time to Recovery.
Therefore, from equation 2.4, to obtain the higher value of A, the higher value of $MTTF$ or the lower value of $MTTR$ is needed. In case of the system with redundancy, A is given by

$$A = \frac{MTTF_{eq}}{MTTF_{eq} + MTTR_{eq}} or \qquad A = \frac{\mu_{eq}}{\lambda_{eq} + \mu_{eq}} \tag{2.5}$$

Where $\lambda_{eq} and \mu_{eq}$ are the equivalent failure rate and repair rate of the system respectively.

- Interval Availability , A_i(t) The interval availability can be defined as the average availability or the expected fraction of the time that the system is up in a given interval $(0, t]$. The interval availability is given by

$$A_I(t) = \frac{1}{t}\int_0^t A(x)dx \tag{2.6}$$

The relationship between three definitions of availability can be given as follows:

$$A(t) = \lim_{t\to\infty} A_I(t) = \lim_{t\to\infty} A(t) = \frac{\mu}{\lambda + \mu} \tag{2.7}$$

- Availability class: the availability class of the system will define how long the system will continuously function without interruptions. The availability can be classified by orders of magnitude as shown in 2.2 below [9].

Table 2.2: Classification of degree of availability and system type [9]

System Type	**Unavailability (minutes/years)**	**Availability (percentage)**	**Availability Class**
Unmanaged	50	90	1
Managed	5000	99	2
Well-managed	500	99.9	3
Fault-tolerant	50	99.99	4
High Availability	5	99.999	5
Very High Availability	0.5	99.9999	6
Ultra Availability	0.05	99.99999	7

For example, the availability of the unmanaged system can be calculated and explained as the computer system that has 90 % availability. So the unavailability (downtime) is equal to 0.1 per year or equal to 52.560 minutes ($0.1 \times 365 \times 24 \times 60 = 52.560$ minutes) or approximately 50.000 minutes per year. Moreover, also for the fault tolerant systems that have 99.99 percent availability means that the system fails 0.0001 or ($0.0001 \times 365 \times 24 + 60 = 52$ minutes) per year. Or the system will be repaired within a few hours.

- Affected Traffic or Traffic Loss: this value represents how much or how many of the services or signals will be disturbed by a failure. The failure is normally a temporal value that can be related only to the fault detection time. Occasionally the notion of the affected traffic is associated with the duplicated traffic. If it is only of a temporal character, the term can be dismissed and extend the recovery time to the point where the whole traffic is properly ordered and not duplicated. However, in such a case some information about the behavior of the selected scheme is lost.

- Cost of Recovery: this parameter consider the redundancy of the network resources that is utilized for recovery as a cost such as backup capacity or link bandwidth.

- Resilience to Multiple Failures: this factor will measure how many failures (single or multiple) that a network can handle with in term of the degree of value.
- Preemption: the preemption is any processes that take away backup path or recovery resources from one entity to provide another entity as another resource for recovery propose. There are two levels of preemption. First, the resource can be removed by others if simultaneous failures occurred, and the preemption one will not be recovered. The second level of preemption means that the resources will be taken not only when the failure occurred at the preemption connection but also when the failure occurred at other path and will, later on, effect the preemption connection. The level indicates when large failures occur and represent a low level of reliability.
- Failure Coverage: the failure coverage is the fraction of recovered traffic or connections and the failure circumstances. The value can represent the recovery efficiency. The high value of failure coverage means the high efficiency of recovery.

2.4.2 Short term quality measurement

The parameters in this group will be used to measure the instant service quality that is perceived by a user. Most of the network that consist of an IP service quality will mainly consider IP packet loss ratio for a short-term quality assessment.

- Packet loss ratio, Packet loss rate or IP Packet loss ratio (IPLR).
  These parameters will represent the ratio of loss packet out of transmitted packet. Alternatively, the ratio of the total packet loss per a total number of transmitted packet or interested transmitted packet. The network factors or parameters mentioned above are important and mainly used for QoR analysis. They can be classified and summarized in Table 2.3 below. The parameters in bold are considered to be the main parameters that will be used to analyzed or represent the reliability of the system. For example, the availability which can be represented by MTTF is normally used to analyze the reliability of the system [3, 5]. Also the packet loss probability or delay can represent the capability of the packet based network or can be used to evaluate and represent the performance of the system [3]. Therefore, this Ph.D. study will focus on these important QoR parameters. Its value will be varied by different types of service or application that need a different level of service quality.

Table 2.3: Summary of the QoR network parameters which is QoS network parameters that related to reliability

Long-term quality network parameters (QoR parameters)	**Short-term quality network parameters (instantaneous availability criteria)**
Mean Up Time (MUT)	Packet loss ratio or IP packet loss ratio (IPLR)
Mean Time to Recovery (MTTR)	Packet delay or latency
Mean Time Between Failure (MTBF) or Mean Time Between Interruptions (MTBI)	Packet loss probability
Mean Time to Failure (MTTF)	Instantaneous Availability (Availability at the instant time, t)
Mean Down Time (MDT)	Data rate, Throughput
Steady State Availability (A)	Bit error rate
Affected Traffic or traffic loss	Jitter (voice, video)
Cost	
Resilience of multiple failures	
Preemption	
Failure coverage	
Criteria on downtime distribution	

2.5 IMS and QoS

The traditional IP-based services have to face many problems: high delay and variable during packet transmission, out of order packet arriving, lost and discarded packets. In order to be the central of all IP-based services of next-generation networks (NGN), the end-to-end QoS is needed to be well defined and addressed in Service Level Agreement (SLM) to efficiently provide and guaranteed services between users and service providers. The network parameters or metrics are normally used to point out or measure QoS levels such as availability, bandwidth or throughput, response time, delay, jitter, and packet loss. The requirement of each parameter can be varied depending on application services types. The IMS is designed and developed to provide end-to-end QoS. The QoS requirements will be expressed via negotiation processes of UE during a SIP session setup or session modification procedure. The requirement of network parameters such as media type, traffic direction, bit rate, packet size, or packet bandwidth can be negotiated via UE. After complete negotiating processes at the application level, UE can reserve suitable resources from access network if not available in case of mobile access. After end-to-end QoS is created, the UE encode and packetize individual media types with an appropriate protocol and send these media packets to the access and transport network by using a transport layer protocol over IP. It is assumed that operators negotiate service-level agreements for guaranteeing the required QoS in the interconnection backbone.

2.6 Reliability and Performance Analysis of IMS

Reliability or resilience of the computer and communication system has become an important factor for future services and networks. Due to users increasingly depend on Internet technology which support and provide various important applications such as financial, medical, or even disaster warning systems. However, the quantification of resilience has not been fully studied especially for IMS or NGN.

The failures in telecommunication networks can be measured by different metrics such as availability, reliability, performance, performability (reliability and performance), and survivability. Availability is the ability of the system to perform required function at a given instant of time within a given time interval. Reliability is the ability of the system to perform a required function under given condition for a given time interval. Survivability is the ability of the system to perform a required function when some components in the system failed. The performability is the measurement technique that considers both reliability and performance aspects of the system. Dependability is another term that defined as the umbrella term of the trustworthiness of the computer or communication system. The dependability will represent all terms such as availability, reliability, performance or survivability of the system. The term security is another term that can be considered as the factor that affect and relate to reliability. Then confidentiality and safety can be considered together when looking at the overall resilience of the system. The dependability terms can be quantified as resilience or performance quantity by using various types of probabilistic analytic models [4].

The performance evaluation of the IMS has been proposed by using the central server Queueing Network Model (QNM) [10]. The central server queueing network model [11] has been applied based on the assumption that UE has to register with the IMS core per every multimedia session setup. Moreover, also the communications between UE and Application Server (AS) always go through or manage by the S-CSCF proxy according to the 3GPP specification [2]. The simulation and experimental setup for the registration and multimedia session setup were shown that the S-CSCF proxy or module was the bottleneck when the number of workload or session request increased. The research validates the central server queueing network model by showing that analytical and simulation results are comparable to each other. The same method had applied with handover mechanisms in IMS for performance evaluation [10].

The Queuing Petri Nets model (QPN) was formerly used to evaluate the performance of the IMS system [12, 13]. The QPN combines both queuing networks and stochastic petri nets which are the most popular theory tools of

network modeling formalisms. The model simply adds queuing and timing aspects to the places of traditional Stochastic Petri Nets model. So that QPN can integrate some advantages and eliminates some disadvantages of both methods. Then, different SIP schemes can be setup; for example, at SIP registration process, the HSS is needed for the signaling only for the first round but not for the next round of SIP signaling process. Moreover, the access and network domain security [14, 15] has been considered if the different cryptographic algorithm will affect the performance of the IMS. The studied shown that different security coding schemes will result in different delay and server utilization time and finally affect the throughput or performance of the network [14, 15, 16]. The methodology and designed of the service IMS reliability has been recently researched. The service reliability matrix, a significant point of failure which mainly impacts the network, and the source of failure has been addressed as the essential keys for analyzing IMS reliability [17].

The reliability of the IMS core network was analyzed by focusing on the failover success rate which can represent the impact of end-to-end reliability in operation at the network level of the IMS architecture by using three-state Markov model [18]. The research was shown that failover duration time and coverage factor is the main factors that impact expected downtime of the system with redundancy. The improvement of end-to-end reliability and performance has been performed by focusing on core IMS architecture [19]. The signaling setup delay is improved by co-locating the IMS servers on the same host. Moreover, IMS reliability analysis based on service in the context of IMS architecture, components and service distribution or customer usage pattern has been researched [20, 21]. The so-called hierarchical methodology was used by applying Discrete Time Markov Chain (DTMC) model that combine the aspects of component reliability, IMS architecture, and service distribution. The method was influenced by the previous work of architecture-based software reliability analysis [21].

Recent availability analysis model of the network with internal and external redundancy has been achieved by using Markov model. The geographical redundancy model shown to improve system availability for both user and system initiated redundancy mechanisms [22]. Moreover, reliability mechanism of IMS network had been proposed by having one redundancy of the S-CSCF unit in the system, and the mechanism was evaluated with the OpenIMS. The results showed that the IMS system seamlessly affect when having redundancy [23]. Also, the performance of IMS core network with parallel redundancy is evaluated [24]. The IMS core elements were modeled as a multi-state (failure and recovery) by using continuous-time Markov chain (CTMC). The performance is simulated by means of system availability and also optimization of parallel redundancy has been performed via Universal Generating Function (UGF). From above, a combination or different model-based analysis techniques have

Table 2.4: Summary of related works

	Model	Objective	Evaluated parameters	Modeling Scenario	Main research assumption
1.	Central Server Queuing Network (QNM) [16, 17].	Performance Evaluation.	Delay (response time) based on server utilization time of different specification of the server's hardware.	Ten users (UEs) requested multimedia session setups from five different application servers (APs) in the same home network.	1. Fixed different numbers of requests. 2. Server service time follow exponential distribution.
2.	Queuing Petri Nets (QPN) [12, 13].	Performance Evaluation.	Delay due to server utilization of register and session setup processes and different amount of arrival rate. 2. Delay plus delay due to different encryption mechanisms.	Two users (UEs) each located on two different domain (domain A and B).	1. Incoming traffic follow Poisson distribution. 2. Server service time follow exponential distribution.
3.	Network Level failover analysis using three-state Markov model [18] based on IMS architecture capabilities [17].	Reliability analysis with redundancy consideration Hardware and Software.	Mean down time with and without automatic recovery (failover)	Registration and multimedia session setup of a UE in their home network.	1. All components have no network level failover capability. 2. If failover is implemented, the failure recovery behavior will follow three-state Markov model
4.	Reliability analysis with hierarchical method. using Discrete Time Markov Chain (DTMC) [20, 17].	Reliability analysis.	1. Service path reliability. 2. Components reliability. 3. Session reliability which calculated as a weighted average of the path reliability.	Registration and multimedia session setup of a UE in their home network with different service or session requested. The basic call setup service to Presence service with three value of service path probability is evaluated.	Assumed that the reliability of each IMS component is 0.9999 (four nines).

been applied for reliability, availability and performance evaluation of IMS system. However, end-to-end reliability or availability based on different failure scenarios or different IMS architecture has not been adequately investigated.

From Table 2.4, there are many challenges and opportunities to perform the reliability or performance evaluation of the IMS architecture due to its complex architecture and different conditions of architecture, scenario, and security options. Therefore, the analysis of the IMS reliability or reliability and performance (performability) has not been fully researched.

2.6.1 End-to-end QoS and QoR.

The definition of the network component reliability can be defined as the probability of a network element (a node or a link) to be fully operational during a certain time frame [7]. Availability can also be defined as the instantaneous counterpart of reliability. Therefore, a reliability of the network can be measured via availability parameter or other network parameters that will affect service availability of the networks. The IMS use SIP or SDP protocol for signaling session flows across all networks between UE. As mentioned before, the signaling will be managed by the IMS layer or mainly by the CSCF modules. The data transmission path will be carried on a separate link by using real-time transport protocol and real-time transport control protocol. Besides, the data flow can be controlled by media gateways or media servers that are managed by IMS. To analysis the end-to-end reliability of the IMS system, the end-to-end QoS framework of the IMS need to be studied. Due to the QoR parameters are considered as a part of the QoS parameters that are related to reliability or availability of the considered network.

The figure 2.8 shows the end-to-end QoS architecture that is originally stated by the 3GPP document [25].The end-to-end QoS is defined to be the point between terminal equipment. If considering mobile terminal or a combination of computer and mobile as a user, the end-to-end QoS can be considered as the area between terminal equipment (TE) and mobile terminal (MT) local bearer service and external bearer service. Therefore, end-to-end IMS QoS will be mainly depended on the general packet radio service (GPRS) service. The GPRS QoS consist of two QoS parts: The Radio Access Network (RAN) QoS and the core network QoS. From the figure 2.9, the core network is defined between a core edge and a core network gateway that will provide connectivity to other or global networks. The RAN can be divided into two segments: the connection between MT and universal terrestrial GSM GPRS EDGE radio access networks (UTRAN/GERAN) and the connection between UTRAN/GERAN and the core edge node as shown in figure 2.9.

From figure 2.9, this can be noted that it will be very complex to control or guaranteed certain end-to-end QoS especially when the mobile users are moving. This due to the packets will be moved to different paths, and this will affect the network parameters and end-to-end QoS.

2.6.2 QoS scheme

The communication data in the IMS network can be classified into four traffic classes due to 3GPP QoS purpose which are:

- The conversation class for voice and real-time multimedia messages.
- The streaming class for streaming applications such as video on demand.

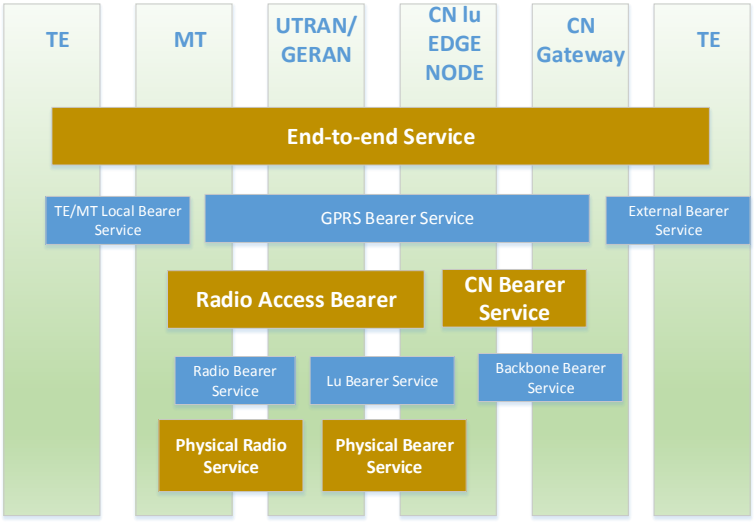


Figure 2.8: End-to-end QoS architecture

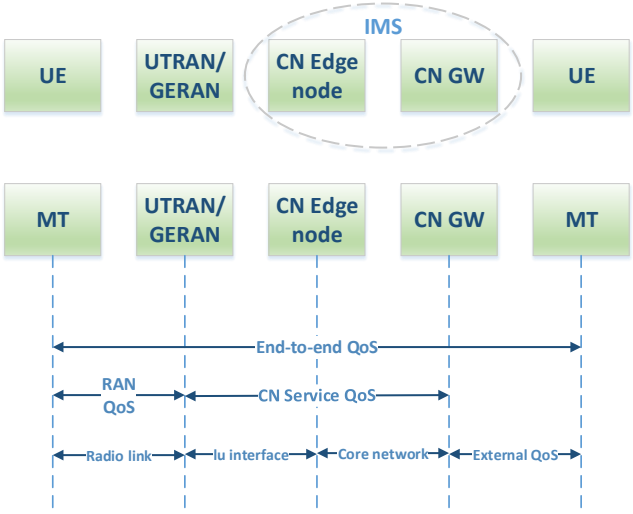


Figure 2.9: QoS characteristic between MT of the IMS architecture

- An interactive class for interactive applications such as web-browsing.
- Background class for background applications such as e-mail and FTP.

Therefore, the packet data can be treated or prioritized according to its QoS aspects. The policy-based control in the IMS can be used to manage or forward the packets by controlling the set of configuration parameters, and forwarding data based on its classes and QoS schemes: Resource Reservation Protocol (RSVP) or Differentiated Services (DiffServ). The session setup processes are needed to establish an NGN multimedia service. The UE need to negotiate its required QoS parameters to establish a service and resources reservation with other UE. So the QoS policies must be enforced during UE registration session setup procedures. Many IMS functional modules will be used during QoS parameter negotiation such as CSCF, HSS. The SIP and the SDP will be used as a signaling protocol to establish a session setup between UE. The SDP will be embedded with required QoS parameters such as bandwidth, type of media, transport protocol, codec type, loss and delay requirements. Considering figure 2.10 and 2.11 about registration and session setup procedures [26].

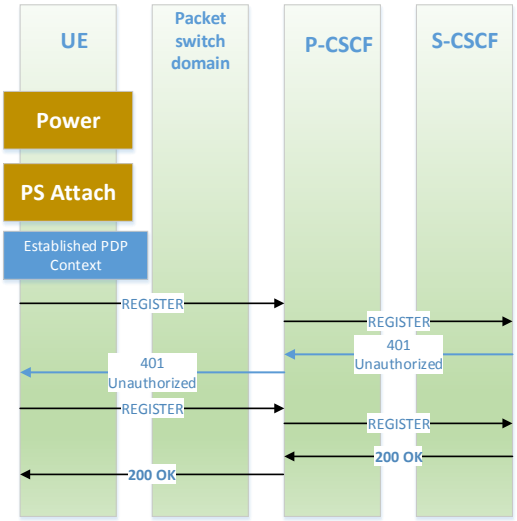


Figure 2.10: (a) Basic IMS registration diagram [27]

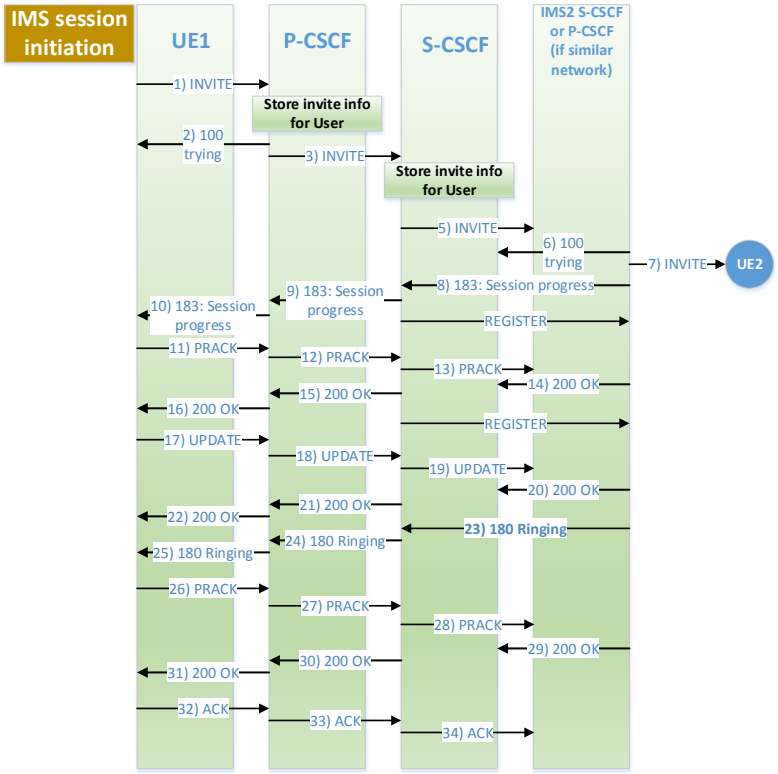


Figure 2.11: (b) Basic IMS session setup diagram [27]

When the UE number one (UE1) want to contact and communicate with the UEnumber two, UE1 sends SIP message (INVITE) to the PCSCF1 in which it specifies its QoS parameters. The P-CSCF authenticates UE1, checks the security of the SIP message, and forwards it to its S-CSCF. The S-CSCF authorizes the multimedia service requested by UE1 based on the service policy and the registration status of UE1 stored in the HSS. Then, the S-CSCF forwards the SIP message to the I-CSCF2 that is the entry point to UE2 if it is on a different network. Then I-CSCF2 will search for the S-CSCF2 that controls UE2 and forwards the SIP message to designated location. The S-CSCF2 in turn forwards the SIP message to the P-CSCF2. Then the UE2 receives the SIP message from its P-CSCF2 and sends a response SIP message (e.g., 183 Session Progress) to UE1 via the same IMS signaling path by specifying its desired QoS parameters. If UE2 is in the same network, the SIP message will forward from S-CSCF to the P-CSCF located near UE2 directly. This SIP message exchange between UE1 and UE2 is repeated until the QoS parameters are determined. When P-CSCF forwards a response SIP message from UE2 to UE1 with the

final negotiated QoS parameters, it will consult with the policy decision point (PDP) to verify the resource availability for the negotiated QoS parameters. If the PDP grants the necessary resources to UE1, P-CSCF forwards the SIP message to UE1 by informing that it can make a resource reservation. Further, when UE1 starts its resource reservation, it sends a SIP message (i.e., PRACK) to UE2 so that UE2 can start making its resource reservation after sending a response SIP message (i.e., 200 OK). The resource control and admission are commonly implemented by two functional modules which are PDP and the PEP. The PDP is co-located with a P-CSCF. Due to recent 3GPP release, a charging rule is added to a PDP and is referred as a PCRF [25]. The PEP is normally located at the edge router of UE. When the P-CSCF request resource from PDP after the final QoS parameter negotiation, PDP will determine how many resources should be allocated according to the registration status of the UE, network resource availability, and network policy. Then, the PDP maps the negotiated QoS parameters specified in SDP to some specific QoS parameters which are used and sends them to the PEP of the access network that serves the UE in order to enforce and allocate the determined resources to the UE when making a resource reservation. However, the Policy function will be different based on different standards such as IETF use PDP, 3GPP use policy decision function (PDF) and 3GPP version 7 use PCRF [28].

2.7 Conclusions

This chapter examined the state of the art in QoS and QoR with regards to the key network parameters for evaluating the reliability and performance of the IMS network architecture. In particular, this chapter focuses on various important network parameters and characteristics toward the classification of these parameters into the short-term and long-term reliability related parameters. This chapter also investigates the IMS architectures, the key IMS components, and the IMS signaling schemes to study how to provide end-to-end QoS and QoR based on the IMS architecture. Therefore, in order to evaluate end-to-end reliability and quality of the system, it is necessary to study how the key network parameters are measured and how the parameters affect system reliability for short-term and long-term operating period.

2.8 References

[1] Jonathan Rosenberg, Henning Schulzrinne, Gonzalo Camarillo, Alan Johnston, Jon Peterson, Robert Sparks, Mark Handley, Eve Schooler, et al. Sip: session initiation protocol. Technical report, RFC 3261, Internet Engineering Task Force, 2002.

[2] IP Multimedia Subsystem. stage 2. 3gpp ts 23.228, 2006.

[3] Piotr Cholda, János Tapolcai, Tibor Cinkler, Krzysztof Wajda, and Andrzej Jajszczyk. Quality of resilience as a network reliability characterization tool. *Network, IEEE*, 23(2):11–19, 2009.

[4] János Tapolcai, Piotr Cholda, Tibor Cinkler, Krzysztof Wajda, Andrzej Jajszczyk, and Dominique Verchere. Joint quantification of resilience and quality of service. In *Communications, 2006. ICC'06. IEEE International Conference on*, volume 2, pages 477–482. IEEE, 2006.

[5] János Tapolcai, Piotr Cholda, Tibor Cinkler, Krzysztof Wajda, Andrzej Jajszczyk, Achim Autenrieth, Stefan Bodamer, Didier Colle, Giuseppe Ferraris, Håkon Lonsethagen, et al. Quality of resilience (qor): Nobel approach to the multi-service resilience characterization. In *Broadband Networks, 2005. BroadNets 2005. 2nd International Conference on*, pages 1328–1337. IEEE, 2005.

[6] P Chołda, Andrzej Jajszczyk, and Krzysztof Wajda. A unified quality of recovery (qor) measure. *International Journal of Communication Systems*, 21(5):525–548, 2008.

[7] ITUT Recommendation. E. 800: Terms and definitions related to quality of service and network performance including dependability. *ITU-T August*, 1994, 1994.

[8] Robin A Sahner, Kishor Trivedi, and Antonio Puliafito. *Performance and reliability analysis of computer systems: an example-based approach using the SHARPE software package*. Springer Science & Business Media, 2012.

[9] Jim Gray and Daniel P Siewiorek. High-availability computer systems. *Computer*, 24(9):39–48, 1991.

[10] IM Mkwawa and DD Kouvatsos. Performance modelling and evaluation of handover mechanism in ip multimedia subsystems. In *Systems and Networks Communications, 2008. ICSNC'08. 3rd International Conference on*, pages 223–228. IEEE, 2008.

[11] Forest Baskett III. Mathematical models of multiprogrammed computer systems. 1970.

[12] Chuang Lin, Kai Wang, Lei Lei, Chanfang Liu, et al. Quality of protection analysis and performance modeling in ip multimedia subsystem. *Computer Communications*, 32(11):1336–1345, 2009.

[13] An'an Luo, Chuang Lin, Kai Wang, Fengyuan Ren, and Limin Miao. Performance modeling and evaluation using queuing petri nets in ims.

In *Communications and Networking in China, 2009. ChinaCOM 2009. Fourth International Conference on*, pages 1–5. IEEE, 2009.

[14] IP Multimedia Subsystem. version 13.1.0, 3gpp ts 33.203, 2015.

[15] IP Multimedia Subsystem. version 13.0.0 3gpp ts 33.210, 2015.

[16] Kai Wang, Chuang Lin, and Fangqin Liu. Quality of protection with performance analysis in ip multimedia subsystem. In *Computer and Information Science, 2009. ICIS 2009. Eighth IEEE/ACIS International Conference on*, pages 234–239. IEEE, 2009.

[17] Himanshu Pant, Chi-Hung Kelvin Chu, Steven H Richman, Ahmad Jrad, and Gerard P O'Reilly. Reliability of next-generation networks with a focus on ims architecture. *Bell Labs Technical Journal*, 12(4):109–125, 2008.

[18] Veena B Mendiratta and Himanshu Pant. Reliability of ims architecture. In *Telecommunication Networks and Applications Conference, 2007. ATNAC 2007. Australasian*, pages 1–6. IEEE, 2007.

[19] Thierry Bessis. Improving performance and reliability of an ims network by co-locating ims servers. *Bell Labs Technical Journal*, 10(4):167–178, 2006.

[20] Swapna S Gokhale and Veena B Mendiratta. Architecture-based assessment of software reliability. In *Quality Software, 2008. QSIC'08. The Eighth International Conference on*, pages 444–444. IEEE, 2008.

[21] Q Zhu, S Gokhale, and VB Mendiratta. Reliability analysis of ip multimedia subsystem. In *Proceedings of the international conference on contemporary computing (IC3) 2008*, pages 141–150, 2008.

[22] Xuemei Zhang, Hoang Pham, and Carolyn R Johnson. Reliability models for systems with internal and external redundancy. *International Journal of System Assurance Engineering and Management*, 1(4):362–369, 2010.

[23] Yong Liu, YiHong Liu, and Xingwei Wang. Reliability mechanism for the core control network-element s-cscf in ims. In *Network Computing and Information Security (NCIS), 2011 International Conference on*, volume 1, pages 57–61. IEEE, 2011.

[24] Maurizio Guida, Maurizio Longo, and Fabio Postiglione. Performance evaluation of ims-based core networks in presence of failures. In *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, pages 1–5. IEEE, 2010.

[25] IP Multimedia Subsystem. version 13.0.0 3gpp ts 23.207, 2015.

[26] V Chandrakhumar, Óscar González de Dios, Juan Fernández Palacios, R Gruenzinger, Jordi Perelló Muntan, Salvatore Spadaro, IE Svinnset, E Zouganeli, P Cholda, Andrzej Jajszczyk, et al. The nobel2 approach to resilience in future transport networks. 2008.

[27] Travis Russell. *The IP multimedia subsystem (IMS): session control and other network operations*. McGraw-Hill, Inc., 2007.

[28] Christian Esteve Rothenberg and Andreas Roos. A review of policy-based resource and admission control functions in evolving access and next generation networks. *Journal of Network and Systems Management*, 16(1):14–45, 2008.

Chapter 3

The Proposed IMS Reliability and Availability Models.

3.1 Introduction

The computer network and system comprise of various components which perform several individual functions to support the main function of the system. These components may operate with various reliability and performance level. Evaluation of system reliability is, then, quite complicated and challenging. The model-based and simulation techniques are one of the foremost solutions for representing system reliability and performance. This due to, measuring and handling the system reliability results from a real network and system is very tough and costly. Moreover, the model-based techniques are useful for designing modern network.

In this thesis chapter, the IMS-based reference communication networks are proposed. Besides, end-to-end reliability and availability evaluation framework and model are developed by using a combination of reliability block diagram (RBD) and continuous-time Markov chain (CTMC). The term "end-to-end" refers to the connection between UEs across an IMS-based network. Moreover, the proposed models are analyzed and compared with the state-of-the-art models for both with and without redundancy cases. Besides, The simulation results are proven that the proposed models outperform the state-of-the-art models by representing better-detailed failures and recovery characteristics.

3.2 The Proposed IMS Setup Scenarios

The proposed analysis scenario idea is that the end-to-end reliability should be analyzed from the network scenario that has, at least, two users. So the scenario will have user types of equipment locate on different or similar home network (for example domain A and B). Further, the sub scenario can be considered in a case of different access technologies, for example, wireline, wireless or legacy public switched telephone network (PSTN) network at one end or another end as shown in 3.1.

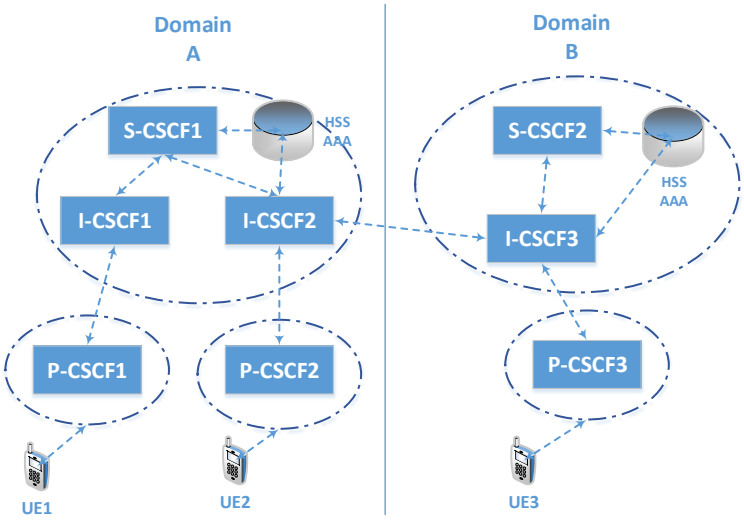


Figure 3.1: The IMS communication setup scenario between two UEs at different visited network and registered home domains

3.3 Reliability Analysis and Model

For essential of an analysis, the functional module or network component state can be assumed with two states (up and down) and the system can assume to have an independent property that a failure of one component did not depend on a failure of other components. The RBD can be used to quantify the reliability and availability of the IP multimedia subsystem (IMS) setup scenarios per figure 3.1. For example, the first scenario, registration and multimedia session setup between UE1 and UE2, and between UE1 and UE3. The RBD can be presented by figure 3.2 and 3.3.

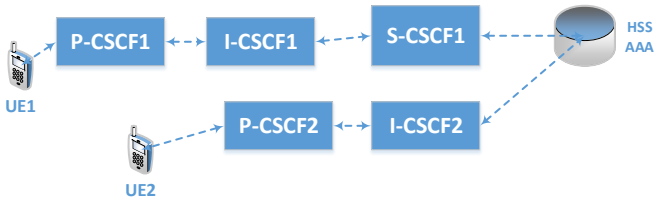


Figure 3.2: The RBD between UE1 and UE2

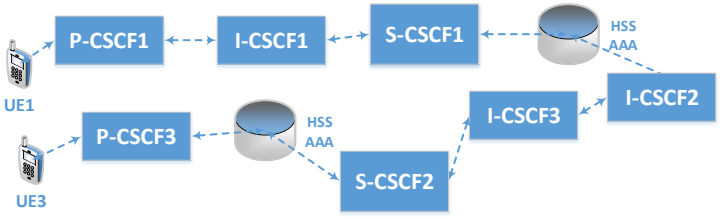


Figure 3.3: The RBD between UE1 and UE3

The reliability and signaling paths of two sample registration and multimedia session setups between UE1 and UE2, and between UE1 and UE3 are presented by figure 11 (a) and 11 (b) respectively. Under the operating period, assuming that all equipment and links (both wired or wireless) are reliable. Accordingly, all network elements or proxy servers need to be in the available state to successfully perform the registration or session services.

The reliability function at time $[0, t]$, between UEs can be represented by each reliability of the block diagram. Let $R_{ue}(t)$,$R_p(t)$,$R_i(t)$,$R_s(t)$,$R_A(t)$ represent the reliability of the user terminal, proxy-call state control function (P-CSCF), interrogating-call state control function (I-CSCF), serving-call state control function (S-CSCF) and HSS/AAA server respectively. Therefore the total reliability of the signaling or session setup path can be computed and given as $R_{(System,a)}(t)$ and $R_{(System,b)}(t)$ [1, 2].

$$R_{System,a}(t) = [R_u(t)]^2[R_p(t)]^2[R_i(t)]^2[R_s(t)][R_A(t)] \tag{3.1}$$

$$R_{System,b}(t) = [R_u(t)]^2[R_p(t)]^2[R_i(t)]^3[R_s(t)]^2[R_A(t)]^2 \tag{3.2}$$

The Poisson processes and exponential distribution can be assumed to represent the traffic between user and servers, and utilization characteristics of the servers respectively. Then the reliability of the signaling path of two scenarios

can be given by the failure parameters [3] as follow:

$$R_{System,a}(t) = e^{-(2\lambda_\mu+2\lambda_p+2\lambda_i+\lambda_s+\lambda_A)t} \tag{3.3}$$

$$R_{System,b}(t) = e^{-(2\lambda_\mu+2\lambda_p+3\lambda_i+2\lambda_s+2\lambda_A)t} \tag{3.4}$$

Where $\lambda_u, \lambda_p, \lambda_i, \lambda_s$, and λ_A are the failure constant parameters of the user end device (UE), P-CSCF, I-CSCF, S-CSCF and the HSS/AAA respectively. The mean time to failure, MTTF, which represent the average time of the system until the failure occurred can be given by equation 3.5.

$$MTTF = \int_0^\infty R_{System}(t)dt \tag{3.5}$$

Therefore, the mean time to failure (MTTF) of the two scenarios can be given by equation 3.6 and 3.7.

$$MTTF_a = \int_0^\infty R_{System,a}(t)dt = -(2\lambda_u + 2\lambda_p + 2\lambda_i + \lambda_s + \lambda_A) \tag{3.6}$$

$$MTTF_b = \int_0^\infty R_{System,b}(t)dt = -(2\lambda_u + 2\lambda_p + 3\lambda_i + 2\lambda_s + 2\lambda_A) \tag{3.7}$$

From the equation 3.6 and 3.7, we can clearly analyze that the reliability of scenario (a) will have a higher value than scenario (b). Considering the IMS architecture, It clearly acknowledges that the reliability of the complex system would be less reliable than a less complex scenario as the scenario (a). In other words, the system composes of many pieces of equipment or operating functions. The analytical equations showed explicit reliability understanding without providing any numerical results. The future state, the communication scenarios can be simulated for validation of the model analysis.

3.4 IMS Reliability via Markov Model

Considering the realistic IMS system where the components can be repaired, or some component may experience some trivial software error and can be recovered by rebooting or some recovery schemes. The system function can be represented by using the CTMC model. Instead of using two-state (up and down) or three-state Markov model as performed by [4], the CTMC model below is proposed and applied for reliability and performance evaluation as shown in figure 3.4.

The idea of the proposed model is influenced by [5] which is used to model a multiprocessor system with different failure types. In presence, the network

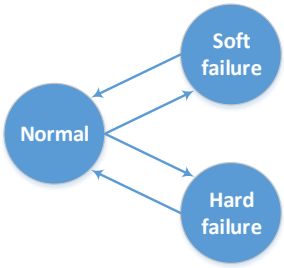


Figure 3.4: The proposed Markov model for IMS reliability and performance analysis

elements can have more than two failure modes or states. The normal working state, the soft failure which similar to failover or auto-recovery state, hard failure state that may end up to long time failure that may need hand-operated repaired to go back to the normal state. The system can be recovered back to normal state when recovery schemes or redundancy triggered after the failure. The quantification of availability or reliability of the proposed model can be mathematically given based on the reliability theory.

3.5 Availability Analysis

As mentioned above, realistic system condition involves repairable components or states. The availability concept can be used for analysis reliability of the system. Two main possible states of the IMS functional modules or components can be represented as in failure or normal state. The two-state Markov model can be used to represent this IMS system's condition as shown by figure 3.5.

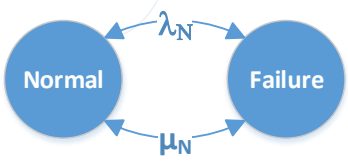


Figure 3.5: The two-state Markov model represents the IMS system behaviors

From the figure 3.5, the state of IMS functional module or server will be represented as in "Normal" or "Failure" state. The normal state outlines the state that all functional units or servers are working properly. However when the failure occurred, the server state will change from "Normal" to "Failure"

state with the probability or failure constant rate equal to λ_N where subscription N is used to represent each IMS functional units or possible failure constant rates and can be formulated as $\{\lambda_N | N \in (u, p, i, s, A)\}$. In contrast, the failure state represents the state with failure and can be recovered with repair rate equal to μ_N where subscription N is used to represent each IMS units or $\{\mu_N | N \in (u, p, i, s, A)\}$. Therefore, using this model, the server state will either be in Normal or Failure state. According to the definition of the availability which is the probability that physical path or service is in working condition after a failure. Therefore, from equation 2.1 can be written in term of the failure probability as

$$\begin{aligned} A(t) =& P_r(fully\, and\, functioning\, in\, [0,t]) \\ &+ P_r(one\, failure\, and\, one repair\, in\, [0,t]) \\ &+ P_r(two\, failure\, and\, two\, repaired\, in\, [0,t]) \\ &+ ... + P_r(N\, failure\, and\, N\, repaired\, in\, [0,t]). \end{aligned}$$

Therefore, the value of the availability is always greater than or equal to the reliability. Then considering the availability of each IMS component, the availability can be written as $A_N(t)$ where $N \in (u, p, i, s, A)$ and assuming that the failure and repair time is characterized by the exponential distribution function. Then $A_N(t)$ [6] is given by equation 3.8.

$$A_N(t) = \frac{\mu_N}{\lambda_N + \mu_N} + \frac{\lambda_N}{\lambda_N + \mu_N} e^{-(\lambda_N + \mu_N)t} \tag{3.8}$$

Also, the steady state availability, availability (A) or the availability when considering typical system operating time where $A = \lim_{t\to\infty} A_N(t)$ and is given by equation 3.9.

$$\begin{aligned} A &= \lim_{t\to\infty} A_N(t) \\ &= \frac{\mu_N}{\lambda_N + \mu_N} + \frac{\lambda_N}{\lambda_N + \mu_N} e^{-(\lambda_N + \mu_N)t} \\ A_N(t) &= \frac{\mu_N}{\lambda_N + \mu_N} \end{aligned} \tag{3.9}$$

The equation 3.9 is similar to the steady state availability previously derived and shown by equation 2.4 where $A = \frac{\mu}{\mu+\lambda} = \frac{MTTF}{MTTF+MTTR}$. The availability only depends on the MTTF and the mean time to recovery (MTTR). This shows that the availability value does not depend on the characteristic of the failure or repair time distribution when considering at steady state condition. These availability equations represent the system with a single Up (normal) state and Down (failure) state and can not be applied to the system with internal redundancy.

Based on end-to-end IMS setup scenario 3.2, the availability analysis can be done by applying availability equation 3.8 and 3.9. Similar to reliability

analysis 3.3 the availability of the two signaling or session setup paths is given by $A_{System,a}(t)$ and $A_{System,b}(t)$ for the IMS setup scenario per figure 3.2 and 3.3 respectively.

$$A_{System,a}(t) = [A_u(t)]^2[A_p(t)]^2[A_i(t)]^2[A_s(t)][A_A(t)] \tag{3.10}$$

$$A_{System,a}(t) = [A_u(t)]^2[A_p(t)]^2[A_i(t)]^3[A_s(t)]^2[A_A(t)]^2 \tag{3.11}$$

Where

$$A_N(t) = \frac{\mu_N}{\lambda_N + \mu_N} + \frac{\lambda_N}{\lambda_N + \mu_N} e^{-(\lambda_N + \mu_N)t}; N \in (u, p, i, s, A)$$

Then the steady state availability can be calculated when $A = \lim_{t\to\infty} A_N(t) = A_N$ and are given by equation 3.12 and 3.13.

$$A_{System,a} = [A_u]^2[A_p]^2[A_i]^2[A_s][A_A] \tag{3.12}$$

$$A_{System,b} = [A_u]^2[A_p]^2[A_i]^3[A_s]^2[A_A]^2 \tag{3.13}$$

Where $A_N(t) = \frac{\mu_N}{\lambda_N + \mu_N}; N \in (u, p, i, s, A)$. Considering the availability between IMS setup scenario (a) and (b) from 3.10 and 3.11, the system (a) has fewer multiplication terms than system (b), we can directly conclude that availability of system (a) would be less than availability of the system (b). However, in practical, for a good system, the failure rate should be less than the recovery rate or $\lambda < \mu$ or $\frac{\lambda}{\mu} < 1$. In this case, from 3.8, the availability value will be less than one. Therefore, the availability of system (a) will be greater than the availability of system (b). The equivalent results can be observed when applying the concepts with steady state availability. Again when comparing together between availability analysis with the reliability analysis of 3.3, it can be concluded and well agree that IMS setup scenario (a) will most likely to have higher availability or reliability than the IMS setup scenario (b). Moreover from [6], considering when the value of recovery rate, μ is close to zero. Then from equation 3.3, 3.4, 3.10, and 3.10, each availability and reliability terms can be given by.

$$A(t) = R(t) = e^{-\lambda t}; \mu \to 0 \tag{3.14}$$

Equation 3.14 proves that when there is no maintenance of the system, $\mu \to 0$, the availability and reliability are identical. We can consider other possibilities of recovery rates, if the value of $\mu \to 0$, The availability value will be close to one. This yields higher availability system due to higher recovery

rate. On the other hand, considering when the system having a very high failure rate, $\lambda \to 0$, the availability and reliability value will moving towards zero. In this case, it represents the system with poor performance or unreliable. From an analysis above, by varying the value of λ and μ, the results agree well with the definition of availability and reliability of the system and also can be used to verify the above equations.

3.6 Two-state CTMC Analysis

Assume a random variable $X(t)$, which will be used to represent states of the system. When the value of $X(t)$ changed, the system is called having a state transition. Let $P_r(X(t_n) = j)$ represents the probability that the system is in state j at an instant time t_n. Then the Markov process refers to stochastic process at times: $t_1 < t_2 < ... < t_n$, with the conditional probability of being in state j from the previous state i is given by equation 3.15.

$$\begin{aligned} P_r\{X(t_n) = j | X(t_{n-1}) = i_{n-1}, X(t_{n-2}) = i_{n-2}, ..., X(t_0) = i_0\} \\ = P_r\{X(t_n) = j | X(t_{n-1}) = i_{n-1}\} \end{aligned} \tag{3.15}$$

Equation 3.15 points that Markov chain, the state transition probability may only depend on the state shortly before it except any other states. The conditional probability is said to be homogeneous or invariant with respect to time t_n, if for any t and t_n.

$$P_r\{X(t) = j | X(t_n) = i_n\} = P_r\{X(t - t_n) = j | X(0) = i_n\}$$

This homogeneous property shows that probability of the system in state j at the time $t > t_n$ does not depend on how long the system has been in this state (sojourn time) or the process is said to be memory less. Therefore, the probability of being in that state does not only depend on the previous states but also the amount of time spending in the current state. Let $\pi_j(t)$ represent the state probabilities and is equal to $P_r(X(t) = j)$ at the time, t. The transition probability that the system is in state j at the time t after being in previous state i at the time u is defined as

$$P_{ij}(t - u) = P_r[X(t) = j | X(u) = i].$$

Normally the value of u is set to be zero. There might be zero or more than one state transitions before moving from the previous state i at the time u to state j. This can be given in the form of Chapman-Kolmogorov equation as in equation 3.16.

$$\pi(t) = \pi(u).\mathbb{P}(t - u);\ where\ \mathbb{P}\ is\ matrix\ form\ of\ P \tag{3.16}$$

For the CTMC, let $u = (t - \Delta t))$ and subtract $\pi(t - \Delta t)$ from both sides of 3.16, then divide by Δt and take limit $\Delta t \to 0$. Then we get equation 3.17.

$$\frac{d\pi(t)}{dt} = \pi(t) \lim_{\Delta t \to 0} \frac{\mathbb{P}(\Delta t) - \mathbb{I}}{\Delta} : where\ \mathbb{I}\ is\ identity\ matrix \tag{3.17}$$

Let $\delta_{ij} = 1$ if $i = j$, zero otherwise, and let define matrix $\mathbb{Q}$ as

$$q_{ij} = \lim_{\Delta t \to 0} \frac{P_{ij(\Delta t) - \delta_{ij}}}{\Delta t}$$

Where q_{ij} is the transition rate going from state i to j. Therefore, equation 3.17 can be written as equation 3.18

$$\frac{d\pi(t)}{dt} = \mathbb{Q}.\pi(t) \tag{3.18}$$

The equation 3.18 is known as Kolmogorov differential equation and is used to represent the state probabilities of the system. If there is a solution when $\lim_{\Delta t \to 0} \pi(t)$, the derivation of 3.18 will be zero. The we can have equation 3.19 and 3.20.

$$\mathbb{Q}.\pi = 0 \tag{3.19}$$

$$\mathbb{E}.\pi = 1 \tag{3.20}$$

Where $\mathbb{E} = [1, 1, ..., 1]^T$, superscript T denote the transpose. From the figure 3.5, based on our assumption of two possible states of the system components (normal and failure), therefore, the values zero and one can be used to represent normal and failure state respectively. The two state Markov chain can be represented by figure 3.6. Figure 3.6 represents homogeneously and continuous

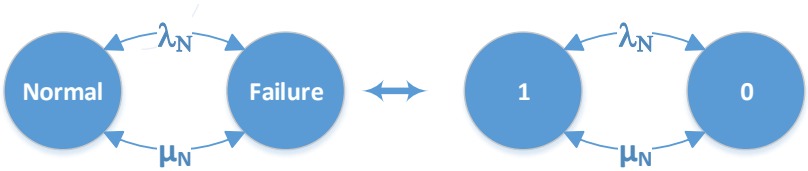


Figure 3.6: The two-state CTMC model representation

time Markov chain model for the IMS components that have two states: normal and failure (one and zero). Then let the failure rate $q_{10} = \lambda$ and the failure rate $q_{01} = \mu$. Therefore, the transition matrix $\mathbb{Q}$ is given by

$$\mathbb{Q} = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \begin{bmatrix} -\mu & \mu \\ \lambda & -\lambda \end{bmatrix}$$

From equation 3.18, $\frac{d\pi(t)}{dt} = \mathbb{Q}.\pi(t)$, the system state probability equation is

$$\pi_0^{'}(t) = -\mu\pi_0(t) + \lambda\pi_1(t)\pi_1^{'}(t) = \mu\pi_0(t) + \lambda\pi_1(t) \tag{3.21}$$

Base on summation of probabilities is equal to one; therefore,$\pi_0(t) + \pi_1(t) = 1$, equation 3.21 is given by equation 3.22.

$$\begin{aligned} \pi_1^{'}(t) &= \mu(1 - \pi_1(t)) + \lambda\pi_1(t) \\ \pi_1^{'}(t) &+ (\mu + \lambda)\pi_1(t) = \mu \end{aligned} \tag{3.22}$$

Then solving the equation by multiplying both sides of equation 3.22 by $e^{\int(\mu+\lambda)dt} = e^{(\mu+\lambda)t}$ then we get equation 3.23.

$$\begin{aligned} e^{(\mu+\lambda)t}\pi_1^{'}(t) + (\mu + \lambda)e^{(\mu+\lambda)t}\pi_1(t) &= \mu e^{(\mu+\lambda)t} \\ \left(e^{(\mu+\lambda)t}\pi_1(t)\right)^{'} &= \mu e^{(\mu+\lambda)t} \\ \left(e^{(\mu+\lambda)t}\pi_1(t)\right) &= \frac{\mu e^{(\mu+\lambda)t}}{\mu + \lambda} + c \end{aligned}$$

$$\pi_1(t) = \frac{\mu}{\mu + \lambda} + ce^{-(\mu+\lambda)t}; where\, c = constant \tag{3.23}$$

If our system condition is assumed to be in normal state (state 1) at the beginning (time =0) or $\pi_1(0) = 1$, then the constant, c can be calculated by using equation 3.23 and is given by $\pi_1(0) = 1 = \frac{\mu}{\mu+\lambda} + c$; $c = \frac{\lambda}{\mu+\lambda}$
Then equation 3.23 can be given as equation 3.24.

$$\pi_1(t) = \frac{\mu}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda} + e^{-(\mu+\lambda)t} = A(t)c = constant \tag{3.24}$$

The equation 3.24 shows the transient probability function and the instantaneous availability, $A(t)$, of the system component. Therefore, the steady state availability of the system is given by taking the limit of t to infinity.

$$A = \lim_{t\to\infty} \pi_1(t) = \frac{\mu}{\mu + \lambda} \tag{3.25}$$

From equations: 3.24 and 3.25, the transient and steady state availability given by using the CTMC model are similar to transient and steady state availability previously derived and shown by equation 3.9 due to the failure and recovery are assumed to be exponential distributions which are the same for both cases. This can validate availability analysis by using the CTMC. Moreover the expected uptime, $E(U(t))$, in the interval $(0, t)$ can be calculated by integration of the normal state probability,$\pi_1(t)$.

$$E(U(t)) = \int_0^t \pi_1(x)dx = \frac{\mu}{\mu + \lambda} + \frac{\lambda}{(\mu + \lambda)^2}(1 - e^{-(\mu+\lambda)t}) \tag{3.26}$$

By using the same method as $\pi_1(t)$ for finding$\pi_0(t)$ from equation 3.21.

$$\pi_0^{'}(t) + (\mu + \lambda)\pi_0(t) = \lambda \tag{3.27}$$

$$\pi_0(t) = \frac{\lambda}{\mu + \lambda} + ce^{-(\mu+\lambda)t};\ where\ c = constant \tag{3.28}$$

Finding the constant, c, by considering $\pi_0(t)$ at time zero, $\pi_0 = 0$.

$$\pi_0(0) = \frac{\lambda}{\mu + \lambda} + c = 0$$
$$c = -\frac{\lambda}{\mu + \lambda}$$

then

$$\pi_0(t) = \frac{\lambda}{\mu + \lambda} - \frac{\lambda}{\mu + \lambda}e^{-(\mu+\lambda)t}. \tag{3.29}$$

The equation 3.29, represents the instantaneous unavailability equation, UA, of the system components. Considering at steady state condition when the function is finite or have the value when the value of time t close to infinity.

$$UA\ at\ statedy\ state\ condition = \lim_{t\to\infty} \pi_0(t)$$
$$= \frac{\lambda}{\mu + \lambda}$$

The unavailability value is simply calculated by subtraction of the availability by one.

$$1 - A = 1 - \frac{\mu}{\mu + \lambda}$$
$$= \frac{\lambda}{\mu + \lambda}$$
$$= UA$$

Moreover, the expected downtime, $E(D(t))$, in the interval $(0, t)$ can be calculated and given by equation 3.30

$$E\left(D\left(t\right)\right) = \int_0^t \pi_0\left(x\right) dx = \frac{\lambda t}{\mu + \lambda} - \frac{\lambda}{\left(\mu + \lambda\right)^2}(1 - e^{-(\mu+\lambda)t}) \tag{3.30}$$

3.7 Simulation of Transient Availability, instantaneous availability (A(t))

The figures below show the simulation results of the transient or instantaneous availability, $A(t)$ per equation 3.24 of the system at different value of failure

rates (λ) and recovery rate (μ). The values of λ and μ are chosen in the way that $\frac{\lambda}{\mu} < 1$ due to the realistic assumption of a good environment system which explained in item 3.5.

- Transient availability, $A(t)$ versus time , (t) when $\mu = 0.99$ and at different values set of $\lambda = \{0.01,\ 0.09,\ 0.2,\ 0.5,\ 0.7\}$

  The simulation results show that when increasing the failure rate at a fixed value of recovery rate, the $A(t)$ will decrease both for both transient and steady states. As we can see from table 3.1 and the figure 3.8, the availability value is decreased when the value of λ is increased. The trend of decreasing is proportional to the polynomial order three. That means in transient state the decreasing of availability value will be more than twice when increasing the failure rate value. The simulation results show that at a high value of λ, the system availability characteristic take longer time to reach the steady state liked region. Moreover, the availability value will not much vary when the value of λ is closed to μ.

Table 3.1: $A(t)$ at different values of λ and a fixed value of μ=0.99

$A(t), \mu = 0.99$	λ	% of increasing λ from B	Maximum value, $A(t)$ Max	% of decreasing Max $A(t)$ from B	Minimum value, $A(t)$ Min or Steady State Availability	% of decreasing Min $A(t)$ from B
A	0.01	-	0.9937	-	0.99	-
B	0.2	1	0.8831	-	0.8319	-
C	0.5	2.5	0.7401	16.1929566	0.6644	20.134632
D	0.8	4	0.6277	28.920847	0.5531	33.513643
E	0.9	4.5	0.5957	32.5444457	0.5238	37.035701
F	0.98	4.9	0.5972	32.3745895	0.5263	37.735185

- Transient availability, $A(t)$ versus time , (t) when $\lambda = 0.01$ and at different values set of $\mu = \{0.4,\ 0.6,\ 0.8,\ 0.9,\ 0.99\}$
  As we can see from the figure 3.10 and the table 3.2, at the given value of $\lambda = 0.01 and 0.2$ and at a different value of μ. As increasing the value of μ, the trend of the availability value is increased. Also, the increasing trend is proportional to the polynomial order three. This shows that increasing the recovery rate will results in increasing availability value more than twice of the previous value. Moreover, the simulation results show that at a higher value of μ, the availability of the system will reach steady state faster than the system that has a lower value of μ.

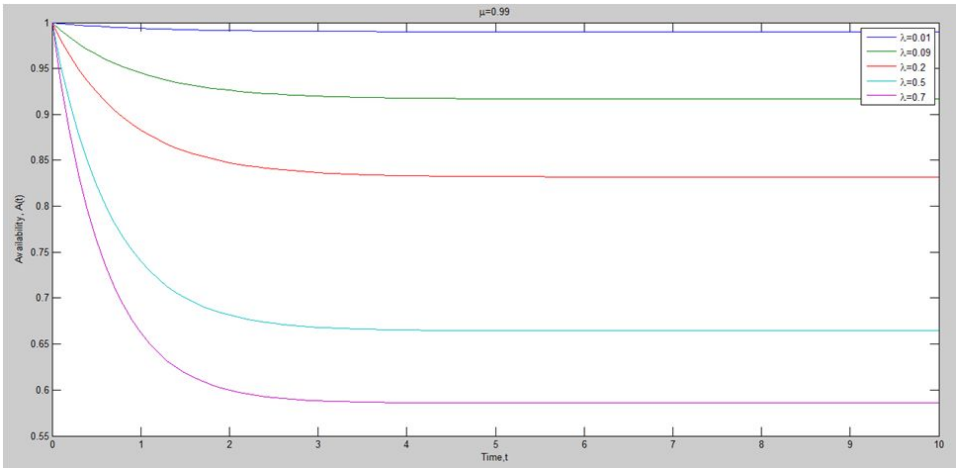


Figure 3.7: The $A(t)$ different values of λ at a fixed value of $\mu = 0.99$

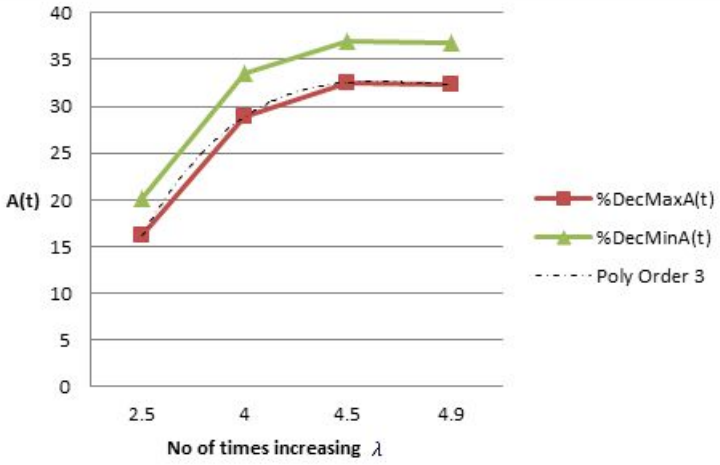


Figure 3.8: The trend of decreasing $A(t)$ when λ is increased at a fixed value of $\mu = 0.99$

- Transient availability, $A\,(t)$ versus time , (t) when $\lambda = 0.9$ and at different values of $\mu = \{0.99,\ 0.999,\ 0.9999\}$
  From figure 3.11, the results show that increasing of the recovery rate will not highly affect the maximum availability value due to a pretty close value between μ and λ. In conclusion from the above transient availability results at the different value of μ and λ, If the system or components of the system have a very high value of recovery rate or a

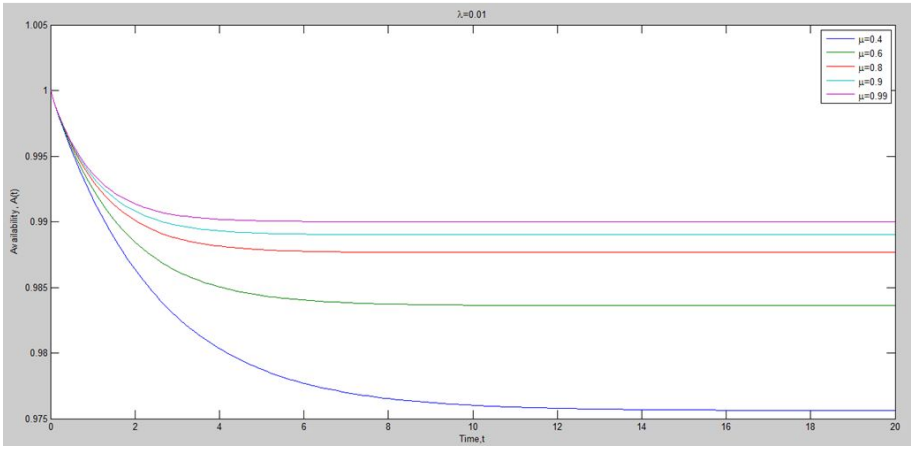


Figure 3.9: The $A(t)$ different values of μ at a fixed value of $\lambda = 0.01$

Table 3.2: $A(t)$ at different values of μ and a fixed value of $\lambda = 0.2$

$A(t), \lambda = 0.2$	Maximum value, $A(t)$ Max	Minimum value, $A(t)$ Min	μ	% of increasing μ	% of increasing Max A(t) from a	% of increasing Min $A(t)$ from a
a	0.8496	0.6667	0.4	1	0	0
b	0.8623	0.75	0.6	1.5	1.494821092	12.494375
c	0.8736	0.8	0.8	2	2.824858757	19.994
d	0.8787	0.818	0.9	2.25	3.425141243	22.693865
e	0.8831	0.8319	0.99	2.475	2.412153543	10.92

very low value of the failure rate, the availability value of the system will be close to one. That refers to an ideal system with a high value of availability or reliability. Nevertheless, if the system components have a very low value of recovery rate or have a very high value of failure rate, the availability value will be exponentially decreased with operating period. Also, the simulation results reveal that when both recovery rate and failure rate values are similar or closed to each other, the availability of the system will not be much varied and the availability value at steady state condition will be $\approx 50\%$, this due to the system characterized by fifty percent of failure and recovery.

- The simulation results of end-to-end availability of the system (a) and (b) based on the proposed scenario analysis per section 3.2 Figure 3.12 shows the simulation results of the end-to-end availability versus time between two session setup scenarios (a and b) which are represented by $A_(System, a)(t)$ and $A_(System, b)(t)$ according to the equation 3.10 and

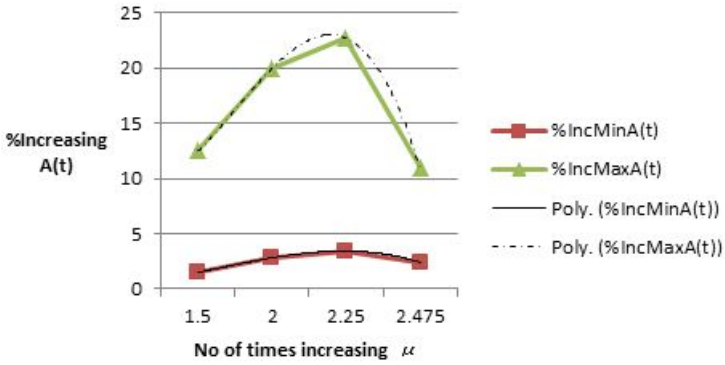


Figure 3.10: The trend of increasing $A(t)$ when the values of μ is increased at fixed value of $\lambda = 0.2$

Table 3.3: The $A(t)$ at different values of μ and a fixed value of $\lambda = 0.9$

Availability $A(t), \lambda = 0.9$	Maximum value, $A(t)$ Max	Minimum value, $A(t)$ Min	μ
a	0.5957	0.5238	0.99
b	0.5970	0.5261	0.999
c	0.5971	0.5263	0.9999

3.11 respectively. The simulation results were plotted at two values of $\lambda = \{0.01, 0.02\}$ and $\mu = 0.9$ for both system (a) and (b). From the results, availability of system (a) is higher than system b, about two percentage difference. Moreover, when increasing the failure rate of the system by approximately two times, the results as seen by figure 3.12 (a) and (c) and the table 3.4, the maximum availability value is decreased approximately five percent comparing to the previous failure rate. Besides, the steady state availability in decreased for more than seven percent. This due to the end-to-end availability of scenario (a) concerns session setup paths less than the scenario (b). The results agree well with the analysis in section 3.5 that end-to-end availability value of session setup scenario of within one communication domain (system a) is higher than availability value of session setup between two communication domains (system b). Moreover, the results validate end-to-end availability analysis of the IMS system which are based on equations 3.10 to 3.13.

Recently the availability analysis model of a computer system and network with internal and external redundancy has been performed by using the Markov chain. The geo-redundancy model with Markov chain was simulated and shown higher availability value than the system without redundancy both client and

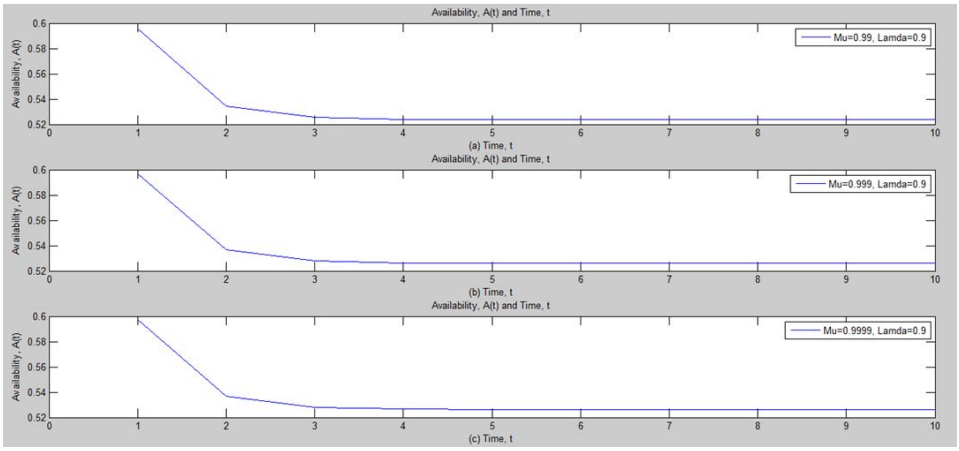


Figure 3.11: The relationship of $A(t)$ at the different values of μ at the fixed value of $\lambda = 0.9$

Table 3.4: The $A(t)$ of system (a) and (b) at different values of $\lambda = 0.1, 0.2$ and a fixed value of $\mu = 0.99$

Availability, $A(t)$	λ	μ	Maximum value, $A(t)$ Max	Minimum value, $A(t)$ Min
a	0.01	0.99	0.9505	0.9227
b	0.01	0.99	0.9326	0.8953
c	0.02	0.99	0.9036	0.8521
d	0.02	0.99	0.8699	0.8025

system part [33]. Moreover, based on the work of Mkwawa and Kouvatsos [18], the performance of IMS system was evaluated by considering the delay of the S-CSCF due to its function that handle many SIP signaling per each session setups more than other functional IMS units. Also, the reliability mechanism of the IMS network had been proposed and studied by having one redundancy S-CSCF units in the system and examined with the open source OpenIMS. The results revealed that the IMS system seamlessly affects when one unit of the S-CSCF is failed [34]. The S-CSCF is one of the most important functional units of IMS system due to it manage and process many SIP signalings during registration and session setup processes. It also provides other important information such as maintenance and service charge. Therefore, this work focuses on performing availability or reliability analysis of the core IMS system by using the two-state CTMC and the proposed five-state CTMC when having a redundancy of the S-CSCF.

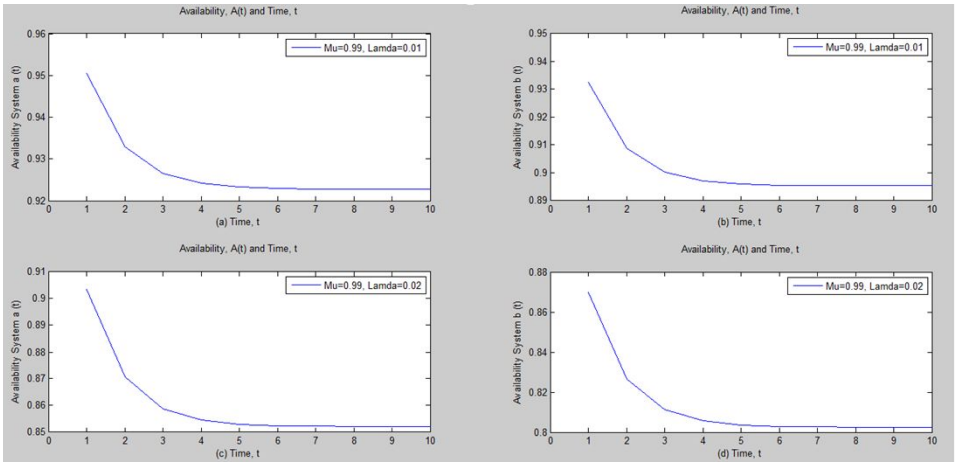


Figure 3.12: The end-to-end availability of the system scenario a and b at different values of λ and μ

3.7.1 Availability analysis with one redundancy of the S-CSCF unit

The figure 3.13 shows the CTMC model of a redundancy unit with failure, degraded, and repair states. From figure 3.13, the model shows the system that

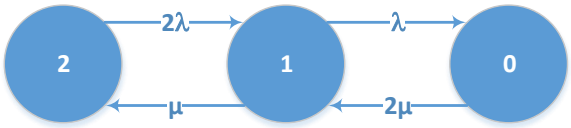


Figure 3.13: The CTMC model with one redundancy unit

composes with one redundancy unit. Assumed that the failure rate does not depend on the status of the unit and the failure and recovery rate is exponentially distributed. There are total three possible states spaces, $S = \{0, 1, 2\}$. State 2 represents two units with both in normal working condition. State 1 represents one unit with the normal working condition and another failed unit. State 0 represent two failure unit or system failed. This occurs when one unit failed while another unit is in a failure state or two units failed at the same time. The variable λ and μ represent the failure and recovery rate of the system respectively. From the figure 3.13 and equation 3.18, $\frac{d\pi(t)}{dt} = \mathbb{Q}.\pi(t)$, it can be

rewritten in the vector differential equation format as $\pi^{'}(t) = \mathbb{Q}\pi(t)$. where

$$\begin{bmatrix} \pi_2^{'}(t) \\ \pi_1^{'}(t) \\ \pi_0^{'}(t) \end{bmatrix} = \begin{bmatrix} -2\lambda & \mu & 0 \\ 2\lambda & -(\mu+\lambda) & 2\mu \\ 0 & \lambda & -2\mu \end{bmatrix} . \begin{bmatrix} \pi_2(t) \\ \pi_1(t) \\ \pi_0(t) \end{bmatrix}$$

$$\pi_2^{'}(t) = -2\lambda\pi_2(t) + \mu\pi_1(t) \tag{3.31}$$

$$\pi_1^{'}(t) = 2\lambda\pi_2(t) - (\lambda+\mu)\pi_1(t) + 2\mu\pi_0(t) \tag{3.32}$$

$$\pi_0^{'}(t) = \lambda\pi_1(t) - 2\mu\pi_0(t) \tag{3.33}$$

The system variables can be rearranged as $\frac{1}{\pi(t)}d\pi(t) = \mathbb{Q}dt$. Then integrate both sides of the equation and let $\pi(0)$ and $\pi(t)$ represent an initial and normal condition of variable t respectively.

$$\begin{aligned} \int_{\pi(0)}^{\pi(t)} \frac{1}{d\pi(t)} &= \int_0^t \mathbb{Q}dt \\ ln\,\pi(t)|_{\pi(0)}^{\pi(t)} &= \mathbb{Q}t \\ ln\,\pi(t) - ln\,\pi(0) &= \mathbb{Q}t \\ ln\,\frac{\pi(t)}{\pi(0)} &= \mathbb{Q}t \end{aligned}$$

$$\pi(t) = \pi(0)e^{\mathbb{Q}t} \tag{3.34}$$

$$e^{\mathbb{Q}t} = I + \mathbb{Q}t + \frac{1}{2!}(\mathbb{Q}t)^2 + ... \tag{3.35}$$

The general solution is given by finding $e^{\mathbb{Q}t}$. Also by taking Laplace transform of equation 3.18, the matrix can be written as

$$s\boldsymbol{\pi}(s) - \boldsymbol{\pi}(0) = \mathbb{Q}.\boldsymbol{\pi}(s)$$

$$\boldsymbol{\pi}(s) = \boldsymbol{\pi}(0)(s\mathbf{I} - \mathbb{Q})^{-1} \tag{3.36}$$

Where $\mathbf{I}$ is the identity matrix with identical size of $\boldsymbol{\pi}$. The $\boldsymbol{\pi}(0)$ is the initial condition value. With state probability conditions: $\pi_2(t) + \pi_1(t) + \pi_0(t) =$

1 and assuming both units are up for initial condition: $\pi_2(0) = 1, \pi_1(0) = 0, \pi_0(0) = 0$. Therefore, the general solutions can be calculated by finding all the eigenvalues of $\mathbb{Q}$ and the inversion of Laplace transform and the solutions are given following equations.

$$\pi_2(t) = \frac{(\lambda^2 + 2\lambda\mu)e^{-(\lambda+\mu)t}}{(\lambda+\mu)^2} + \frac{\mu^2}{(\lambda+\mu)^2} \tag{3.37}$$

$$\pi_1(t) = \frac{\left(2\lambda^2 - 2\lambda\mu\right)e^{-(\lambda+\mu)t}}{(\lambda+\mu)^2} - \frac{2\lambda^2 e^{-2(\lambda+\mu)t}}{(\lambda+\mu)^2} + \frac{2\lambda\mu}{(\lambda+\mu)^2} \tag{3.38}$$

$$\pi_0(t) = \frac{\lambda^2}{(\lambda+\mu)^2} - \frac{2\lambda^2 e^{-(\lambda+\mu)t}}{(\lambda+\mu)^2} + \frac{\lambda^2 e^{-2(\lambda+\mu)t}}{(\lambda+\mu)^2} \tag{3.39}$$

Then, the transient availability is a summation of all availability states and is given by

$$A_{Redundant}(t) = \pi_2(t) + \pi_1(t)$$

$$A_{Redundant}(t) = \frac{(\lambda^2 + 2\lambda\mu)e^{-(\lambda+\mu)t}}{(\lambda+\mu)^2} + \frac{\mu^2}{(\lambda+\mu)^2} + \frac{\left(2\lambda^2 - 2\lambda\mu\right)e^{-(\lambda+\mu)t}}{(\lambda+\mu)^2} - \frac{2\lambda^2 e^{-2(\lambda+\mu)t}}{(\lambda+\mu)^2} + \frac{2\lambda\mu}{(\lambda+\mu)^2}$$

$$A_{Redundant}(t) = \frac{3\lambda^2 e^{-(\lambda+\mu)t}}{(\lambda+\mu)^2} - \frac{2\lambda^2 e^{-2(\lambda+\mu)t}}{(\lambda+\mu)^2} + \frac{\mu^2 + 2\lambda\mu}{(\lambda+\mu)^2} \tag{3.40}$$

From the proposed architecture model, the RBD of figure 3.2 and 3.3 can be rewritten when having redundant unit of S-CSCF as figure 3.14 and 3.15 below.

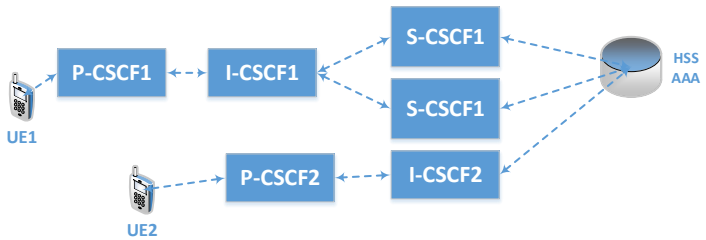


Figure 3.14: The RBD between UE1 and UE2

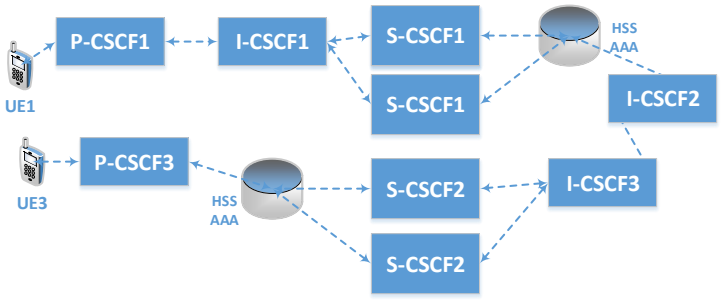


Figure 3.15: The RBD between UE1 and UE3

The combination of stochastic models is chosen to calculate the availability of the IMS system. Each RBD represents subsystem or individual components of the IMS system. The RBD is used to estimate the end-to-end availability of the whole system. Also, the CTMC process model which represents more detailed failures and recovery characteristic of the components will be used to evaluate the availability of each subsystem. The RBD can avoid constructing large state space or complexity of the Markov process for the whole system. However, detailed failures and recovery can be characterized as a second level of the combination techniques with the Markov process. Therefore, the system model can be created and evaluated with more detailed failure and recovery characteristics with less complexity. The other functional units except S-CSCF assume to have no redundancy. From the RBD theory and independence assumption of the system components, the total availability function of series and parallel connection is given by [6].

$$A(t) = \left\{ \begin{array}{c} \prod_{j=1}^{N} A_i(t) : for\ a\ series\ structure \\ 1 - \prod_{j=1}^{N} (1 - A_i(t)) \ : for\ a\ parallel\ structure \end{array} \right\} \tag{3.41}$$

Therefore, end-to-end availability of the proposed model architecture per figure 3.14 and 3.15 are

$$\begin{aligned} A_{Total\_scenario\_a}(t) = [A_u(t)]^2 [A_p(t)]^2 [A_i(t)]^2 \\ [A_A(t)] \left[1 - ((1 - A_s(t))^2\right] \end{aligned} \tag{3.42}$$

$$\begin{aligned} A_{Total\_scenario\_b}(t) = [A_u(t)]^2 [A_p(t)]^2 [A_i(t)]^3 \\ [A_A(t)]^2 \left[1 - ((1 - A_s(t))^2\right]^2 \end{aligned} \tag{3.43}$$

The steady state probability vector of the system can be obtained by considering transient state probability at the time (t) is closed to infinity. If the steady state exists, equation 3.18 can be rewritten as, $\frac{d\pi(t)}{dt} = 0 = \mathbb{Q}.\boldsymbol{\pi(t)}$. This condition also referred to as Ergodic Markov chain where the average behavior of the system is identical over time. Therefore, the steady state probability can be analyzed from the equation below.

$$\mathbb{Q}.\boldsymbol{\pi(t)} = 0 \tag{3.44}$$

Based on the model architecture per figure 3.14 and 3.14, the steady state availability of the components with no redundancy and redundancy can be given by considering $\lim_{t\to\infty} A(t)$ then

$$A_{Total\_scenario\_a} = [A_u]^2[A_p]^2[A_i]^2\,[A_A]\left[1-(1-A_s)^2\right] \tag{3.45}$$

$$A_{Total\_scenario\_b} = [A_u]^2[A_p]^2[A_i]^3[A_A]^2\left[1-(1-A_s)^2\right]^2 \tag{3.46}$$

Where the steady state availability of the components with no redundancy and redundancy can be given by considering $\lim_{t\to\infty} A(t)$, then equation 3.24 and equation 3.40 are equal to

$$\begin{aligned}\lim_{t\to\infty} A\,(t) &= \lim_{t\to\infty}\left(\frac{\mu}{\mu+\lambda}+\frac{\lambda}{\mu+\lambda}e^{-(\mu+\lambda)t}\right)\\ &= \frac{\mu}{\mu+\lambda}\end{aligned}$$

$$\begin{aligned}\lim_{t\to\infty} A_{Redundant}\,(t) &= \lim_{t\to\infty}\left(\frac{3\lambda^2 e^{-(\lambda+\mu)t}}{(\lambda+\mu)^2}-\frac{2\lambda^2 e^{-2(\lambda+\mu)t}}{(\lambda+\mu)^2}+\frac{\mu^2+2\lambda\mu}{(\lambda+\mu)^2}\right)\\ &= \frac{\mu^2+2\lambda\mu}{(\lambda+\mu)^2}\end{aligned} \tag{3.47}$$

Assuming the following values of recovery and failure rates: $\mu = 0.99$, and $\lambda = 0.01$, then the simulation results of end-to-end availability of the model architecture (a) and (b) which is based on the proposed scenario analysis per figure 3.14 and 3.15 can be given by 3.16.

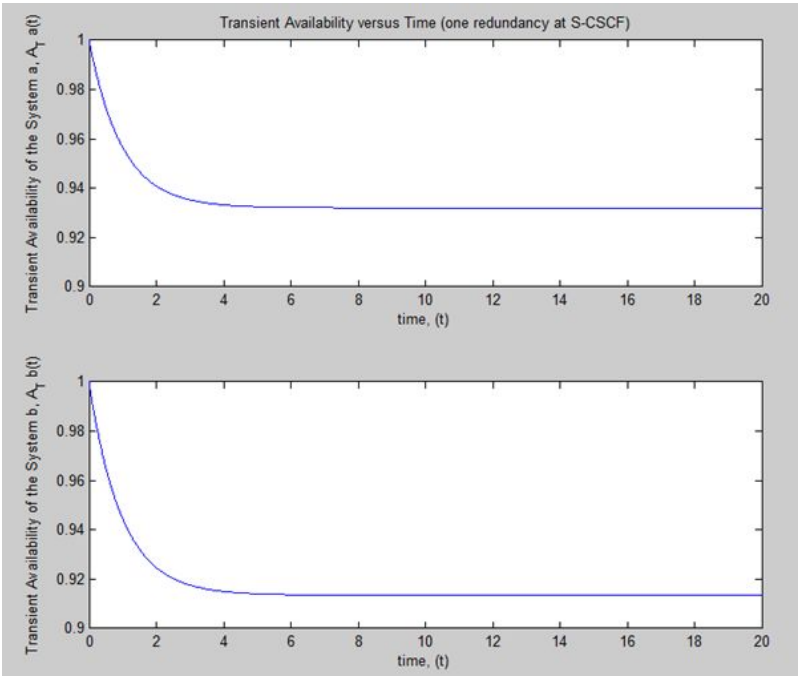


Figure 3.16: End-to-end availability of system scenario a and b at different with one redundancy at S-CSCF

From the figure 3.16 the end-to-end availability (with one redundancy of the S-CSCF) simulation results compare with the results shown by 3.13 which is the end-to-end availability without redundancy unit, the results clearly show that the availability of system with just one redundancy of the S-CSCF is higher than the system without redundancy for both transient and steady state characteristics. The results also show that for inter-domain communication or long connection of signaling path such as the IMS model architecture (b) or b-R will improve end-to-end availability value more than twice at steady state condition (see also the table 3.5). These interesting results imply that to increase availability or reliability of long signaling connection paths or large system architecture, redundancy is needed to be achieved.

3.7.2 Availability analysis with one redundancy of the S-CSCF unit while considering coverage factor

Figure 3.17 shows the system that compose with one redundancy unit which is similar to the model of figure 21. This model includes the effect of coverage factor, c, into the model. The coverage factor represents the percentage of failure or the probability that the system component may fail and can be

Table 3.5: The end-to-end availability $A(t)$ of system (a-R) and (b-R) with a redundant unit of the S-CSCF and the end-to-end availability of system a and b, without redundancy at $\lambda = 0.01$ and $\mu = 0.99$

Availability, $A(t)$	λ	μ	Maximum value, $A(t)$ Max	Minimum value $A(t)$ Min	Min % of increasing when comparing with the system with non redundancy (steady state)
a-R	0.01	0.99	1	0.9321	-
b-R	0.01	0.99	1	0.9135	-
a	0.01	0.99	1	0.9227	1.007912
b	0.01	0.99	1	0.8953	2.021669

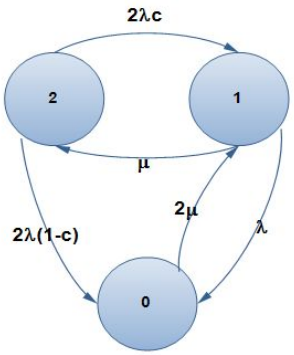


Figure 3.17: The continuous Markov model with one redundancy unit and coverage factor

automatically recovered. Assumed that the failure rate does not depend on the status of the unit and the failure and recovery rate is exponentially distributed. There are three state spaces, $S = \{0, 1, 2\}$. State 2 represents two units with both in normal working condition. State 1 represents one unit with the normal working condition while another one is failed and in the recovery process. State 0 represents two failure unit or system failed. From figure 3.17 and equation 3.18, $\frac{d\pi(t)}{dt} = \mathbb{Q}.\boldsymbol{\pi}(t)$, the system state probability equation can be given by

$$\begin{bmatrix} \pi_2^{'}(t) \\ \pi_1^{'}(t) \\ \pi_0^{'}(t) \end{bmatrix} = \begin{bmatrix} -2\lambda & \mu & 0 \\ 2\lambda c & -(\mu+\lambda) & 2\mu \\ 2\lambda(1-c) & \lambda & -2\mu \end{bmatrix} . \begin{bmatrix} \pi_2(t) \\ \pi_1(t) \\ \pi_0(t) \end{bmatrix}$$

$$\pi_2^{'}(t) = -2\lambda\pi_2(t) + \mu\pi_1(t) \tag{3.48}$$

$$\pi_1^{'}(t) = 2\lambda c(t) - (\lambda+\mu)\pi_1(t) + 2\mu\pi_0(t) \tag{3.49}$$

$$\pi_0^{'}(t) = 2\lambda(1-c)\pi_2(t) + \lambda\pi_1(t) - 2\mu\pi_0(t) \tag{3.50}$$

With state probability conditions: $\pi_2(t)+\pi_1(t)+\pi_0(t) = 1$ and assuming both units are up for initial condition: $\pi_2(0), = \pi_1(0) = 0, \pi_0(0) = 0$. The transient and steady state probabilities can be determined. Due to long expression of transient solutions, only the steady state probability will be presented as follow.

$$\pi_{2\_st}(t) = \frac{\mu^2}{(3\lambda\mu + \lambda^2 + \mu^2 - C\lambda\mu)} \tag{3.51}$$

$$\pi_{1\_st}(t) = \frac{2\lambda\mu}{(3\lambda\mu + \lambda^2 + \mu^2 - C\lambda\mu)} \tag{3.52}$$

$$\pi_{0\_st}(t) = \frac{\lambda(\lambda+\mu-C\mu)}{(3\lambda\mu + \lambda^2 + \mu^2 - C\lambda\mu)} \tag{3.53}$$

The transient and steady state availability is summation of all availability states which is state 2 and 1 and is given by

$$A_{Redundant\_C}(t) = \pi_2(t) + \pi_1(t). \tag{3.54}$$

Also the steady state availability of the components with no redundancy and redundancy will be given by considering $lim_{t\to\infty}A(t)$ of equation 3.24 and 3.54 and is equal to

$$\begin{aligned} lim_{t\to\infty}A(t) &= lim_{t\to\infty}(\frac{\mu}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda}e^{-(\mu+\lambda)t}) \\ &= \frac{\mu}{\mu+\lambda} \end{aligned}$$

Therefore,

$$\begin{aligned} \lim_{t\to\infty} A_{Redundant}(t) &= \frac{\mu^2}{(3\lambda\mu + \lambda^2 + \mu^2 - C\lambda\mu)} + \frac{2\lambda\mu}{(3\lambda\mu + \lambda^2 + \mu^2 - C\lambda\mu)} \\ &= \frac{\mu^2 + 2\lambda\mu}{(3\lambda\mu + \lambda^2 + \mu^2 - C\lambda\mu)} \end{aligned}$$

(3.55)

Assuming that the recovery and failure rates are $\mu = 0.99, \lambda = 0.01$. Most of the coverage factor value is close to one. Based on previous research work [7, 4, 8], assuming that $c = 0.95$ then the simulation results of end-to-end availability which is based on the proposed scenario analysis per figure 3.14 and 3.15, and is given by table 3.6 and figure 3.18. The simulation results

Table 3.6: End-to-end availability $A(t)$ of system $(a - R - C)$ and $(b - R - C)$ with a redundant unit of S-CSCF and considering coverage factor in the model. And also end to end availability of system $a - R$, $b - R$, a and b when $\lambda = 0.01$, $\mu = 0.99$ and $c = 0.95$

Availability, $A(t)$	λ	μ	Coverage factor, c	Maximum value,$A(t)$ Max	Steady state),$A(t)$
a-R-C	0.01	0.99	0.95	1	0.9321
b-R-C	0.01	0.99	0.95	1	0.9135
a-R	0.01	0.99	0	1	0.9321
b-R	0.01	0.99	0	1	0.9135
a	0.01	0.99	0	1	0.9227
b	0.01	0.99	0	1	0.8953

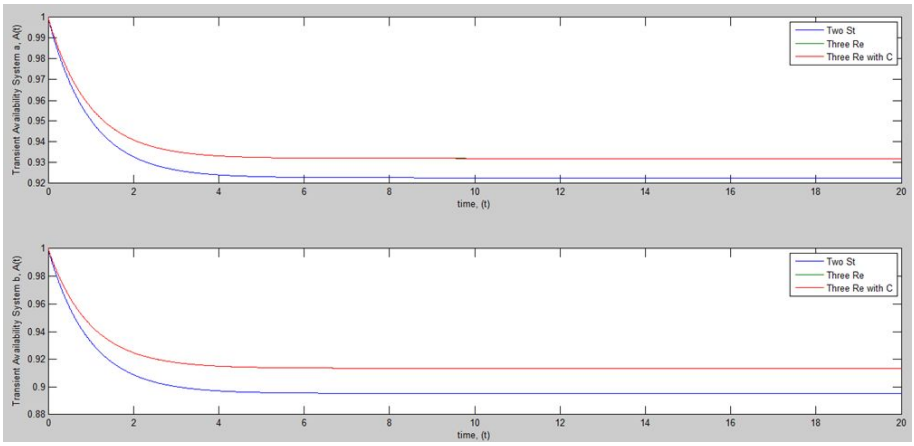


Figure 3.18: End-to-end availability, based on scenario a and b, of the three different models

show that the redundancy model with coverage factor provides similar end-to-end availability results to the redundancy model without coverage factor. This due to most characteristic of the computer system, the value of c is closed to one. However, combining the coverage factor into the model could provide insight failure behavior details of the system that is affected by c [8].

3.7.3 Availability analysis with one redundancy of S-CSCF unit via the five-state Markov model

Due to a present network system and equipment technology (both Hardware and Software) have high capabilities and stability enough for the core network components of repairable and redundancy system so that there are a few chances that the system will completely fail or goes back out from normal to completely failed state. Therefore, the five-state model with the assumption of soft and hard failure state is proposed. The proposed reliability analysis based on Markov model with redundancy. The system assumes to have a soft failure that can be recovered automatically and hard failure that mostly need to be manually repaired and take longer repairing period than soft failure state.

In reality, typically, two main types of failures can be observed; instant and degradation failures. Therefore, the proposed CTMC model comprise of two main types of failures: Soft Failure (SF) and Hard Failure (HF). SF can be defined as degradation or ordinary failure types (for both Hardware and Software) that can be automatically or manually recovered. HF is defined as instant or severe failure types that need longer recovery time than SF. The HF types comprise failures that require manual repair processes with many hours of recovery time. Each unit may fail with the failure rate (λ) and recover with recovery rate (μ). The coverage factor or the probability that the system will be recovered at a given fault is represented by (c). When SF occurred, the Soft failure recovery rate is represented by α. More failures can occur if the system cannot be recovered automatically, this lead to HF state. The recovery rate is represented by β. Then the recovery time spending in this state which is in time $\frac{1}{\beta}$ is longer than $\frac{1}{\alpha}$. Therefore there are possible five state spaces, $S = \{0, 1, 2, SF, HF\}$. State 2 represent two components both in working or normal condition. State 1 represents one working and one fail. State 0 represents failure state or both components failed. SF and HF represent Soft and Hard Failure state respectively. The system availability and reliability will be further derived from above model. From figure 3.19 and equation (25), $\frac{d\pi(t)}{dt} = \mathbb{Q}.\pi(t)$, the system state probability vector can be given by

$$\begin{bmatrix} \pi_2^{'}(t) \\ \pi_{HF}^{'}(t) \\ \pi_{SF}^{'}(t) \\ \pi_1^{'}(t) \\ \pi_0^{'}(t) \end{bmatrix} = \begin{bmatrix} -2\lambda & 0 & 0 & \mu & 0 \\ 2(1-c)\lambda & -\beta & 0 & 0 & 0 \\ 2\lambda c & 0 & -\alpha & 0 & 0 \\ 0 & \beta & \alpha & -(\lambda+\mu) & \mu \\ 0 & 0 & 0 & \lambda & \mu \end{bmatrix} \begin{bmatrix} \pi_2(t) \\ \pi_{HF}(t) \\ \pi_{SF}(t) \\ \pi_1(t) \\ \pi_0(t) \end{bmatrix}$$

$$\pi_2^{'}(t) = -2\lambda\pi_2(t) + \mu\pi_1(t) \tag{3.56}$$

$$\pi_{HF}^{'}(t) = 2(1-c)\lambda\pi_2(t) - \beta\pi_{HF}(t) \tag{3.57}$$

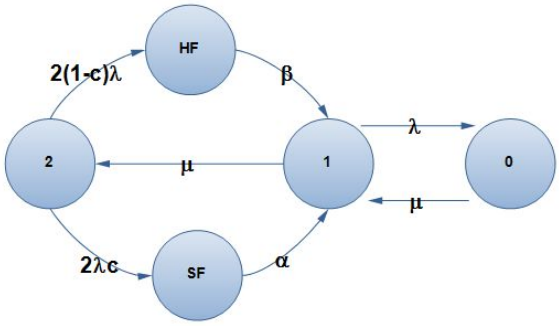


Figure 3.19: The continuous Markov model with one redundancy and two failure types

$$\pi_{SF}^{'}(t) = 2\lambda c\pi_2(t) - \alpha\pi_{SF}(t) \tag{3.58}$$

$$\pi_1^{'}(t) = \beta\pi_{HF}(t) + \alpha\pi_{SF}(t) - (\lambda+\mu)\pi_1(t) + \mu\pi_0(t) \tag{3.59}$$

$$\pi_0^{'}(t) = \lambda\pi_1(t) - \mu\pi_0(t) \tag{3.60}$$

With state probability conditions: $\pi_2(t) + \pi_{HF}(t) + \pi_{SF}(t) + \pi_1(t) + \pi_0(t) = 1$ and assuming both units are up for initial condition: $\pi_2(0) = 1, \pi_{HF}(0) = \pi_{SF}(0) = \pi_1(0) = \pi_0(0) = 0$. The transient and steady state probabilities can be calculated. Due to the long expression of transient solutions, only steady state probability will be presented as follow.

$$\pi_{2\,st} = \frac{\alpha\beta\mu^2}{2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2} \tag{3.61}$$

$$\pi_{HF\,st} = \frac{2\mu^2\alpha\lambda(1-c)}{2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2} \tag{3.62}$$

$$\pi_{SF\,st} = \frac{2\beta c\lambda\mu^2}{2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2} \tag{3.63}$$

$$\pi_{1\,st} = \frac{2\alpha\beta\lambda\mu}{2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2} \tag{3.64}$$

$$\pi_{0\,st} = \frac{2\alpha\beta\lambda^2}{2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2} \tag{3.65}$$

The transient and steady state availability is summation of all availability states which is state 2 and 1 and is given by

$$A_{Redundant-five-state-c}(t) = \pi_{2st}(t) + \pi_{1st}(t) \tag{3.66}$$

Moreover, the steady state availability of the components with no redundancy and redundancy will be given by considering $lim_{t\to\infty}A(t)$ of equation 3.24 and 3.66 and is equal to

$$\begin{aligned} lim_{t\to\infty}A(t) &= lim_{t\to\infty}(\frac{\mu}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda}e^{-(\mu+\lambda)t}) \\ &= \frac{\mu}{\mu+\lambda} \end{aligned}$$

$$lim_{t\to\infty}A_{redundant}(t) = \frac{\alpha\beta\mu^2 + 2\alpha\beta\lambda\mu}{(2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2)} \tag{3.67}$$

For comparison propose with the previous end-to-end availability models analysis, assuming the following recovery, failure rate and coverage factor values: μ=0.99,λ=0.01 ,and c=0.95. The soft failure recovery rate may refer to failure that can be recovered by rebooting, and the hard failure recovery rate may take longer time due to manually recovered. Then, assuming that α=0.99 and β=0.90. Therefore, the simulation results of end-to-end availability which is based on the proposed scenario analysis per figure 21 (a) and (b) can be given by table 3.7 and figure 3.20.

As shown from figure 3.20 and table 3.7, the end to end availability results from five-state Markov model is quite close to three-state and three-state model with coverage factor. From table 11, the availability of five-state model differs from the three-state model at the third precision. Therefore, the simulation results proved that five-state models can represent the availability of the system. Moreover with more details of different recovery and failure rate of the system, five-state Markov model will best represent system behaviors with a carefully selected value of λ, μ, c, α, and β.

Table 3.7: End-to-end availability $A(t)$ of the five state model (a-F-R-C) and (b-F-R-C) with a redundant unit of S-CSCF comparing with other models while λ=0.01, μ=0.99,c=0.95,α=0.99 and β=0.90

Availability, $A(t)$	λ	μ	Coverage factor, c	Maximum value,$A(t)$ Max	Minimum value,$A(t)$ Min
a-F-R-C	0.01	0.99	0.95	1	0.9316
b-F-R-C	0.01	0.99	0.95	1	0.9127
a-R-C	0.01	0.99	0.95	1	0.9321
b-R-C	0.01	0.99	0.95	1	0.9135
a-R	0.01	0.99	0	1	0.9321
b-R	0.01	0.99	0	1	0.9135
a	0.01	0.99	0	1	0.9227
b	0.01	0.99	0	1	0.8953

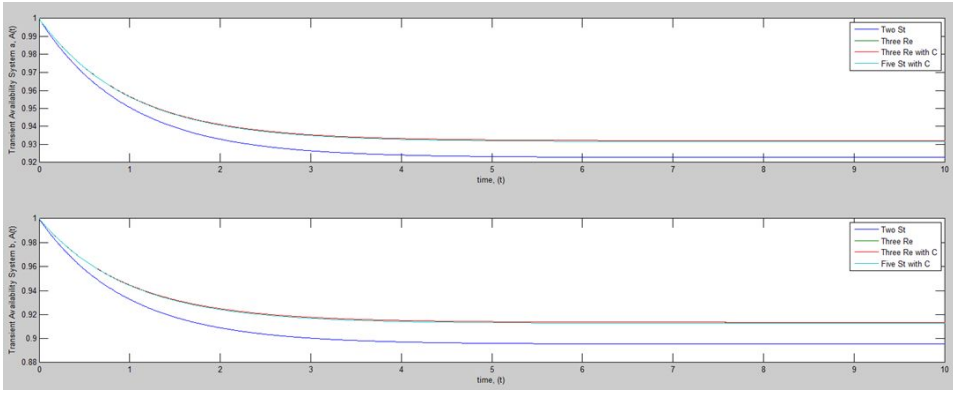


Figure 3.20: End-to-end availability, based on the communication scenario a and b, of four different models

3.8 Comparing Availability Analysis of a Redundancy of the S-CSCF Unit by Using the Five-state Markov Model and The Three-state Markov Model [7, 8]

Figure 3.21 shows the three-state Markov model [7, 8] and the proposed five-state Markov model with applied redundancy. The three-state Markov model which based on figure 3.21 (a) has been employed for a modeling reliability of repairable system due to the effect of coverage factor [8]. Also based on the same coverage concept, the model shown by figure 3.21 (b) has been used to evaluate the reliability of the IMS system. The coverage factor with duplex

failure or redundancy condition is also assumed in the model. In this section,

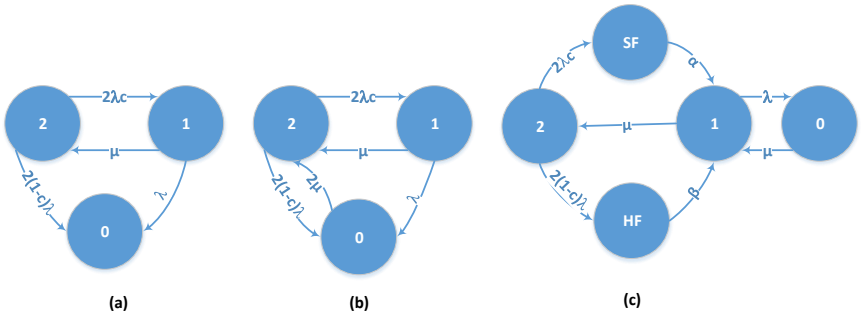


Figure 3.21: Comparisons between the three [(a) and (b)], and the five-state continuous time Markov model (c) with redundancy and coverage factor

the comparison of the proposed five-state and three-state Markov model [7] will be performed and simulated based on the same assumption of independent failure characteristics and exponentially distributed of failure and recovery rate. Moreover, the affect of coverage factor will be compared. Besides the effect of Soft failure and Hard failure recovery rates (α and β) characteristics of the proposed model to end-to-end availability will be simulated and analyzed for both network communication scenarios (a) and (b).

From the figure 3.21(b), there are three states, $S = \{0, 1, 2\}$. State 2 represents two units with both in normal working condition. State 1 represents one unit with a normal working condition while another one is failed and in the recovery process. State 0 represent two failure unit or system failed. From figure 3.21 (b) and equation 3.18, $\frac{d\boldsymbol{\pi} t}{dt} = \mathbb{Q}.\boldsymbol{\pi}(t)$, the system state probability equation can be given by

$$\begin{bmatrix} \pi_2^{'}(t) \\ \pi_1^{'}(t) \\ \pi_0^{'} \end{bmatrix} = \begin{bmatrix} -2\lambda & \mu & 2\mu \\ 2\lambda c & -(\mu+\lambda) & 0 \\ 2\lambda(1-c) & \lambda & -2\mu \end{bmatrix} . \begin{bmatrix} \pi_2(t) \\ \pi_1(t) \\ \pi_0(t) \end{bmatrix}$$

$$\pi_2^{'}(t) = -2\lambda\pi_2(t) + \mu\pi_1(t) + \mu\pi_0(t) \tag{3.68}$$

$$\pi_1^{'}(t) = 2\lambda c\pi_2(t) - (\mu + \lambda - \lambda c)\pi_1(t) \tag{3.69}$$

$$\pi_0^{'}(t) = -2\lambda(1-c)\pi_2(t) + \lambda\pi_1(t) - 2\mu\pi_0(t) \tag{3.70}$$

With state probability conditions: $\pi_2(t) + \pi_1(t) + \pi_0(t) = 1$ and assuming both units are up for initial condition: $\pi_2(0) = 1, \pi_1(0) = 0, \pi_0(0) = 0$. The transient and steady state probabilities can be calculated. Due to long expression of transient solutions, only steady state probability will be displayed as follow.

$$\pi_{2st}(t) = \frac{\mu^2 + \lambda\mu}{(2\lambda\mu + \lambda^2 + \mu^2 + c\lambda\mu)} \tag{3.71}$$

$$\pi_{0st}(t) = \frac{2c\lambda\mu}{(2\lambda\mu + \lambda^2 + \mu^2 + c\lambda\mu)} \tag{3.72}$$

$$\pi_{0st}(t) = \frac{\lambda(\lambda + \mu - c\mu)}{(2\lambda\mu + \lambda^2 + \mu^2 + c\lambda\mu)} \tag{3.73}$$

The transient and steady state availability are summation of all availability states which is state 2 and 1 and is given by

$$A_{Three-redundant-c}(t) = \pi_2(t) + \pi_1(t) \tag{3.74}$$

Also, the steady state availability of the components with no redundancy and redundancy will be given by considering $lim_{t\to\infty}A(t)$ of the equation 3.74, and is given by

$$\begin{aligned} lim_{t\to\infty}A_{Three-redundant-c}(t) &= \frac{\mu^2 + \lambda\mu}{(2\lambda\mu + \lambda^2 + \mu^2 + c\lambda\mu)} + \frac{2c\lambda\mu}{(2\lambda\mu + \lambda^2 + \mu^2 + c\lambda\mu)} \\ &= \frac{\mu^2 + (1+2c)\lambda\mu}{(2\lambda\mu + \lambda^2 + \mu^2 + c\lambda\mu)} \end{aligned} \tag{3.75}$$

Comparing with the proposed five-state Markov model which has the steady state availability per equation 3.67 below

$$lim_{t\to\infty}A_{redundant}(t) = \frac{\alpha\beta\mu^2 + 2\alpha\beta\lambda\mu}{(2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2)}$$

3.8.1 The effect of c to the steady state availability of repairable redundancy system using three-state model

First let consider steady state availability of the three-state model, by dividing both numerator and denominator by μ^2 equation 3.75 can be re-written as

$$A_{Steady-Three-redundant-c}(t) = \frac{1 + \frac{\lambda}{\mu}(1+2c)}{1 + (\frac{\lambda}{\mu})^2 + \frac{\lambda}{\mu}(3+c)} \tag{3.76}$$

As discussed by [8], normally the failure rate will be a lot less than recovery rate ($\lambda \lll \mu$), then the value of the ratio $\frac{\lambda}{\mu}$ will be very small and can be ignored. And also the multiplication of $\frac{\lambda}{\mu}$ with other terms can also be ignored. This lead to the steady state value can be approximated by

$$A_{Steady-Three-redundant-c}(t) \approx 1$$

That means the steady state value will mainly depend on the failure and recovery rate of the system. If the system characterizes by having the value of ($\lambda \lll \mu$) , the coverage factor c will have no effect with the steady state availability of the system. That means non-covered fault will not affect steady state availability due to high value of recovery rate μ. The conclusion of three-state model discussed by [8] or per figure 3.21(a) said that in case of non-covered fault (the system has low value of c), the recovery factor will have less effect to the mean time to first failure of the system due to the system is most likely to fail due to non-covered fault. However, this considered three-state model [7] and from the derived equation 3.76 shown that the steady state availability is influenced by the ratio of $\frac{\lambda}{\mu}$ event if non-recovered fault is occurred. Due to the system can be recovered back to the normal state from the complete failure or state zero. Moreover, the same analysis conclusion can be applied to the model represented by 3.7.3.

3.8.2 The effect of c,α,and β to the steady state availability of repairable redundancy system using the five-state model

The steady state of the five-state model, per equation 3.67, can also be rewritten when dividing both numerator and denominator by μ^2 as

$$A_{Five-redundant-c}(t) = \frac{\alpha\beta + \frac{\lambda}{\mu}(2\alpha\beta)}{\frac{\lambda}{\mu}(2\alpha\beta + \frac{\lambda}{\mu}2\alpha\beta) + 2\lambda(\alpha + \beta c - \alpha c) + \alpha\beta} \quad (3.77)$$

Let's consider the normal fact of the failure and recovery rate value that $\lambda \lll \mu$, then the second term of the numerator and the first term of the denominator can be ignored. Therefore, the ideal steady state when coverage factor c=1 is

$$A_{Five-redundant-c=1}(t) \approx \frac{\alpha}{2\lambda + \alpha}$$

Consequently the imperfect coverage factor can affect the ideal steady state availability by

$$\frac{A_{Five-redundant-c=1}(t)}{A_{Five-redundant-c}(t)} = \frac{2\lambda(\alpha + \beta c - \alpha c) + \alpha\beta}{\beta(2\lambda + \alpha)} \quad (3.78)$$

From equation 3.78, considering both of the numerator and denominator term, not only the coverage factor c but also the failure rate λ and the hard and soft failure recovery rates (α) are the main factors that can impact steady state availability represented by five-state repairable redundancy system. Therefore, to achieve high steady state availability, a high value of c,α or β is needed. Please be noted that at some point μ may have less effect with the steady state availability. However, the normal condition of the system that $\lambda \lll \mu$ need to exist in the first place.

3.8.3 The numerical example of steady state availability and the effect of c,α,and β of three and five-state model

For realistic and simplicity of the comparison and simulation, the failure and recovery parameters are approximated based on the statistical value of failure and recovery of optical and IP networks [9] and assumed for all IMS components. Accordingly, the fault rate is assumed to occur ten times per year or by average once every one and a half months (1104 hrs). The normal fault can be repaired in an average of two hours. And also the SF and HF are assumed to be recovered in an average of quarter-hour and five hours respectively. The comparison of the simulation results between two models at different coverage factor values are given below.

Table 3.8: End-to-end availability $A(t)$ of the five state model (a-F-R-C) and (b-F-R-C) with a redundant unit of S-CSCF comparing with other models while λ=0.01, μ=0.99,c=0.95,α=0.99 and β=0.90

Coverage factor	Five-state model	%increasing of As comparing at c = 0.9 (five-state)	Three-state model	%increasing of As comparing at c = 0.9 (three-state)
c=0.9	As=0.998710 $As-a$=0.990941 $As-b$= 0.988367	0%	As=0.999818 $As-a$= 0.998731 $As-b$= 0.998369	0%
c=0.95	As=0.999123 $As-a$=0.993882 $As-b$= 0.992141	0.042254%	As=0.999907 $As-a$= 0.999354 $As-b$= 0.999170	0.008901%
c=0.99	As=0.999462 $As-a$=0.996244 $As-b$= 0.995173	0.075598%	As=0.999979 $As-a$= 0.999853 $As-b$= 0.999811	0.016102%
c=0.995	As=0.999505 $As-a$=0.996540 $As-b$= 0.995554	0.080504%	As=0.999987 $As-a$= 0.999915 $As-b$= 0.999891	0.016903%
c=0.999	As=0.999539 $As-a$=0.996777 $As-b$= 0.995858	0.083908%	As=0.999994 $As-a$= 0.999961 $As-b$= 0.999950	0.017603%

As shown by the table 3.8, when increasing the coverage factor values, c, both steady state availability of the system itself (As) is increased. By considering only at the As value, the three-state model seems to have higher As than the five-state model. However, when considering the increasing percentage of c and As, the five-state model will have higher increasing percentage than the three-state model. The results correspond to an analysis per equations 3.76 and 3.78 that at a certain value of c, the steady state availability of the three-state model will not likely be affected by the value of c but will mostly depending on the ratio of λ and μ. Besides, the five-state model can not only represent a suitable value of c in the model but also the effect due to soft and hard failure can be classified, and the availability will also depend on the recovery factors α and β of the system. For an end-to-end steady state availability, when increasing the value of c, the availability is also increased for both models. The communication system scenario a is shown to have higher availability than the system scenario b. From the results, the five-state Markov model of repairable redundancy system is proved to show a suitable characteristic of both transient and steady state availability of the system. And also with two classifications of the system failures (SF and HF), more characteristics of failure and recovery rates can be included to evaluate the complex system behavior which is better than the normal three-state model.

3.9 Availability of The Simplex System

The simplex system can be modeled and presented by figure 3.22(a) that used for reliability and performance analysis [10, 11]. Our proposed idea of having two main types of failures: HF and SF can be used to model the simplex system as shown by figure 3.22(b). This item will consider and compare the existing model of the simplex system and the new proposed simplex model with incorporated HF and SF concepts. From the figure 3.22, let 1 represent working

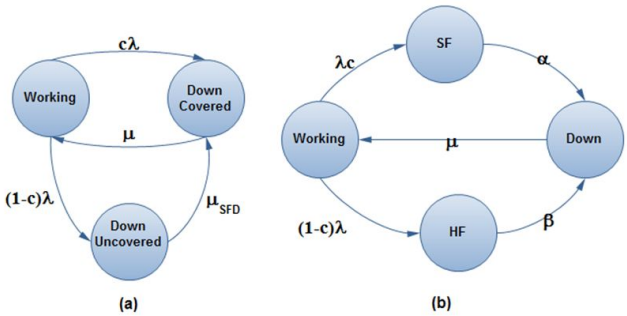


Figure 3.22: (a) The Markov model of the simplex system with coverage factor, (b) the proposed model with two main failures and coverage factor

state and 0_i represent down state with state descriptions:i. For example, figure 3.22(a), has three-state spaces, $S = \{0_c, 0_{uc}, 1\}$ where 0_c and 0_uc represent down covered and down uncovered states respectively. In the same way, figure 3.22(b) will have four-state spaces $S = \{0, 1, HF, SF\}$ where 0 refer to down state; HF and SF represent hard and soft failure respectively. By using the Kolmogorov differential equation, the system state probability equation of the simplex model per figure 3.22(a) can be given by

$$\begin{bmatrix} \pi_1^{'}(t) \\ \pi_{0_c}^{'}(t) \\ \pi_{0_{uc}}^{'}(t) \end{bmatrix} = \begin{bmatrix} -\lambda & \mu & 0 \\ \lambda c & -\mu & \mu_{SD} \\ \lambda(1-c) & 0 & -\mu_{SD} \end{bmatrix} . \begin{bmatrix} \pi_1(t) \\ \pi_{0_c}(t) \\ \pi_{0_{uc}}(t) \end{bmatrix} \tag{3.79}$$

or,

$$\begin{aligned} \pi_1^{'}(t) &= -\lambda\pi_1(t) + \mu\pi_{0_c}(t) \\ \pi_{0_c}^{'}(t) &= \lambda c\pi_1(t) - \mu\pi_{0_c}(t) + \mu_{SD}\pi_{0_{uc}}(t) \\ \pi_{0_{uc}}^{'}(t) &= \lambda(1-c)\pi_1(t) - \mu_{SD}\pi_{0_{uc}}(t) \end{aligned} \tag{3.80}$$

Where q represents self recovery rate from the down uncover state. By assuming both units are up for initial condition: $\pi_1(t) = 1, \pi_{0_c}(t) = 0, \pi_{0_{uc}}(t) = 0$ and with state probability conditions: $\pi_1(t) + \pi_{0_c}(t) + \pi_{0_{uc}}(t) = 1$, the transient and steady state probabilities can be calculated. Due to long expression of the transient solutions, only the steady state probabilities are given as follow.

$$\pi_1 = \frac{\mu_{SD}\mu}{\mu_{SD}(\lambda+\mu) + \lambda(\mu - c\mu)} \tag{3.81}$$

$$\pi_{0_c} = \frac{\lambda\mu_{SD}}{\mu_{SD}(\lambda+\mu) + \lambda(\mu - c\mu)} \tag{3.82}$$

$$\pi_{0_{uc}} = \frac{\lambda\mu(c-1)}{\mu_{SD}(\lambda+\mu) + \lambda(\mu - c\mu)} \tag{3.83}$$

Then, the steady state availability is summation of all availability states; therefore steady state availability of the simplex model per figure 3.22(a) is given by

$$A_{sp} = \pi_1 = \frac{\mu_{SD}\mu}{\mu_{SD}(\lambda+\mu) + \lambda(\mu - c\mu)} \tag{3.84}$$

Correspondingly, the system state probabilities of the proposed simplex model per figure 3.22(b) can be given by

$$\begin{bmatrix} \pi_1^{'}(t) \\ \pi_{HF}^{'}(t) \\ \pi_{SF}^{'}(t) \\ \pi_0^{'}(t) \end{bmatrix} = \begin{bmatrix} -\lambda & 0 & 0 & \mu \\ \lambda c & 0 & -\alpha & 0 \\ (1-c)\lambda & -\beta & 0 & 0 \\ 0 & \beta & \alpha & -\mu \end{bmatrix} \begin{bmatrix} \pi_1(t) \\ \pi_{HF}(t) \\ \pi_{SF}(t) \\ \pi_0(t) \end{bmatrix}$$

With the initial condition: $\pi_1(t) = 1, \pi_{HF}(t) = \pi_{SF}(t) = \pi_0(t) = 0$ and satisfy the probability conditions: $\pi_1(t) + \pi_{HF}(t) + \pi_{SF}(t) + \pi 0(t) = 1$. The steady state probabilities can be calculated and given as follow.

$$\pi_1 = \frac{\alpha\beta\mu}{\alpha\beta(\lambda+\mu) + \lambda\mu(\alpha + \beta c - \alpha c)} \tag{3.85}$$

$$\pi_{HF} = \frac{\mu\alpha\lambda(1-c)}{\alpha\beta(\lambda+\mu) + \lambda\mu(\alpha + \beta c - \alpha c)} \tag{3.86}$$

$$\pi_{SF} = \frac{\beta c\lambda\mu}{\alpha\beta(\lambda+\mu) + \lambda\mu(\alpha + \beta c - \alpha c)} \tag{3.87}$$

$$\pi_{0_{st}} = \frac{\alpha\beta\lambda}{\alpha\beta(\lambda+\mu) + \lambda\mu(\alpha + \beta c - \alpha c)} \tag{3.88}$$

Then, the steady state availability of the proposed simplex model is given by

$$A_{psp} = \pi_1 = \frac{\alpha\beta\mu}{\alpha\beta(\lambda+\mu) + \lambda\mu(\alpha + \beta c - \alpha c)} \tag{3.89}$$

3.9.1 The effect of c to the steady state availability of the simplex system using three-state model

Considering the steady state availability of the simplex system per equation 3.85, dividing both numerator and denominator by μ^2, then equation 3.85 can be given by

$$A_{sp} = \frac{\frac{\mu_{SD}}{\mu}}{\frac{\lambda\mu_{SD}}{\mu^2} + \frac{\mu_{SD}}{\mu} + \frac{\lambda}{\mu}(1-c)} \tag{3.90}$$

Due to the fact that normally $\lambda \lll \mu$, then the value of $\frac{\lambda}{\mu}$ is very small and can be ignored, and also the multiplication of $\frac{\lambda}{\mu}$ with other terms can also be ignored. Therefore, A_{sp} can be approximated and given by $A_{sp} \approx 1$. The result is similar to the analysis per 3.8.1 which means that the steady state availability mainly depends on the failure and recovery rate of the system. If the system characterize by having the value of $\lambda \lll \mu$, the coverage factor c will have no influence to the steady state availability of the system. That means non-covered fault will not affect steady state availability due to a high value of recovery rate μ. Another interesting fact is that the self-discovery rate μ_{SD} or normal recovery rate μ will have no influence to the steady state availability as long as the ratio of $\frac{\lambda}{\mu}$ corresponds to the fact that $\lambda \lll \mu$.

3.9.2 The effect of c,α,and β to the steady state availability of the proposed four states simplex model

The steady state availability of the four states simplex model per equation 3.90 can be rewritten by dividing both numerator and denominator by μ^2 as

$$A_{psp} = \frac{\frac{\alpha\beta}{\mu}}{\frac{\alpha\beta\lambda}{\mu^2} + \frac{1}{\mu} + \frac{\lambda}{\mu}(\alpha + \beta c - \alpha c)} \tag{3.91}$$

Similarly to 3.9.1, considering the normal system fact that $\lambda \lll \mu$, then equation 3.91 can be given as

$$A_{psp} \approx \alpha\beta \tag{3.92}$$

From equation 3.92, the steady state availability value of the proposed simplex model will not only be influenced by $\frac{\lambda}{\mu}$ but also the recovery factors, α and β. This means all failures and recovery rates will reasonably affect availability of the system for both transient and steady states of the system.

3.10 Reliability Analysis

The reliability of a unit or system can be defined as the probability that the unit or system can fully operate without interruption. The interruption of a service may occur due to different type of system failures for both hardware and software. Also, failure events that refer to interruption may vary depending on the considered system. Typically, two main types of failures can be perceived: instant and degradation failures. Therefore, the proposed CTMC model includes two main types of failures: Soft Failure (SF) and Hard Failure (HF). The SF can be defined as degradation failure types that can be automatically or manually recovered. The HF is defined as instant or severe failure types that sometimes need longer recovery time than SF. The failure regularly concern standard repaired processes with many hours of recovery time. The absorbing states of the CTMC model refer to the failure in operation of the system or unit. These states will be used to evaluate system reliability. In order to model the absorbing states, the failure characteristics of the system need to be defined. For instance, the figure 3.23, the SF or system reconfiguration may be considered as an absorbing state for real time communication services such as telephone or video conference. In addition, the HF and failure of the last unit events can be considered as an interruption for the redundancy system. Therefore, the proposed model can be classified into different reliability models based on different absorbing states scenarios. The advantage of this classification is that all feasible failure events or absorbing states can be estimated and

analyzed. The proposed model can be classified into three reliability models per figure 3.23 below. These models as well can adapt to represent reliability according to the QoS concepts. For example, the reliability models can represent reliability for three main QoS levels: low, medium, and high. The high-quality service should refer to the service that has no interruptions due to any faults. Therefore, the reliability model can be represented by figure 3.23 (a). Figure 3.23 (b) represents the medium-quality level that refers to services interruption due to HF and the repeated failure of the last unit. The low-quality level may refer to the services that have no recovery process. These services can apply re-transmission procedures in case of communication failures such as FTP and E-mail services. The reliability model of the low-quality is represented by figure 3.23(c).

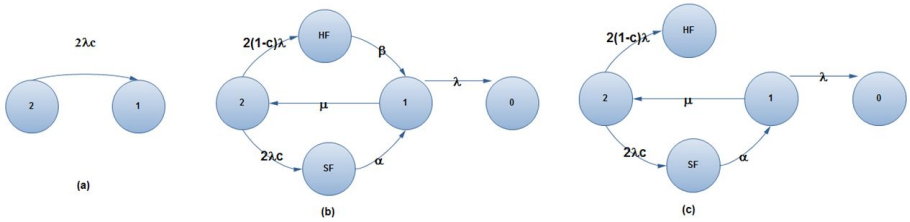


Figure 3.23: (a) Interruption case 1, any failure events can cause service interruption (b) Interruption case 2, only failure of the last active unit cause service interruption (c) Interruption case 3, HF and failure of the last active unit cause service interruption

3.10.1 Markov Reward Models

Due to the complexity of present communication systems and the need of fault-tolerant features, an effective evaluation technique is needed for such system. The Markov model and especially markov reward models (MRM) widely use for the evaluation. Those evaluations include reliability, performance, performability and dependability for the computer and communication systems. The MRMs can be simply described as the extended features model of the CTMC to analytically evaluate reliability, performance and combined measures such as dependability of the system. This can be done by applying a weight or so called a reward rate, r_i for each transition or state i of the CTMC. With this assignment, the CTMC can be redefined as the MRM. Hence, the expected or accumulate statistical values can be calculated per different reward rate values to represent different system characteristics [12].

The MRMs have been applied in Markov decision theory to assign cost and reward structures to states of the Markov processes for an optimization [13]. Moreover, MRMs was applied for an integration of performance and

dependability analysis by [14]. This work also forms the term performability which refers to the ability of the fault tolerant system that can perform some tasks while having failures. By applying MRMs, the rewards can be assigned to the state transition of CTMC. Then the reward can be defined based on the system requirement referring to availability, reliability and system tasks. Therefore, the system specifications can be unified with the model structure of the system by using MRMs. The equations 3.18 to 3.20 that were used to analyze CTMC will be applied as a fundamental equation for the MRMs analysis.

$$\begin{aligned} \frac{d\boldsymbol{\pi}(t)}{dt} &= \mathbb{Q}.\boldsymbol{\pi}(t),\ \pi(0) = \pi_0 \\ \mathbb{Q}.\pi &= 0 \\ \mathbb{E}.\pi &= 1 \end{aligned}$$

Where $\mathbb{E} = \begin{bmatrix} 1, & 1, & \cdots, & 1 \end{bmatrix}^T$, the superscript T denote the transpose. Then, the cumulative probabilities can be given by

$$\mathbb{L}(t) = \int_0^t \pi(u)du \tag{3.93}$$

Let $\mathbb{L}_i(t)$ indicates the expected total time of the CTMC spends in state i during the interval $[0, t)$. Then, the solution of equation 3.93 can be given by solving the differential equation 3.94 [15].

$$\frac{d\mathbb{L}(t)}{dt} = \mathbb{L}(t).\mathbb{Q} + \pi(0), \mathbb{L}(0) = 0 \tag{3.94}$$

where $\mathbb{I}$ is the identity matrix. Let r_i and τ_i represent the reward rate and the sojourn time spending in state $i \in S$. Then $r_i.\tau_i$ define the reward during the period τ_i. Let $\{X(t),\ t \geq 0\}$ represents a homogeneous finite-state CTMC with state space S. Then

$$Z(t) = r_{X(t)} \tag{3.95}$$

Where $Z(t)$ is the instantaneous reward rate of the MRMs at the time t. The overall reward rate $Z(t)$ of the MRMs represent the whole stochastic process of the system model where r_i is assigned as the reward rate of individual states. Therefore, the accumulated reward, $Y(t)$ for a given time $[0, t)$ can be given by

$$\begin{aligned} Y(t) &= \int_0^t Z(\tau)d\tau \\ &= \int_0^t r_{X(\tau)}d\tau \end{aligned} \tag{3.96}$$

Based on the definition of these three non-independent random variables,$X(t)$,$Y(t)$, and $Z(t)$, various measurement terms can be defined such as performability which is a function of $Y(t)$ [50].Another important term is called an expected instantaneous reward rate, $E[Z(t)]$ which can be obtained by solving different equation 3.18 and is given by equation 3.97.

$$E\left[Z(t)\right] = \sum_{i \in S} r_i \pi_i(t) \tag{3.97}$$

The expected reward rate when $t \to \infty$ (assuming the model have finite solution at $t \to \infty$) is calculated by solving the linear equation 3.19 and 3.20 and is given by equation 3.98.

$$\begin{aligned} lim_{t\to\infty}(\frac{1}{t}E[Y(t)]) &= E[Z(\infty)] \\ &= E[Z] \\ &= \sum_{i \in S} r_i \pi_i \end{aligned} \tag{3.98}$$

Where the expected accumulated reward rate,$E\left[Y(t)\right] = \sum_{i \in S} r_i L_i(t)$ can be written as

$$E\left[Y(t)\right] = \sum_{i \in S} r_i L_i(t) \tag{3.99}$$

3.10.2 System reliability by using MRMs

In order to design the resilience system, redundancy, single point of failure, and maintenance procedures are three main factors that need to be considered. Redundancy is the ability to protect and prevent malfunction of a unit by using another replaceable or standby unit. There are many types of redundancy such as standby, parallel, and $N : 1$ redundancy. In standby, the redundant unit starts operating after the primary unit fails, which is called cold standby and the switching period between the failed and the redundant unit is the main disadvantage. The parallel redundancy is called active redundancy. The difference between parallel and standby is that in parallel the redundant unit can operate instantaneously once the primary unit fails as all data and configurations are monitored by the redundant unit in real time. This is often called hot standby. The example of the parallel redundancy is the $1 : 1$ redundancy. The number refers to the one-to-one relationship between the active and redundant unit that supports hot standby. The $N : 1$ redundancy refers to the many to one relationship between the active units and the redundant unit where there is only one redundant unit for all N units. Therefore, any single point of failure which can cause malfunction of the system should be avoided in a resilient system. Then, all possibilities of a single point of failure should be identified and

prevented in the design processes. One method is by using the redundancy techniques. In addition, a good maintenance plan and procedures can also provide resilience of the system. Effective maintenance can reduce the risk of a failure or consecutive failures of the system. In order to evaluate the overall reliability of the system, an end-to-end reliability framework needs to be developed and analyzed along with failure, recovery, and redundancy aspects of the system. The IMS network reference architecture per figure 3.24 is applied for the IMS system reliability analysis. The border gateway control function (BGCF) and media gateway control function (MGCF) are included into the communication paths when considering the user from the legacy networks: a PSTN or a cellular network.

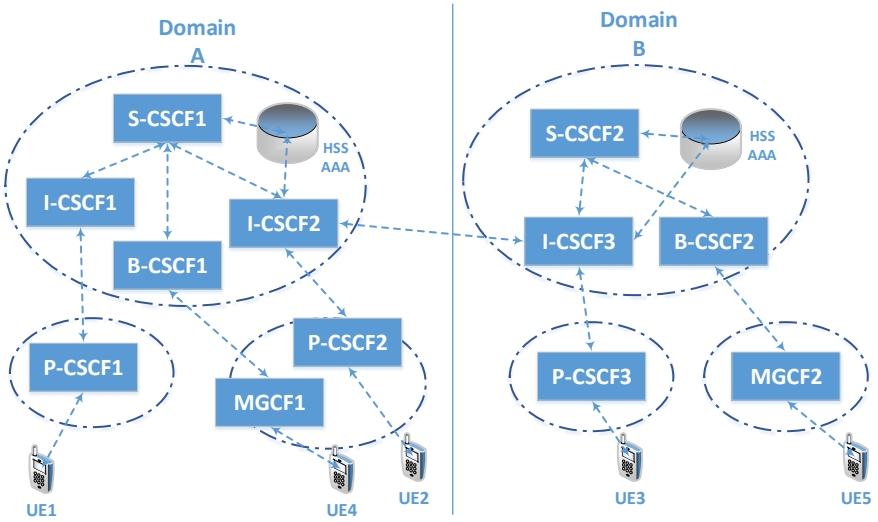


Figure 3.24: The IMS setup scenario between UEs on the different visited network and domains

The binary reward function with two types of reward rates, 1 and 0. The reward rate, 1 and 0, represents up and down states respectively. These values will be assigned for reliability analysis of the reference network. Therefore, the system reliability, $R(t)$ can be given as the probability of $Z(t) = r_{X(t)} = 1$ and is given by

$$\begin{aligned} R(t) &= P[T > t] \\ &= P[Z(t) = 1] \\ &= 1 - P[Z(t) \leq 0] \\ &= E[Z(t)] \end{aligned} \tag{3.100}$$

where T is the random variable of time to the failure event. From figure 3.23, the proposed five-state redundancy model is classified and represented as three reliability models. The models represent different failure conditions which will affect the reliability of the system. These models can represent reliability according to different QoS requirements. Let $R_k(t)$ represent reliability at a different service quality, k. Therefore, $R_k(t)$ can be evaluated through the MRMs per equation 3.100 and is given by

$$\begin{aligned} R_a(t) &= \pi_2(t) \\ R_b(t) &= \pi_2(t) + \pi_{HF}(t) + \pi_{SF}(t) + \pi_1(t) \\ R_c(t) &= \pi_2(t) + \pi_{HF}(t) + \pi_1(t) \end{aligned} \tag{3.101}$$

Where $\pi_2(t)$, $\pi_{HF}(t)$, $\pi_{SF}(t)$, $\pi_1(t)$, and $\pi_0(t)$ represent probabilities of each MRM state. From equation 3.101, the transient reliability represents interesting facts about the three different reliability models which corresponding to three levels of QoS: a,b and c. The $R_a(t)$ has the smallest value out of the three considered models. This due to it represents the reliability of service that needs high QoS level. Likewise, the $R_c(t)$ has the highest reliability value because it represent reliability of the system which require the lowest quality of service (QoS) level. Moreover, $R_b(t)$ represents reliability of the medium QoS service, therefore, its value is lower than $R_c(t)$ but higher than $R_a(t)$. In conclusion, the system reliability per equations 3.101, which are calculated by means of the MRM, can appropriately quantify the reliability of the system if the detailed reliability characteristics of the system are provided.

Therefore, in order to efficiently evaluate the reliability of the system, the main characteristic and requirements of the studied system are needed to be included into the model.

Reliability analysis based on parallel redundancy (1:1 redundancy)

Many research works have been focusing in performance and reliability analysis of the IMS core system with redundancy [10, 4, 16]. Different redundancy strategies had been applied to achieve high service availability and reliability. The Markov model with internal and external redundancy had been applied for reliability and availability analysis [10]. The system reliability was shown to increase with the user and system-initiated redundancy mechanisms. In other words, the redundant unit can be distributed any locations in the IP-based network. Therefore, both scalability and reliability of the IMS architecture can be achieved. Besides, by using the OpenIMS testbed, the system reliability results of the IMS system was shown to increase when having a redundancy of the S-CSCF which is one of the critical core IMS units [4]. Besides, the performance of the IMS core network was evaluated with optimization of the parallel redundancy by using the CTMC and Universal Generating Function

(UGF) [16]. The results showed that the system needs more than two parallel redundancy in order to achieve five nines reliability. However, a number of redundancies may not be required to efficiently run the real system. In reality, the reliability of the system will depend on many factors. For instance, the reliability of the system will not only depend on the core IMS system but its value can vary due to end-to-end communication architectures and resilience of each communication units or sub-system. Recently, the reliability and performability have been measured with the proposed MRMs framework [17]. The measures were presented with a well-suited binary reward structure. The MRMs approach was applied for the dependability and performance assessment of the fault-tolerant systems as an extension framework from CTMC [18].

Above all, it is time-consuming and difficult to collect and measure the exact reliability of the system. Nevertheless, from the above, the Markovian model approach can quantitatively and effectively evaluate reliability and performance behavior of the system if the failure and recovery characteristics are known. Accordingly, to study resilience properties of the IMS system at the minimum number of redundancy and to prevent a single point of failure, 1:1 redundancy is modeled for each IMS core component.

The reliability analysis of different IMS communication scenarios via the MRMs model.

The reliability analysis will apply the failure and recovery rates from [9] where the failure and recovery rate values are approximated based on the statistical values of the optical and internet protocal (IP) networks. These values are assumed for all IMS components. The fault rate is assumed to occur eight times per year, $\lambda = 2.4897 \times 10^{-7}\ sec^{-1}$, $\mu = 1.3889 \times 10^{-}4\ sec^{-1}$ (corresponding to the average time to repair of 120 min). The assumed average recovering time of SF and HF are 15 min and 300 min respectively, $\alpha = 1.1000 \times 10 - 3\ sec^{-1}$ and $\beta = 5.5556 \times 10 - 5$. The reliability model per figure 3.23(b), which is related to the medium QoS level, is assumed for all core IMS domains including the BGCF. The MGCF, interworking with traditional communication networks, is expected to have higher reliability and will be assumed with the reliability model per figure 3.23(c). Based on these assumptions, end-to-end reliability of communications at the same and different communication domains can be evaluated.

The communication scenarios are based on the reference network architecture of figure 3.24. However, with similar assumptions, the difference in reliability results for each scenario is very small. The difference of the each result can be observed only beyond the second or fifth digit for the each communication scenario. Therefore, instead of showing the graphical results of the transient reliability versus time, the major reliability values at the same point

and period are selected for an analysis. The results are presented by Table 3.9. From the table 3.9, the differences in reliability results at different scenarios are

Table 3.9: End-to-end reliability at different communication scenarios

Communication scenarios (similar home domain)	**Reliability**, $R_T(t)$
$UE1 - UE2$	0.998387377
$UE1 - UE3$	0.998387547
Communication scenarios (different home domain)	**Reliability**, $R_T(t)$
$UE1 - UE4$	0.998387037366
$UE1 - UE5$	0.998387037365

very small as expected due to the similar assumption of the failure and recovery rates. Nevertheless, the differences illustrate sufficient interesting end-to-end reliability characteristics of the system.

First, considering the communication scenarios within the same home domain. These are communication between SIP users and between a user from the legacy network and a session initiation protocol (SIP) user, which refer to communication scenarios between $UE1 - UE2$, and $UE1 - UE3$ respectively. The results show that communication between $UE1 - UE3$ has higher reliability than between $UE1 - UE2$. This is due to the legacy network equipment having higher reliability than the pure IP network. Therefore, within the same communication domain and with a similar number of components along the connection paths, the end-to-end reliability is more likely to increase with increasing reliability of any equipment within the connection paths.

When comparing the reliability results between the same and different communication network, the end-to-end reliability is reduced more than twice in term of difference percentage because communication across different home domains involves many components and longer connection paths. In particular, the connection paths between $UE1 - UE4$ involve less equipment than between $UE1 - UE5$. However, the difference percentage of the reliability results for these communication pairs is very small comparing with the same domain case. In other words, end-to-end reliability of long communication paths is affected by both varieties of the connection paths and the reliability of an equipment where the latter will have more effect. The results also support the advantage of performing geographical redundancy for the IMS system.

3.11 Conclusions

The availability and reliability models which are the fundamental requirements for end-to-end reliability evaluation are examined here. To conclude, the IMS communication architecture is studied. The proposed IMS-based reference communication network are presented to explore different IMS-based

communication scenarios for both intra-domain and inter-domain communication scenarios. In addition, the proposed CTMC and MRM models for both simplex and redundancy unit were applied with the RBD for end-to-end availability and reliability estimation. Moreover, the proposed model and method were simulated and compared with the state-of-the-art models. The proposed models were proven to provide better reliability characteristic details of the system than the current models. Moreover, the model simply exhibits reliability and availability outcomes due to the redundancy of various communication scenarios.

3.12 References

[1] Marko Čepin. Reliability block diagram. In *Assessment of Power System Reliability*, pages 119–123. Springer, 2011.

[2] Haitao Guo and Xianhui Yang. A simple reliability block diagram method for safety integrity verification. *Reliability Engineering & System Safety*, 92(9):1267–1273, 2007.

[3] Israel Koren and C Mani Krishna. *Fault-tolerant systems*. Morgan Kaufmann, 2010.

[4] Yong Liu, YiHong Liu, and Xingwei Wang. Reliability mechanism for the core control network-element s-cscf in ims. In *Network Computing and Information Security (NCIS), 2011 International Conference on*, volume 1, pages 57–61. IEEE, 2011.

[5] David Heimann, Nitin Mittal, Kishor S Trivedi, et al. Dependability modeling for computer systems. In *Reliability and Maintainability Symposium, 1991. Proceedings., Annual*, pages 120–128. IEEE, 1991.

[6] Robin A Sahner, Kishor Trivedi, and Antonio Puliafito. *Performance and reliability analysis of computer systems: an example-based approach using the SHARPE software package*. Springer Science & Business Media, 2012.

[7] Veena B Mendiratta and Himanshu Pant. Reliability of ims architecture. In *Telecommunication Networks and Applications Conference, 2007. ATNAC 2007. Australasian*, pages 1–6. IEEE, 2007.

[8] Thomas F Arnold. The concept of coverage and its effect on the reliability model of a repairable system. *Computers, IEEE Transactions on*, 100(3):251–254, 1973.

[9] Sofie Verbrugge, Didier Colle, Mario Pickavet, Piet Demeester, S Pasqualini, A Iselt, A Kirstädter, R Hülsermann, F-J Westphal, and

M Jäger. Methodology and input availability parameters for calculating opex and capex costs for realistic network scenarios. *Journal of Optical Networking*, 5(6):509–520, 2006.

[10] Xuemei Zhang, Hoang Pham, and Carolyn R Johnson. Reliability models for systems with internal and external redundancy. *International Journal of System Assurance Engineering and Management*, 1(4):362–369, 2010.

[11] Eric Bauer, Xuemei Zhang, and Douglas A Kimber. *Practical system reliability*. John Wiley & Sons, 2009.

[12] John I McCool. Probability and statistics with reliability, queuing and computer science applications. *Technometrics*, 45(1):107–107, 2003.

[13] Ronald A Howard. *Dynamic Probabilistic Systems, Volume II: Semi-Markov and Decision Processes*, volume 2. Courier Corporation, 2013.

[14] John F Meyer. Closed-form solutions of performability. *Computers, IEEE Transactions on*, 100(7):648–657, 1982.

[15] Andrew Reibman, Roger Smith, and Kishor Trivedi. Markov and markov reward model transient analysis: An overview of numerical approaches. *European Journal of Operational Research*, 40(2):257–267, 1989.

[16] Maurizio Guida, Maurizio Longo, and Fabio Postiglione. Performance evaluation of ims-based core networks in presence of failures. In *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, pages 1–5. IEEE, 2010.

[17] Kishor S Trivedi, Manish Malhotra, and Ricardo M Fricks. Markov reward approach to performability and reliability analysis. In *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 1994., MASCOTS'94., Proceedings of the Second International Workshop on*, pages 7–11. IEEE, 1994.

[18] Kishor S Trivedi, Jogesh K Muppala, Steven P Woolet, and Boudewijn R Haverkort. Composite performance and dependability analysis. *Performance Evaluation*, 14(3):197–215, 1992.

Chapter 4

High Availability and Reliability Optimization

4.1 Introduction

Service reliability and system reliability concern the consistency of giving its required function under a given condition. These days, it is a standout amongst essential quality terms for most of services and system, particularly considering the way that everything is accessible through full-time online services. Additionally, another imperative quality term called resilience has likewise increased more interest because of the term alludes to the system's capacity to be recouped from an inability to some particular working conditions [1, 2]. Subsequently, availability of the services and systems are critical qualities for the next-generation networks (NGN). System availability is the likelihood that the framework is in a prepared state to perform its capacities; the system reliability is the likelihood that the system can work without failure; plus, system resilience expresses how well the system can overcome the failure. Therefore, availability is part of the reliability furthermore resilience addresses the reliability level of the system.Consequently, these quality terms are firmly identified with one another and can demonstrate the system performance. Accordingly, measuring overall reliability characteristics of the system need to consider end-to-end quality evaluation framework. As specified in the previous chapters, it is hard to ensure end-to-end quality of an IMS framework because of different communication states between end clients and the geometrical of the systems. On the other hand, model-based reliability assessments have given significant information for the IMS system [3, 4]. Although, end-to-end reliability and availability of the IP multimedia subsystem (IMS) system have not been completely examined. Therefore, to increase reliability and availability of the system, this chapter investigate the overall reliability effect due to

the redundancy of the IMS-based network. Moreover, the chapter explored the redundancy optimization to further design high availability and reliability system.The related works are given in the following section.

4.2 Overview

Recently, there is different reliability evaluation techniques have been purposed for the IMS system. For instance, the state space methods such as Queuing Network Model (QNM) [5, 6, 7], Queuing Petri Nets (QPN) [8], and the Markov model [9, 10, 11, 12, 13, 14] were applied for performance and reliability evaluation of the IMS network. The QNM is regularly applied to analyze the processing delay. The QPN can attach some advantages of both Stochastic Petri Nets (SPN) and queuing networks models [8]. The Markov model can consider some detailed working states and system parameters such as failover and recovery rates. Moreover, there are many considerations of the redundancy influence for reliability evaluation [9, 12, 13, 14]. Furthermore, reliability at different failover success rates was evaluated by utilizing three-state Markov model [10]. The system downtime with redundancy was shown to influence by failover period and the coverage factor [11]. Besides, the system availability was shown to improve with redundancy mechanisms by using the Markov model [12]. Moreover, the combination of markov reward models (MRM) and reliability block diagram (RBD) was utilized for evaluating end-to-end reliability and resilience properties of the IMS system [13]. The paper showed that end-to-end reliability will be highly affected by the individual reliability of the system components. Further, the authors [9] proposed a combination of the five-state Markov model and RBD to evaluate end-to-end availability analysis of the intra-domain and inter-domain IMS-based communication network. The proposed model was compared with state-of-the-art Markov models and proved to provide better reliability behaviors of the simplex and redundant systems. In addition, the simulation results of end-to-end availability were shown to significantly improve especially for long distance communication when adding a redundancy of the serving-call state control function (S-CSCF) unit. Further, performance optimization of the IMS core network with parallel redundancy was evaluated by using the Markov model and Universal Generating Function UGF [14]. Furthermore, simulation of reliability improvement with a single redundancy of the S-CSCF unit was shown through the OpenIMS [15]. Later, the well-known network simulator (OPNET) was employed to simulate resilience characteristics of the IMS-based communication within similar and across the registered domain [16, 17]. In particular, the simulation results revealed various resilience behaviors of long distance communications and the redundancy effect. Therefore, a combination of models and simulation can be utilized for analyzing reliability and availability of the IMS-based network. However, an

overall reliability evaluation of the IMS system has not been completely investigated. This chapter extends the investigation of the parallel redundancy effects from the previous chapter (??). Moreover, this chapter exhibits an optimization of end-to-end reliability and availability of the IMS-based network at multiple communication scenarios by using a combination of the proposed three and five-state continuous-time Markov chain (CTMC) models and RBD models.

4.3 The Analysis Model

As mentioned in the previous chapters that availability is a function of reliability. This implies that reliability is a no failure probability of a system or similar to an availability with no single failure. The measurement of system availability solely expresses no meaningful details about the system such as a number of component failures or a number of replaced components. Therefore, both reliability and availability analysis of the system are needed for representing comprehensive reliability quality of the system. For comparison aims, similar reference network architecture to the analysis of 3 will be applied along with the proposed five-state CTMC model [9] as represented by figure 4.1 and figure 4.2 respectively. For comparison intentions, similar reference network

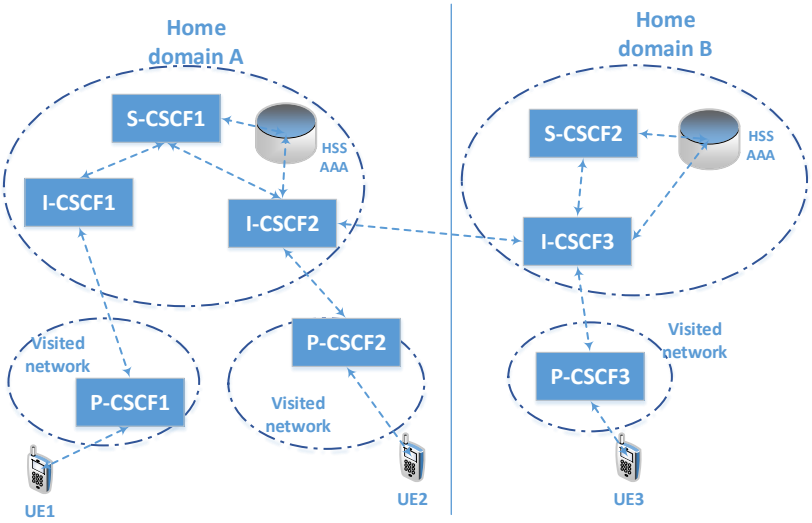


Figure 4.1: The IMS reference network of UEs located at the same and different IMS home domains

architecture to the analysis of 3 is applied along with the proposed five-state CTMC model [9] as represented by Figure and Figure respectively. Further,

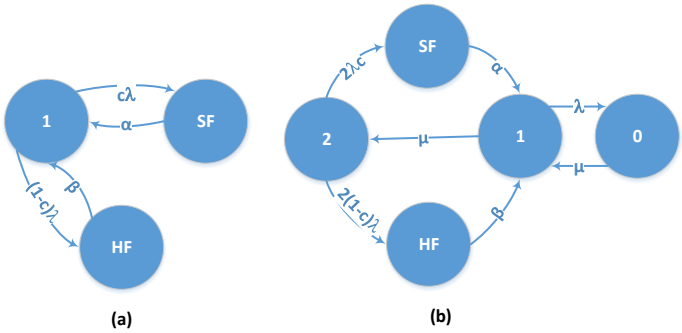


Figure 4.2: The three-state and five-state CTMC model: (a) simplex unit (b) redundant unit

the state probabilities of the proposed model, for both simplex and redundant unit are interpreted for evaluating system availability and reliability.

4.3.1 End-to-end availability analysis

From figure 4.1 (b), the model represents the proposed five-state CTMC of the redundant unit. The system state probabilities of the proposed model can be formulated as equation 4.1.

$$\begin{aligned}
\frac{d\pi_2(t)}{dt} &= -2\lambda\pi_2(t) + \mu\pi_1(t) \\
\frac{d\pi_{HF}(t)}{dt} &= 2(1-c)\lambda\pi_2(t) + \beta\pi_{HF}(t) \\
\frac{d\pi_{SF}(t)}{dt} &= 2\lambda c\pi_2(t) + \alpha\pi_{SF}(t) \\
\frac{d\pi_1(t)}{dt} &= \beta\pi_{HF}(t) + \alpha\pi_{SF}(t) - (\lambda+\mu)\pi_1(t) + \mu\pi_0(t) \\
\frac{d\pi_0(t)}{dt} &= \lambda\pi_1(t) - \mu\pi_0(t)
\end{aligned} \tag{4.1}$$

With the initial working condition where $\pi_2(0) = 1$, then the transient and steady state availability of all available states, state "2" and "1", can be assessed. Let $A_{5R}(t)$ represents the transient availability. Then, the steady state availability is given by equation 4.2.

$$\lim_{t\to\infty} A_{5R}(t) = \frac{\alpha\beta\mu^2 + 2\alpha\beta\lambda\mu}{2\alpha\lambda\mu^2 + 2\alpha\beta\lambda^2 + \alpha\beta\mu^2 + 2\alpha\beta\lambda\mu + 2\beta c\lambda\mu^2 - 2\alpha c\lambda\mu^2} \tag{4.2}$$

For the simplex system model as presented in figure 4.2 (a) where two detailed failures are utilized: state "1" represents the working state and state "0_i "

represents the down state with the state description i. The figure has three-state space S=1,HF,SF where "HF" and "SF" represent hard and soft failures respectively. In other words, both failures represent downstate or state "0" for a normal two-state CTMC. Accordingly, the system state probabilities can be formulated as equation 4.3.

$$\begin{aligned} \frac{d\pi_1(t)}{dt} &= -\lambda\pi_1(t) + \alpha\pi_{SF}(t) + \beta\pi_{HF} \\ \frac{d\pi_{SF}(t)}{dt} &= c\lambda\pi_1(t) - \alpha\pi_{SF}(t) \\ \frac{d\pi_{HF}(t)}{dt} &= (1-c)\lambda\pi_1(t) - \beta\pi_{HF}(t) \end{aligned} \tag{4.3}$$

With the working initial assumption, the transient availability can be measured. Let $A_s(t)$ represents the transient availability of the model, then, the steady state can be given by equation 4.4

$$\lim_{t\to\infty} A_s(t) = \frac{\alpha\beta}{\alpha(\beta+\lambda) - c\lambda(\alpha-\beta)} \tag{4.4}$$

The system availability of communication scenarios between two UEs, similar (UE1 and UE2) and different communication domains (UE1 and UE3), can be assessed by co-operating the RBD. The total availability depends on the analyzed network architecture, which can combine different parallel redundancy into the system. The RBDs of the call setup path for the communication scenarios are exhibited by the figure 4.3 and figure 4.4. Therefore, the total availability of the system is the total production result of unit availability of the series or parallel systems per each communication network scenario. Let $A_{Ts}(t)$ and $A_{Tp}(t)$ represent total availability of the series and parallel connection respectively. Then, the total availability equation can be given by

$$A_{Ts}(t) = \prod_{i=1}^{n} A_i(t) \tag{4.5}$$

$$A_{Tp}(t) = 1 - \prod_{i=1}^{n} (1 - A_i(t)) \tag{4.6}$$

Where $A_i(t)$ is the transient availability of the component i.

4.3.2 End-to-end reliability analysis

Consequently, total reliability of different end-to-end communication scenarios can be evaluated using the RBD. With the initial assumption that all communication units are reliable at the beginning of an operating period. Let $R_i(t)$

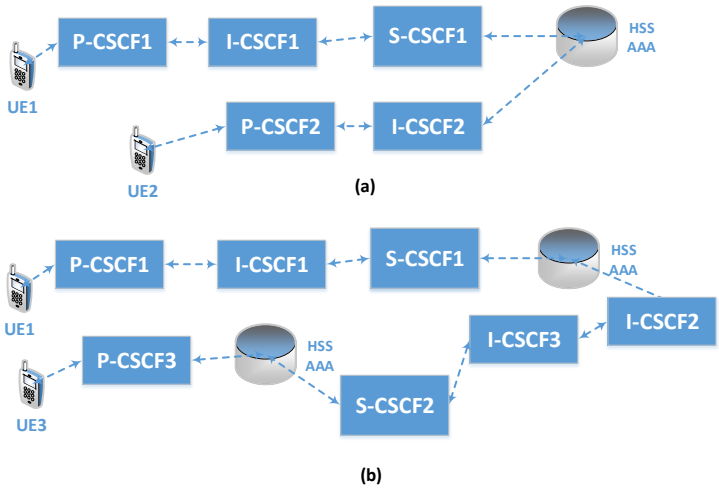


Figure 4.3: RBD of a communication network scenario: (a) similar home domain ($UE1\&UE2$) (b) different home domain ($UE1\&UE3$).

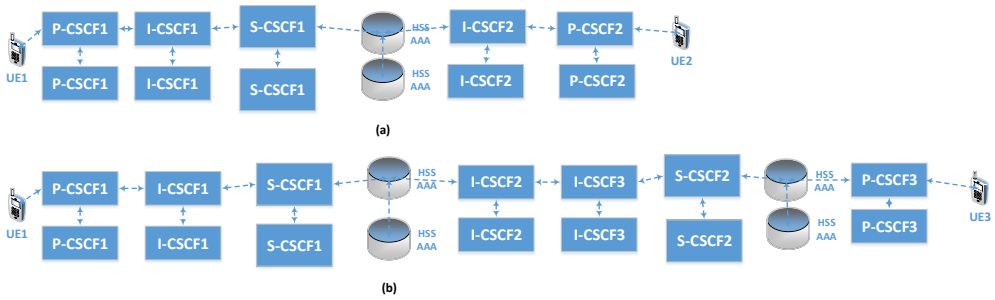


Figure 4.4: RBD of a communication network scenario with parallel redundancy (a) similar home domain ($UE1\&UE2$) (b) different home domain ($UE1\&UE3$)

represent the transient reliability of each component where i indicates each IMS core unit. Therefore, the total reliability for series and parallel systems can be given by R_{Ts} and R_{Tp} respectively.

$$\begin{aligned} R_{Ts}(t) &= \prod_{i=1}^{n} \pi_i(t) \\ &= \prod_{i=1}^{n} R_i(t) \end{aligned} \tag{4.7}$$

$$\begin{aligned} R_{Tp}(t) &= 1 - \prod_{i=1}^{n}[1 - \pi_i(t)] \\ &= 1 - \prod_{i=1}^{n}[1 - R_i(t)] \end{aligned} \tag{4.8}$$

Where the $R_i(t)$ is the transient reliability of unit i. Therefore, with reliable initial condition and exponential distribution assumption, the total reliability for each communication scenarios per figure 4.3 and figure 4.4 can be given by the following equations:

$$R_{T_{3a}}(t) = \left\{[e^{-2\lambda_{UE}(t)}][e^{-2\lambda_P(t)}][e^{-2\lambda_I(t)}][e^{-2\lambda_S(t)}][e^{-2\lambda_H(t)}]\right\} \tag{4.9}$$

$$R_{T_{3b}}(t) = \left\{[e^{-2\lambda_{UE}(t)}][e^{-2\lambda_P(t)}][e^{-3\lambda_I(t)}][e^{-2\lambda_S(t)}][e^{-2\lambda_H(t)}]\right\} \tag{4.10}$$

$$\begin{aligned} R_{T_{4a}}(t) =& [e^{-2\lambda_{UE}(t)}][1 - \left\langle e^{-2\lambda_P(t)} \right\rangle^2]^2[1 - \left\langle e^{-2\lambda_I(t)} \right\rangle^2]^2 \\ & [1 - \left\langle e^{-2\lambda_S(t)} \right\rangle^2][1 - \left\langle e^{-2\lambda_H(t)} \right\rangle^2] \end{aligned} \tag{4.11}$$

$$\begin{aligned} R_{T_{4b}}(t) =& [e^{-2\lambda_{UE}(t)}][1 - \left\langle e^{-2\lambda_P(t)} \right\rangle^2]^2[1 - \left\langle e^{-2\lambda_I(t)} \right\rangle^2]^2 \\ & [1 - \left\langle e^{-2\lambda_S(t)} \right\rangle^2]^2[1 - \left\langle e^{-2\lambda_H(t)} \right\rangle^2]^2 \end{aligned} \tag{4.12}$$

Where λ_{UE},λ_P, λI, λ_S, and λ_A are the failure rate of the core IMS units:UE, P-CSCF, I-CSCF, S-CSCF, and AAA.

4.4 The Fault- Tolerant System Models: The M-out-of-N Reliability Model and Optimization

The number of parallel redundancy can increase the reliability of the subunit or system. Therefore, increasing the number of parallel server units for each subunit can clearly enhance availability and reliability of the subunit and system. Hence, redundancy can produce high availability and reliability or fault-tolerant system. The figure represents the M-out-of-N model for each SIP server unit. The model is a very traditional type of redundancy toward fault-tolerant systems. The model represents the N-component system and the system will fail

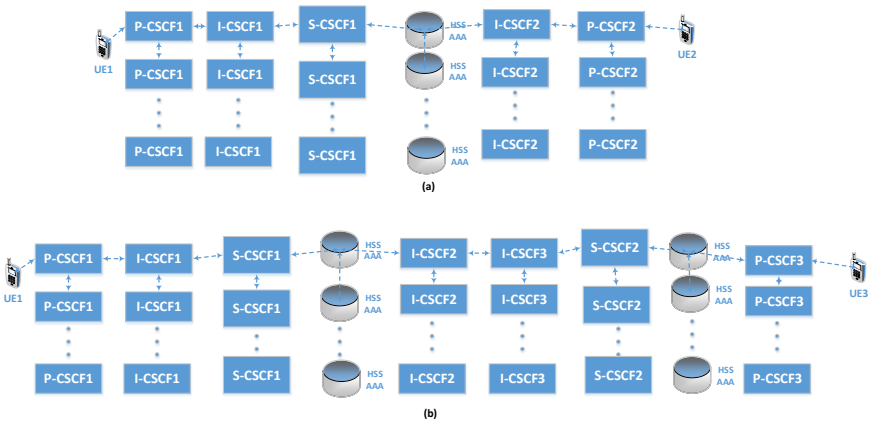


Figure 4.5: RBD of a communication network scenario with N-parallel redundancy (a) similar home domain ($UE1\&UE2$) (b) different home domain ($UE1\&UE3$)

if at least M of the N unit fails. With independent properties assumption, the reliability of the model can be evaluated by using the binomial distribution as

$$R_S(M,N,R) = \sum_{r=M}^{N} R^r(1-R)^{N-r} \tag{4.13}$$

Where R_s is the total transient reliability of the subnit, N is a total number of parallel redundancy, and M is the minimum number of unit required for the subunit to function, and R represents transient reliability of each unit.Then, the total reliability for the fault tolerant model of the communication scenario per figure 4.5 can be given by

$$\begin{aligned} R_{T_{5a}}(t) =& [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^2_{UE} \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^2_{P} \\ & \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]_{I} \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]_{S} \\ & \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^2_{H} \end{aligned} \tag{4.14}$$

$$R_{T_{5a}}(t) = [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^2_{UE} \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^2_{P} \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^2_{S} \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^3_{I} \times [\sum_{r=M}^{n} \binom{N}{r} R^r(1-R)^{N-r}]^2_{H} \quad (4.15)$$

From equations 4.14 and 4.15, in order to gain maximum system availability and reliability, the optimization between a parallel redundancy number and total reliability and availability is needed. This due to, in practical, there are several related factors such as equipment cost, system complexity, system maintenance, and management. According to the high availability system requirement where the system aims to ensure some operational performance level and the performance is represented by different availability values such as three-nines system availability (99.9%) means that the system has only 8.76 hours downtime per year. In this chapter, various high availability and reliability standard values will be examined. The five-nines (99.999%) and 6-nines (99.9999%) system availability, which is replying to the system downtime of 5 minutes per year and 31 seconds per year respectively, may be suitable for the NGN services and system.

4.5 Simulation and Discussion

Due to the fact that a failure rate value is less than one and is significantly lower than the recovery rate ($\lambda \lll \mu$), therefore, from the communication topology and equations 4.5-4.15 , we can simply analyze that long communication path scenario has less reliability than the short communication scenario path: scenario (a) has higher reliability than scenario (b) as shown in figure 4.3, 4.4, and 4.5. This also implies that a less complex communication scenario would have less failure probability or high reliability. However, simulation results are needed to provide an insight system reliability information. The effective five-state redundancy model [9] and the proposed three-state model per figure 4.2(a) and 4.2(b) respectively are applied for the simulation. The simulation results are different from [9] where the five state CTMC model was only applied for the redundancy model. The similar failure and recovery rates of realistic optical and IP networks [18] and the similar assumption with [9] are assumed for comparison purpose where the failure is assumed to occur eight times per year, λ = 2.4897×10-7 sec_{-1}, μ = 1.3889×10-4 sec_{-1} (corresponding to the average time to repair of 120 min). The assumed average recovering time of

SF and HF are 15 min and 300 min. Then, simulation results of different communication scenarios and conditions are given in the following subsections.

4.5.1 Intra domain communication: simplex and redundancy models

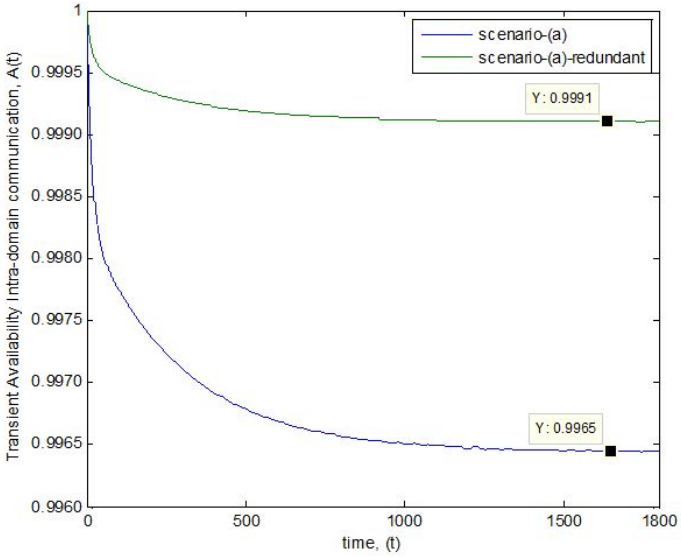


Figure 4.6: End-to-end availability results of intra-domain communications (with and without redundancy)

The simulation results are presented by figure 4.6 and figure 4.7 for intra-domain communication scenarios between UE1 and UE2 with and without single redundancy respectively. With initial reliability assumption, the overall availability is decreased versus time and remained steady at the steady state liked period. The end-to-end availability is increased when having a redundancy. The percentage change at the steady state is about 0.26%. This percentage change level is quite high in term of an overall availability or percentage of uptime per year. Moreover, an overall availability characteristics tend to move toward steady state liked period faster with redundancy. Accordingly, decreasing of end-to-end reliability is directly proportional to an operating period. Unlike availability characteristics, the reliability has no steady state liked region due to it represent a no failure probability. Besides, overall reliability results are not improved with the redundancy case. This implies that involving more equipment into the system increase complexity and decrease overall reliability of the system, even though an overall availability is improved with redundancy from the beginning of an operating period.

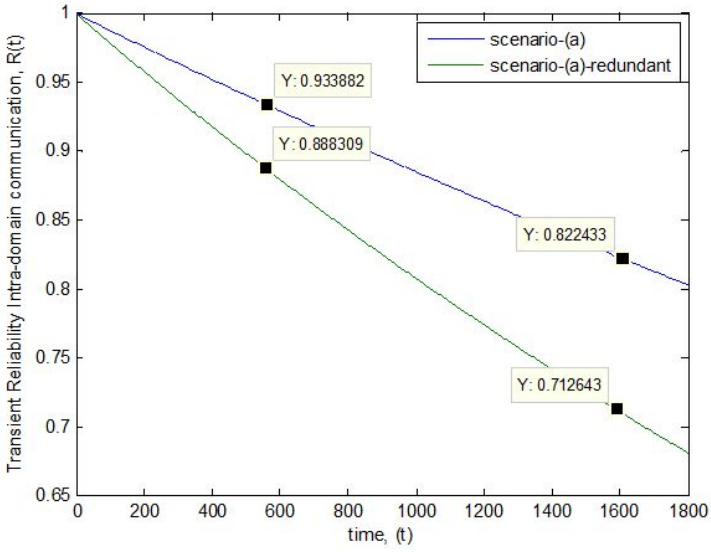


Figure 4.7: End-to-end reliability results of intra-domain communications (with and without redundancy)

4.5.2 Inter domain communication: simplex and redundancy models

The end-to-end availability and reliability results of the inter-communication scenarios are presented by figure and figure . Similar to the intra-domain communication scenarios, the availability results are decreased and running toward steady state liked region at some period. Besides, the results reach steady state liked region faster with redundancy. The percentage change of the availability results when adding redundancy is about 0.40%. The percentage change is quite high and is about two times higher than the intra-domain communication case. Therefore, redundancy is very effective with overall availability characteristics for the inter-domain communication or long communication setup path. For the end-to-end reliability results, it is decreased when increasing operating period. Moreover, the falling slopes of the reliability results are steeper than the intra-domain communication cases. To put it more simply, increasing communication equipment highly affects overall reliability characteristics of the system regardless of the number of redundancies. This effect can be clearly observed at the beginning of the operating period. Moreover, end-to-end quality of both reliability and availability are lower than the intra-domain communication scenarios.

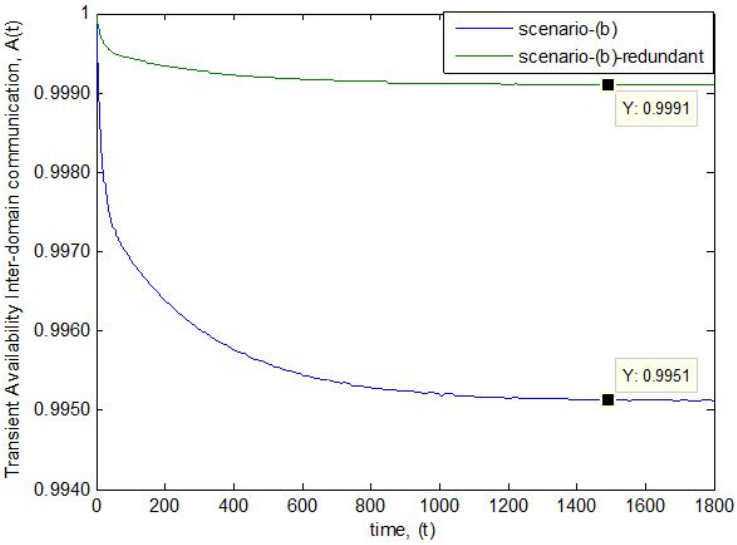


Figure 4.8: End-to-end availability results of inter-domain communications (with and without redundancy)

4.6 Comparison Between Intra-domain and Inter-domain Communications

The transient characteristics of end-to-end availability and reliability of all communication scenarios are presented by figure 4.10 and figure 4.11 respectively.

4.6.1 End-to-end availability

The results prove that intra-domain communication scenario has higher availability than communication across the different home domain. With single redundancy, significant improvement of end-to-end availability can be observed. In particular, redundancy highly influences inter-domain communication scenarios by having an almost similar level with the intra-domain communication cases. The percentage difference of the availability gap between without and with redundancy case of intra-domain communication is $\approx 0.26\%$ and is $\approx 0.40\%$ for inter-communication domain. While, the percentage difference of the availability gap between intra-domain and inter-domain communication of no redundancy case is $\approx 0.14\%$ and is $9 \times 10^{-4}\%$ for the redundancy cases. Besides, the percentage difference of the availability gap between intra-domain

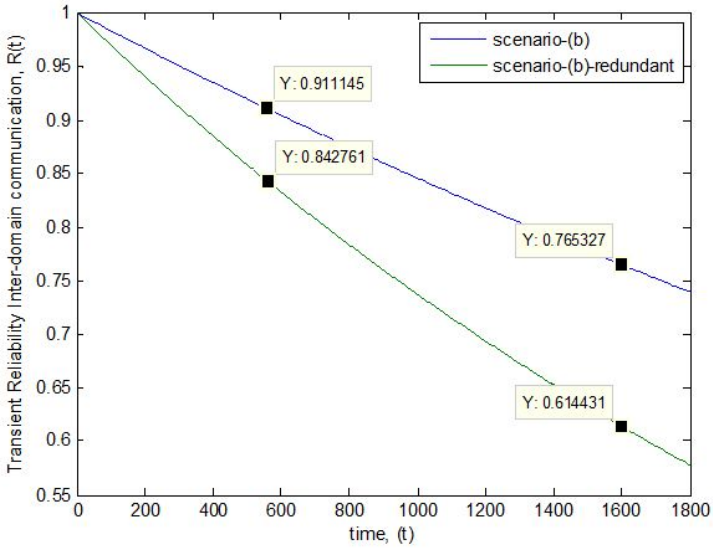


Figure 4.9: End-to-end reliability results of inter-domain communications (with and without redundancy)

and inter-communication domain is reduced by 10^4 times. Therefore, redundancy is needed to improve the end-to-end availability, especially for inter-communication domain or long communication setup paths.

4.6.2 End-to-end reliability

The transient characteristics of end-to-end reliability tend to decrease when the communication scenarios involve many components or signaling setup paths. Moreover, with redundancy, total reliability is decreased more than without redundancy cases. In particular, the declining percentage change of the reliability gap between without and with redundancy case of intra-domain communication is $\approx 11.18\%$ and is $\approx 16.31\%$ for inter-domain communication. For the percentage difference of the reliability gap between intra-domain and inter-domain communication are $\approx 5.93\%$ and $\approx 11.87\%$ for no redundancy and redundancy case respectively. Therefore, redundancy can magnify the complexity and can diminish end-to-end reliability quality of the system.

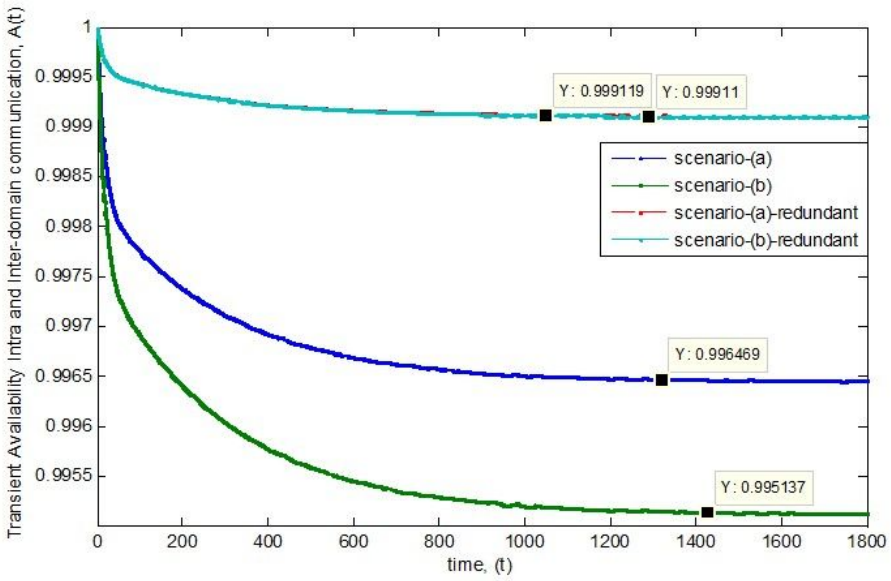


Figure 4.10: End-to-end availability results of the intra and inter domain communication scenarios (with and without redundancy)

4.7 Optimization Results of the Fault-Tolerant System Models

From the previous simulation results of the subsection 4.6.2, increasing redundancy into the system may end up increasing the complexity and decreasing overall reliability of the system. However, adding redundancy has significant effect to an overall availability of the system especially for intra-domain or long distance communication scenarios. Therefore, high availability does not refer to high reliability. On the other hand, high reliability would refer to high availability. With the with initial redundancy or M=2 and similar failure rate assumption for comparison purpose, $\lambda = 2.4897 \times 10^{-7}\ sec^{-1}$, and the high availability concept and from (15) and (16), the minimum redundancy number to achieve different end-to-end reliability as relating to high availability concept can be estimated and given in Table 4.1. From the results of table 4.1, the two and three nines system reliability require at least four parallel redundancy of each core IMS unit for both intra-domain and inter-domain communication cases. This implies that in order to achieve 8.76 hours of the maximum system downtime per year, we need a minimum total number of 18 and 27 core IMS units for intra-domain and inter-domain communication core IMS units

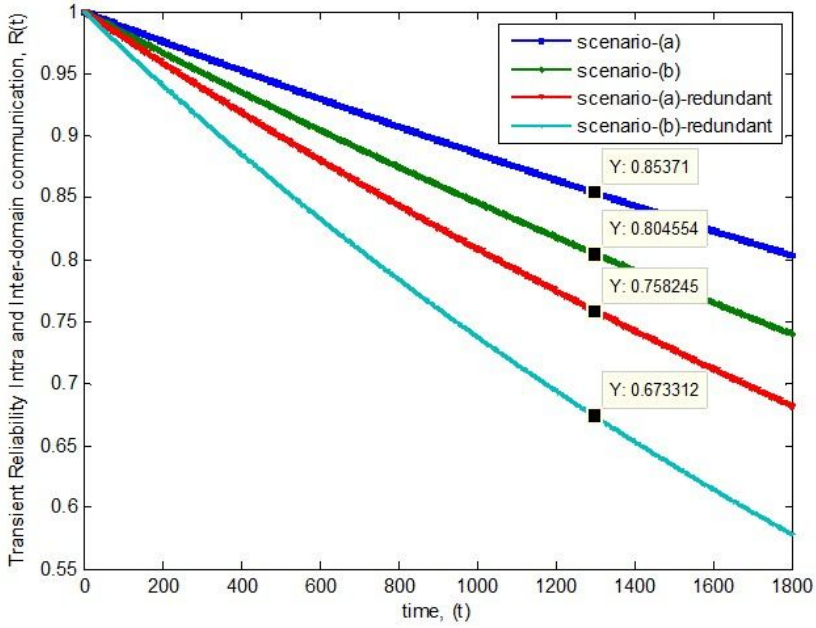


Figure 4.11: End-to-end reliability results of the intra and inter domain communication scenarios (with and without redundancy)

Table 4.1: Minimum redundancy unit at different end-to-end reliability requirement of intra-domain and inter-domain communication scenarios

End-to-end Reliability	Intra-domain (min. no. of redundancies)	Inter-domain (min. no. of redundancies)	% increasing of N
0.99%	4	4	0 %
0.999%	4	4	0 %
0.9999%	5	5	25 %
0.99999%	5	5	0 %
0.999999%	5	5	0 %

respectively. Therefore, the percentage difference of the total unit between the intra-domain and inter-domain system is 40%. So, the required total unit of the inter-domain communication is less than twice of the intra-domain communication case. Moreover, for producing four to six nines system reliability or the system downtime per year equal to 31.5 seconds, we need at least five parallel redundancies for each core IMS units. Comparing with the three nines condition, the different percentage of the increasing number of redundancy is 25%. In addition, the system could support up to six nines system reliability which

is equivalent to the percentage change of 0.1%. This changing amount represents a very significant change of the system reliability in term of probability. Obviously, the average system downtime per year was shown to improve from 8.76 hours to be 31.5 seconds which is almost a thousand times improvement. Moreover, the system needs a total number of 30 and 45 core IMS units for intra-domain and inter-domain communication. In particular, the percentage difference is 40% and is similar to the three nines condition. Therefore, with the similar increasing ratio of the total unit, six nines reliability condition could be managed. These optimization results signify that a particular amount of parallel redundant unit could be evaluated and optimized to produce a desired end-to-end reliability and availability system.

4.8 Conclusions

In this chapter, end-to-end availability and reliability of the IMS system were evaluated by using the proposed reliability and availability models (for both single and redundancy). In particular, the parallel redundancy effects were manifested at different IMS-based communication scenarios. The simulation results show that the IMS-based system availability can be significantly developed by adding redundancy especially for inter-domain or long distance communication scenarios. Nevertheless, adding redundancy unit could end up increasing system complexity and decreasing system reliability. Moreover, an optimization of the number of parallel redundant unit that corresponding to high availability and reliability system at different IMS-based communication scenarios is represented. The results demonstrate an interesting fact that high system availability and reliability can be reached with a proper amount of the IMS core redundant unit. Therefore, there is a possibility to build a fault tolerant or high-reliability system where a high-performance computer cost become cheaper in the near future.

4.9 References

[1] Carlos Queiroz, SK Garg, and Zahir Tari. A probabilistic model for quantifying the resilience of networked systems. *IBM Journal of Research and Development*, 57(5):3–1, 2013.

[2] Achim Autenrieth and Andreas Kirstädter. Engineering end-to-end ip resilience using resilience-differentiated qos. *Communications Magazine, IEEE*, 40(1):50–57, 2002.

[3] Patrick O'Connor and Andre Kleyner. *Practical reliability engineering*. John Wiley & Sons, 2011.

[4] Mohammad Modarres, Mark P Kaminskiy, and Vasiliy Krivtsov. *Reliability engineering and risk analysis: a practical guide*. CRC press, 2009.

[5] Lukas Nagy, Jurgen Tombal, and Vit Novotny. Proposal of a queueing model for simulation of advanced telecommunication services over ims architecture. In *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, pages 326–330. IEEE, 2013.

[6] IM Mkwawa and DD Kouvatsos. Performance modelling and evaluation of handover mechanism in ip multimedia subsystems. In *Systems and Networks Communications, 2008. ICSNC'08. 3rd International Conference on*, pages 223–228. IEEE, 2008.

[7] Wang Jianhui, Jin Hao, and Wu Wenguang. A novel queuing model for ims-based iptv system. In *Broadband Network & Multimedia Technology, 2009. IC-BNMT'09. 2nd IEEE International Conference on*, pages 560–564. IEEE, 2009.

[8] An'an Luo, Chuang Lin, Kai Wang, Fengyuan Ren, and Limin Miao. Performance modeling and evaluation using queuing petri nets in ims. In *Communications and Networking in China, 2009. ChinaCOM 2009. Fourth International Conference on*, pages 1–5. IEEE, 2009.

[9] Chayapol Kamyod, Rasmus Hjorth Nielsen, Neeli Rashmi Prasad, and Ranga Prasad. End-to-end availability analysis of ims-based networks: Simplex and redundant systems. In *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*, pages 1103–1108. IEEE, 2013.

[10] Veena B Mendiratta and Himanshu Pant. Reliability of ims architecture. In *Telecommunication Networks and Applications Conference, 2007. ATNAC 2007. Australasian*, pages 1–6. IEEE, 2007.

[11] Thomas F Arnold. The concept of coverage and its effect on the reliability model of a repairable system. *Computers, IEEE Transactions on*, 100(3):251–254, 1973.

[12] Xuemei Zhang, Hoang Pham, and Carolyn R Johnson. Reliability models for systems with internal and external redundancy. *International Journal of System Assurance Engineering and Management*, 1(4):362–369, 2010.

[13] Chayapol Kamyod, Rasmus Hjorth Nielsen, Neeli Rashmi Prasad, and Ramjee Prasad. Resilience in ims: End-to-end reliability analysis via markov reward models. In *Wireless Personal Multimedia Communications (WPMC), 2012 15th International Symposium on*, pages 564–568. IEEE, 2012.

[14] Maurizio Guida, Maurizio Longo, and Fabio Postiglione. Performance evaluation of ims-based core networks in presence of failures. In *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, pages 1–5. IEEE, 2010.

[15] Yong Liu, YiHong Liu, and Xingwei Wang. Reliability mechanism for the core control network-element s-cscf in ims. In *Network Computing and Information Security (NCIS), 2011 International Conference on*, volume 1, pages 57–61. IEEE, 2011.

[16] Chayapol Kamyod, Rasmus Hjorth Nielsen, Neeli Rashmi Prasad, and Ranga Prasad. Ims intra-and inter domain end-to-end resilience analysis. In *Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 2013 3rd International Conference on*, pages 1–5. IEEE, 2013.

[17] Chayapol Kamyod, Rasmus Hjorth Nielsen, Neeli Rashmi Prasad, and Ranga Prasad. Resilience of the ims system: The resilience effect of inter-domain communications. In *Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 2014 4th International Conference on*, pages 1–4. IEEE, 2014.

[18] Sofie Verbrugge, Didier Colle, Mario Pickavet, Piet Demeester, S Pasqualini, A Iselt, A Kirstädter, R Hülsermann, F-J Westphal, and M Jäger. Methodology and input availability parameters for calculating opex and capex costs for realistic network scenarios. *Journal of Optical Networking*, 5(6):509–520, 2006.

Chapter 5

Modeling and Simulation

5.1 Introduction

There are many methods for evaluating network reliability and performance such as mathematical modeling, network simulation (time-based or discrete event-based simulation), hybrid simulation (both analysis and simulation) and test-bed emulation. Mathematical model analysis can quickly provide insight information, but inaccuracy can occur due to some assumptions or approximations. Besides, not all studied cases can be represented via analytical models. Discrete event simulation (DES) is normally used for large-scale network simulation which provides more realistic and accurate results than the analysis method [1]. However, times and computing power are needed for simulation of large and complex networks. Then, the Hybrid method which combines both analysis and simulation can be applied for time saving. Test-bed emulation is an implementation of real-world hardware but on a smaller scale. The benchmark estimation is needed to verify and validate with the real situation. Due to the cost and difficulties of hardware implementation, test-bed will not be suitable for large-scale network simulation. There are many widely used network simulators such as discrete event simulator OMNet++, Network Simulator (NS), and Optimized Network Engineering Tools (OPNET). The OPNET is a discrete event simulator that supports analytical simulation, hybrid simulation, 32 bit and 64 bit fully parallel simulation and other features [2]. It is extensively used by many researchers due to its user-friendly graphic user interface (GUI) and its comprehensive development environment for modeling and performance analysis. For that reason, this thesis applies the OPNET Modeler v14.5 and the contributed model library of SIP-IMS [3]. The model is modified to simulate reliability and performance parameters of the studied IMS-based network scenarios.

5.1.1 Simulation of the IMS Network

The basic IP multimedia subsystem (IMS) setup model is built at a beginning state by using the basic IMS communication between two IMS users as shown per figure 5.1. The two users connect to the IMS core via layer four Ethernet switch. The application configuration node is where the source of traffic or different applications can be configured such as web, e-mail, or video, etc. The Voice over IP service with popular encoding scheme (G.729A) is chosen for an interactive service between two users by using session initiation protocol (SIP) as a signaling. The profile configuration node is the group of applications that were created by the application node which can be chosen and assigned by each object node such as user end device (UE)1, UE2, proxy-call state control function (P-CSCF), serving-call state control function (S-CSCF), and interrogating-call state control function (I-CSCF).

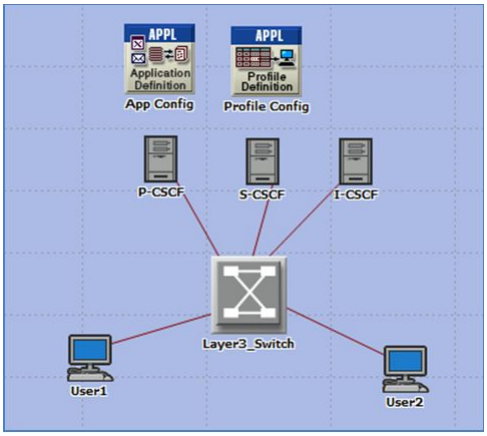


Figure 5.1: The basic SIP simulation of the IMS core network using OPNET

By setting the simulation time equal to fifty minutes and randomly unlimited repeatability call until the end of simulation time, the simulation results are shown by figure 5.2. There are seventeen calls had been created with different calling periods. Please be noted that the simulation time can be varied depending on how many sample data do we need. However, the simulation time is needed to be long enough for the system to generate at least one call. The Global statistic of packet end-to-end delay is monitored where the statistic congregates information of the whole network instead of an individual unit. The results showed that packet End-to-End delay for this case is ≈ 0.0652 seconds.

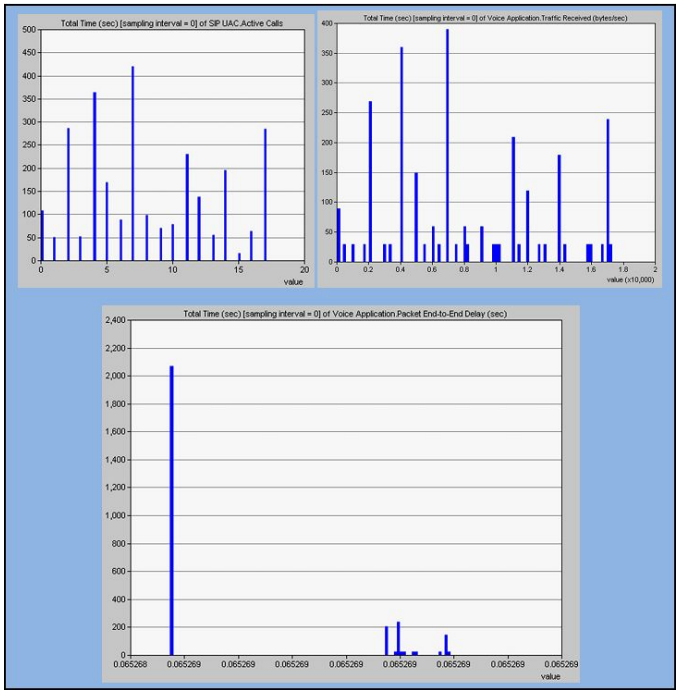


Figure 5.2: The basic SIP simulation results of the IMS core network using OPNET.

5.1.2 Simulation of the IMS Communication Scenarios

The proposed IMS scenarios per figures 5.3,5.4, and 5.5 which represent the communication between two UEs on similar and different visited network and domains are created. Moreover, the 1:1 redundancy of S-CSCF unit is also applied for comparison and verification with previous theoretical Markov model analysis. It is impossible to exactly evaluate or assign the failure behaviors into the simulator. Besides, all links and nodes are assumed or needed to be up during operation or simulation period. Accordingly, equally load balancing of S-CSCF unit is configured by using configured weight in order to see the effect of redundancy when the network traffic is changed. The OPNET modeler supports a variety of network parameters for performance and reliability evaluation. The Table 5.1 shows performance requirement and reliability related to different services [4]. Two keys network parameters; delay and jitter, represent that if either one of them influenced the system, its reliability will be directly affected. The VoIP application is chosen with one and a half hour of simulation time. Accordingly, our different IMS networks setup are tested against delay and jitter. However, to completely observe all dimensions of voice over

IP (VoIP) service, other statistical parameters such as packet delay variation and MOS (Mean opinion score) are also monitored and analyzed. All main attributes configuration remains the same for comparisons propose except for the number of S-CSCF in the case of redundancy.

Table 5.1: Performance requirement for different type of services [4].

Types of Service	Content				Sensitivity			
	Texts	Audio	Video	Image	Delay	Jitter	Bandwidth	reliability
Email	Yes	Yes	Yes	No	Low	Low	Low	High
LMS	Yes	Yes	Yes	No	High	High	High	Low
Video Conferencing	Yes	Yes	Yes	Yes	High	High	High	Low
Online Discussion	Yes	Yes	No	No	Low	Low	Low	High
Virtual Labs	Yes	Yes	Yes	Yes	High	High	High	Low
WAP	Yes	No	No	No	Low	Low	Low	High
VoIP(IP Telephone)	No	Yes	No	No	High	High	Low	Low
Virtual Classroom	Yes	Yes	Yes	Yes	Low	High	High	Low

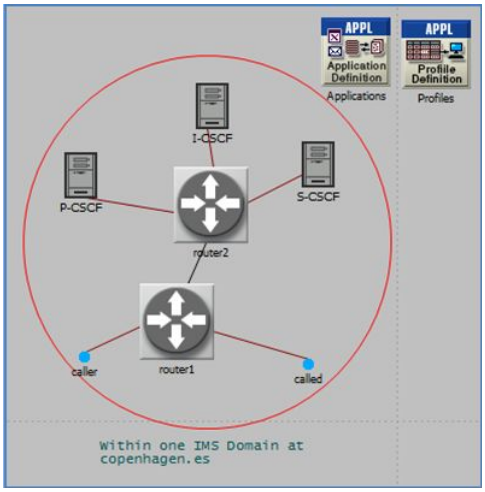


Figure 5.3: The two users are located inside their registered home domain

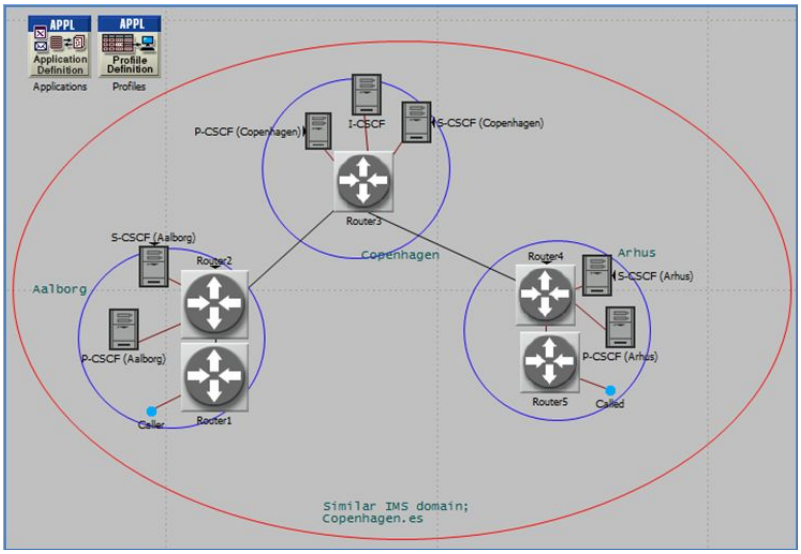


Figure 5.4: The two users are located at different visited network (Aalborg and Arhus)

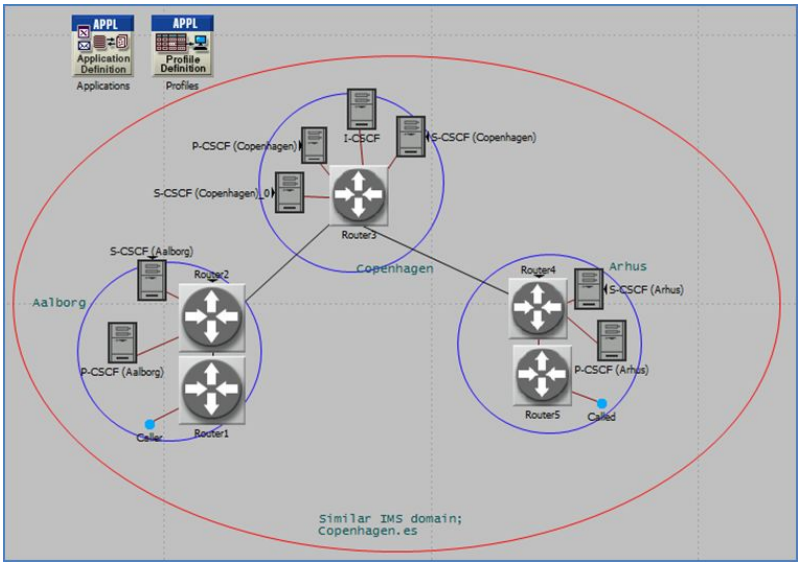


Figure 5.5: The IMS communication scenarios where two users are located at different visited network with redundancy at their registered home domain

End-to-end delay

The packet end-to-end delay represents the total voice packet delay which includes network delay, encoding delay, decoding delay, compression and decompression delay. The network delay is the time at which caller start sending the packet to real-time transport protocol (RTP) to the receiver received. The encoding delay is calculated from the encoding scheme. Decoding delay is assumed to be equal to encoding delay. Compression and decompression delays calculate from the attributes assigned by voice application configuration. The measurement of the variance among end-to-end delay for voice packets which is known as Packet delay variation is also simulated.

Jitter

The Jitter represents the variation in time of the transmitted and received packet due to congestion, bad queuing or error configuration. Therefore, instead of having continuously and constantly packet stream, variation or delay between the packet can occur. The jitter can be calculated by checking two consecutive packets leaving the source node with time stamps $t1$ and $t2$ and reach the destination node at time $t3$ and $t4$. Then, jitter equal to $(t4-t3)-(t2-t1)$.

Mean Opinion Score

Mean Opinion Score (MOS) is a numerical method of representing a perceived voice and video quality from the communication system. The MOS can be calculated by parameters like delay, Jitter, packet loss, and also from the inputs of user ratings. The number is ranged from one to five. Five and one represents the best and worst perceived quality respectively. The VoIP quality can be affected due to many factors such as Bandwidth, Codec scheme, Hardware, Jitter, Latency and Packet Loss. The VoIP calls often are in the range of 3 to 4.2. The table 5.2 below shows the sample MOS guide related to the voice quality [5]. The values need not to be an integer; it can be a decimal value from MOS spectrum such as 4.1 which is a maximum score of VoIP with G.711 codec (integrated services for digital network (ISDN), data rate=64 Kbit/s) and referred to a good quality. The compressed codec (8 Kbit/s) VoIP or G.729 A codec will be used for our simulation which can provide maximum MOS score at ≈ 3.7.

Table 5.2: The MOS and Voice quality [5]

MOS	Quality level	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

For the simulation, the voice quality is simulated and plotted by using the global MOS score. The score is the estimated mean opinion score for all of the demands in the network model. The call quality is simulated and displayed on the scale from one to five as described above. Correspondingly to the MOS, the R factor score also represents the effect of impairments of the voice signal. The score is evaluated from various VoIP metrics, including latency, jitter, and loss. The factor uses a scale from zero to one hundred, where zero and a hundred represent the lowest and highest voice quality respectively. The R factor is given by [6].

$$R = R_0 - I_s - I_d - I_e + A \tag{5.1}$$

Where R_o= Signal to noise ratio, I_s is the simultaneous impairments to voice signal transmission, I_d is the impairments delayed after transmission, I_e is the impairments due to codes and network equipment, A is called the advantage factor or the factor that attempts to account for caller expectations. The conversion of the voice quality factors; from R to MOS or so-called the MOS Conversational Quality Estimate (MOSCQE), and is given by equation 5.2.

$$MOS = 1 + 0.035R + R(R-60)(100-R)7 \times 10^{-6};\ for\, 0 < R < 100 \tag{5.2}$$

Please be noted that equation 5.1 and 5.2 are given by the E-model algorithm and recommended by the international telecommunication union (ITU)-T. It is not the actual customer opinion prediction and is currently under study. However, the factors can be used as additional performance factor or together with other specific performance parameters of the studied network to guarantee the user satisfaction [6].

5.2 IMS Communication Across Similar Domain

The figure 5.3, 5.3 and 5.5 show the communication of two SIP clients. These clients have the similar registered home domain which is located at Copenhagen area. The communication scenarios are divided into three topologies; the users

are located within their registered domain, they are located at visited domains, and they are located at visited domains with a redundancy of the S-CSCF unit at their registered domain. Figure 5.3 shows the scenario where two users located at their home domain area. Figure 5.4 scenario is the communication when both users located at different visited network areas which is similar to the communication scenario per figure 5.5 except that a redundancy of S-CSCF is added at the registered home domain. The internet protocal (IP) gateways (router) are used to connect between sip servers and clients. The Routing Information Protocol (RIP) is applied to dynamically and automatically create the gateway routing table and select routes; routing is based on a first-come-first-serve basis.

5.2.1 End-to-end delay

The packet end-to-end delays for all communication scenarios are shown by figure 5.6. From the figure, it look like all scenarios have the same end-to-end delay which is ≈ 0.06 second and remain constant along the simulation period. However, per figure 5.7, by observing in terms of probability mass function, the calls from visited domain have a higher end-to-end delay than calling from within their home domain around $\frac{(0.060326-0.060146)}{0.060146\times 100} \approx 0.30\%$. Besides, the delay values of calling from the visited home domain are varied around some range of delay values. On the contrary, calling within home domain, most of the calling delay values remain constant. Moreover, the histogram of time distribution plot of the results shown by figure 5.8 also support the steady trend of delay value for calling within their home domain case regardless of calling duration. Small variation of delay values can also observed from calling from visited domain and with redundancy case. From above, the results of calling inside home domain delay can be distinguished from calling from the visited domain cases by considering PMF or the histogram of the simulation results. Moreover, the delay of both visited domain cases (with and without redundancy) spreading at the same region and is difficult to compare. Therefore, differentiation of the packet end-to-end delay is plotted and shown per figure 5.9. Both visited domain cases have higher delay variation than calling within home domain case, which is almost steady. In addition, the visited domain case with redundancy shows smaller and less delay variation than the without redundancy case.

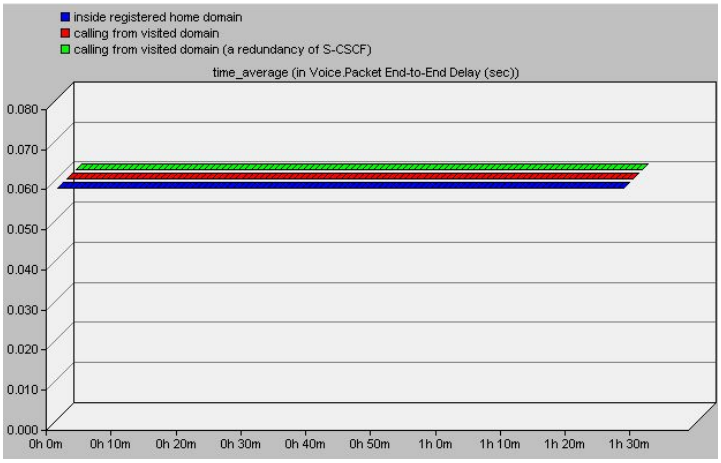


Figure 5.6: Packet end-to-end delay of the IMS users at different locations (Time average).

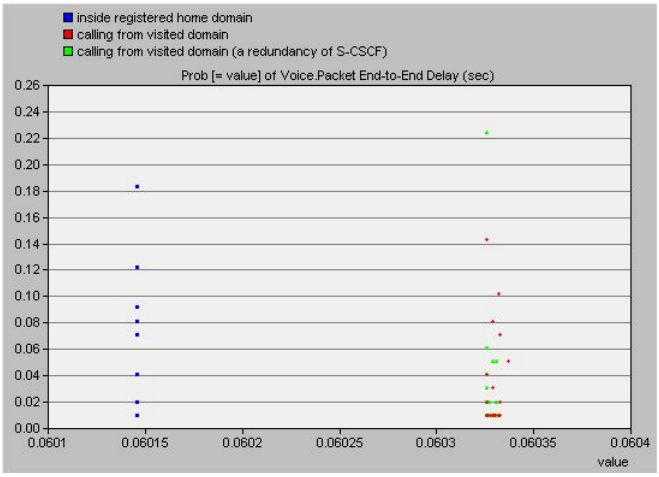


Figure 5.7: Packet end-to-end delay of the IMS users at different locations (Probability Mass Function, PMF)

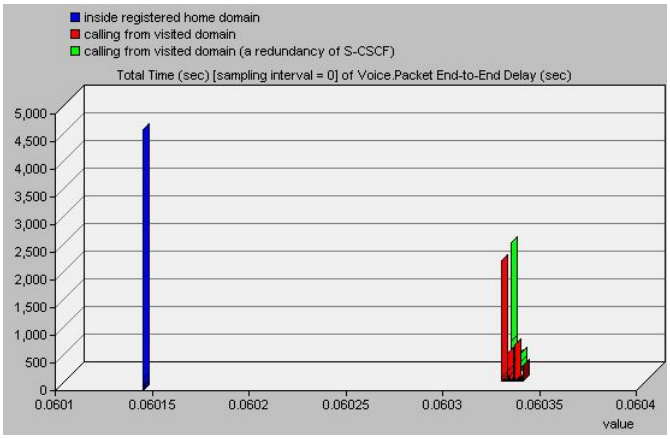


Figure 5.8: Packet end-to-end delay of the IMS users at different locations (Histogram, time distribution).

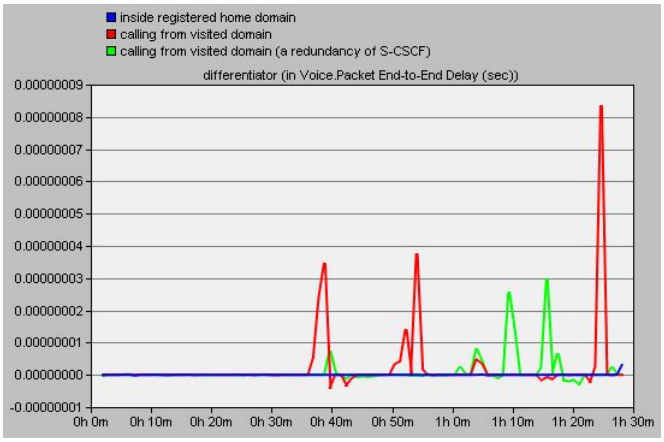


Figure 5.9: Packet end-to-end delay of the IMS users at different locations(differentiator plot)

Figure 5.10 represent the packet delay variation results. The variance of packet end-to-end delay can clearly represent the variation of the delay. The distribution of the delay can be observed for all scenarios. Then again, constant delay can be easily observed from calling inside home domain case. However, unsteady delay variation is observed for both visited domain cases. The fluctuation of the delays can be easily observed for both visited network cases where the without redundancy case shows much higher variation than the redundancy case in the long operation period.

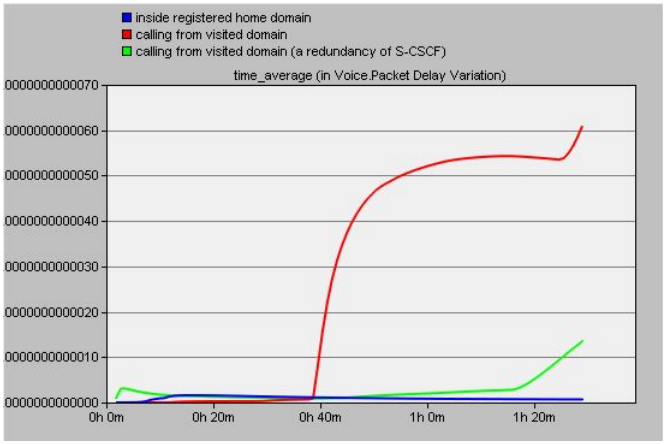


Figure 5.10: Packet delay variation (time average) at different locations

5.2.2 Jitter

Figure 5.11 shows a clear picture of stability of the calling packets. Unlike both visited network cases, the calling inside home domain case shows constant delay variation. The variation of the visited network without redundancy shows higher variation than the redundancy case. Another interesting point is that jitter can represent much more detail of delay variation characteristic at some specific point in time than the packet delay variation plot. Furthermore, the jitter of redundancy case is lower than without redundancy.

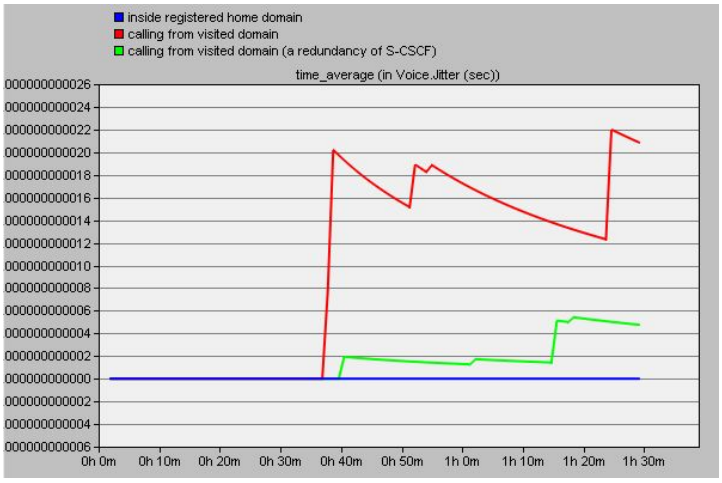


Figure 5.11: The Jitter (time average) of the IMS users at different locations.

5.2.3 MOS

Figure 5.12 shows that all scenarios seem to have the same range of MOS along the simulation period. The histogram plot of the same results is shown by figure 5.13. Most of the calls of calling inside home domain scenario (in dark blue) show a bit higher MOS score than from both visited network scenarios. However, some calls of the visited network scenarios experienced the same score with the calling within home domain scenario. The data can be observed by the green color pattern, which show the overlap colors of blue and yellow. Moreover, both visited domain cases present similar MOS score for most of the calls regardless of calling duration or redundancy.

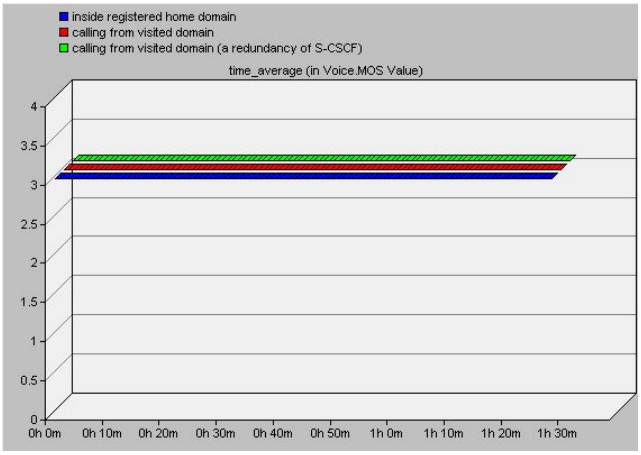


Figure 5.12: The MOS (time average) of the IMS users at different locations

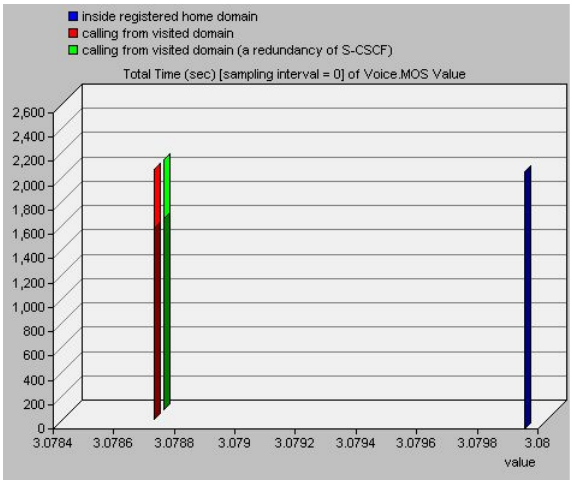


Figure 5.13: The MOS (histogram, sampling on time interval) of the IMS users at different locations

5.3 IMS Comunication Across Different Registered Domain

Figure 5.14: The IMS setup scenarios of communication between IMS users across different home domains

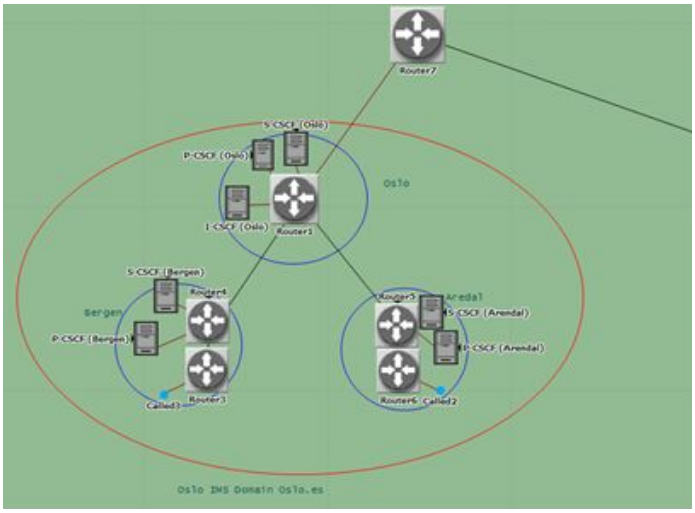


Figure 5.15: The IMS setup scenarios of the Network 2

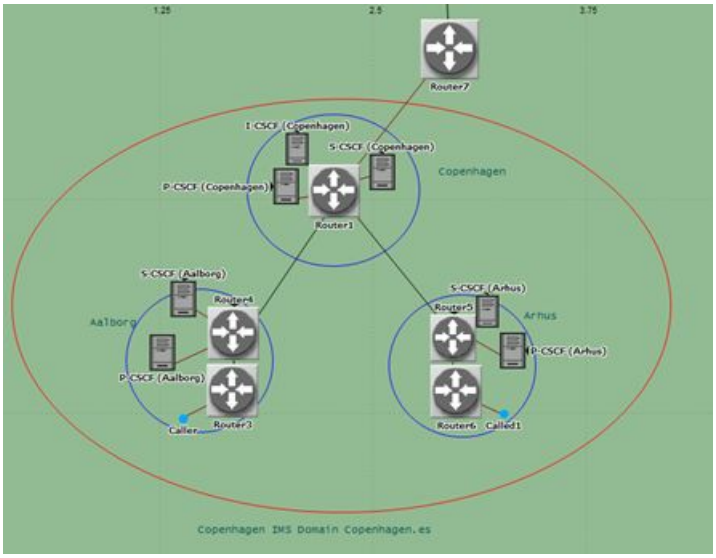


Figure 5.16: The IMS setup scenarios of the Network 1

The scenario per figure 5.14,5.15, and 5.16 represent the proposed analysis topology of the IMS setup communication network. This due to it can analyze overall reliability of the network at different calling locations. Besides, the effect of communication between users that have different registered domains and also similar registered home domains located at different locations can

be studied. In the same way, the keys reliability parameters such as delay, packet delay variation, and jitter are tested and simulated. The 1:1 redundancy scenario cases of S-CSCF are also created at each registered home domain location for both sub networks. As mention earlier, it is impossible to predict or assign exact failure events to the nodes and links. Besides, all nodes and links are assumed to be up during simulation period. Therefore, equally load balancing is also configured in case of 1:1 redundancy of S-CSCF for observing the effective results of the redundancy. The two subnets represent two IMS networks with different registered home domains at Copenhagen and Oslo. The communication across home domain or between subnets is routed through an IP cloud via gateway routers. As seen from figure 5.15 and 5.16, the windows with red lines represent the local network topology inside each subnet from the top view of the network hierarchy. The duplex point to point link, which can support up to 44.736 Mbps, is used to connect two subnets with an IP cloud. The same VoIP application and attributes configuration with communication of similar home domain case are assigned for comparison purpose.

5.3.1 End-to-end delay

As shown by figure 5.17, the packet end-to-end delay of calling across different home domains (caller to called2 or called3, in dark blue) is quite high (≈ 0.56 second) comparing with the delay of communication of similar home domain case (≈ 0.06 second). The average delay (in dark blue) is constant for the first twenty-four minutes. Then, it is slowly decreased and remained constant again at fifteen minutes before ending of the simulation. The same calling scenarios with one redundancy of the S-CSCF unit (in red) clearly demonstrate a smaller level of end-to-end delay. At the first ten minutes of the simulation period, rapidly decreasing of the delay is observed. Then the delay is slightly varied and almost constant at ≈ 0.35 second until the end of the simulation time. After the first five minutes, the delay is decreased from no redundancy (in dark blue) for almost 42 % when adding just one redundancy of the S-CSCF unit to the registered IMS home domain. Moreover, by looking at the steady state liked period or one hour of the simulation period, ≈ 18 % delay decreasing is observed.

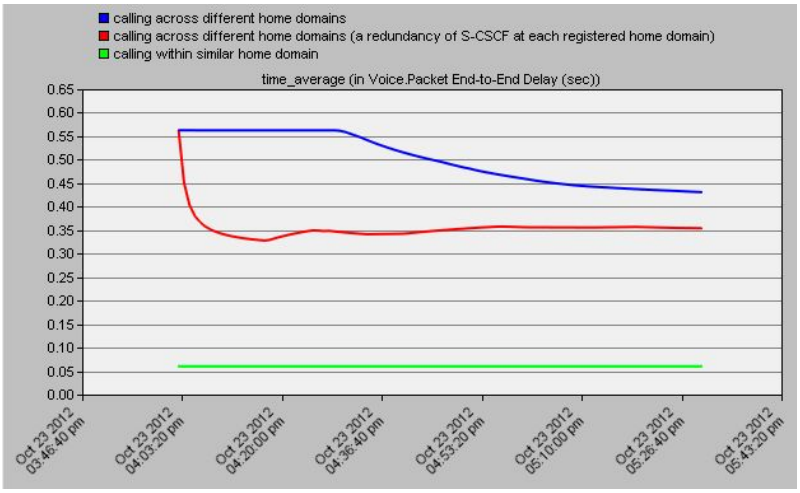


Figure 5.17: The IMS setup scenarios of the Network 1

The results of calling within the similar home domain (caller to called1, in yellow) have a very small delay value (≈ 0.06 second) and remained constant along the simulation period. When comparing with calling across home domain cases, at fifteen minutes of the simulation time, 832.5% and 443.6% of increasing delay are observed for without and with redundancy cases respectively. In the same way, by looking at the steady state liked region or after one hour of the simulation time, almost 637% and 490% of increasing delay still can be observed. The figure 5.18 shows the packet delay variation, the less delay variation is from calling within their registered home domain case (in the dark blue). The interesting results of calling across registered home domain shows that the redundancy case has highest delay variation than other cases. Besides, the delay variation of calling across the home domain (without redundancy) is almost at the same level with calling within home domain case.

5.3.2 Jitter

Figure 5.19 shows the jitter of three calling scenarios. Similarly to the packet delay variation results, the redundancy case of calling across different home domain has higher jitter value at some period. The similar trend of jitter results is observed between calling across the different home domain (without redundancy) and calling within home domain case. Even though there are some peaks of jitter values for the redundancy case, the rapid exponential decay is observed after the peak. That means just only a point in time that a few delay packets may cause high jitter value. At the long run, the jitter trend is reduced

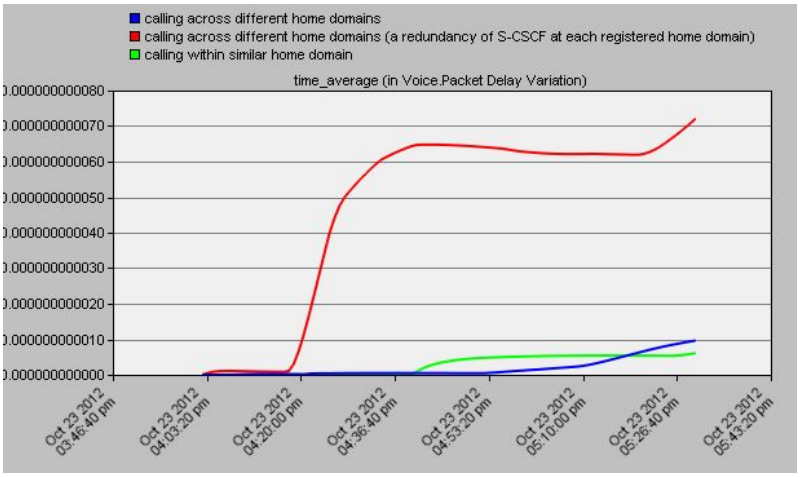


Figure 5.18: The Packet delay variation of different IMS setup scenarios (communication within and across communication domains)

and quite close to the calling without redundancy case. The constant period of jitter value is mainly observed from calling within home domain case.

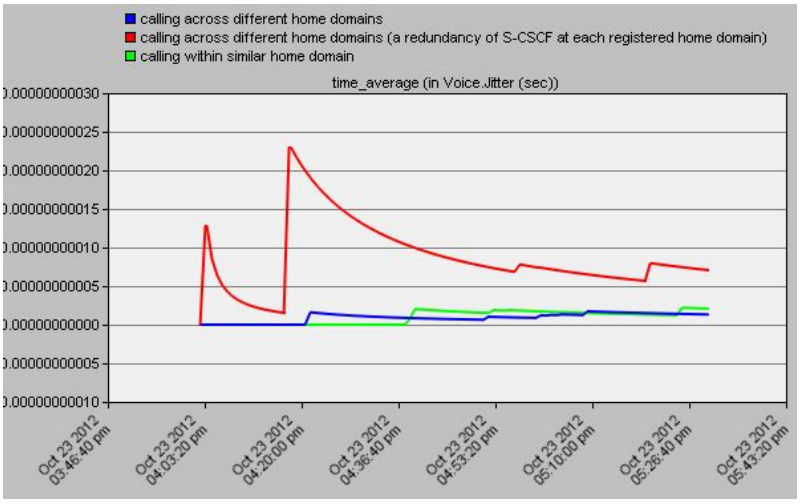


Figure 5.19: Jitter of different IMS setup scenarios (communication within and across communication domains)

5.3.3 MOS

By looking at the MOS score per figure 5.20, calling within the similar home domain has constant MOS score which is ≈ 3.1. Another interesting MOS scores are observed from calling across different home domain cases. The MOS of redundancy case is rapidly increased and reached the equivalent level of calling within home domain after fifteen minutes of the simulation period. However, without redundancy, the lowest MOS score (MOS=1) is observed for almost twenty-five minutes of the simulation period. Then, it is slowly increased and the results are likely to reach and stable at MOS ≈ 2.4.

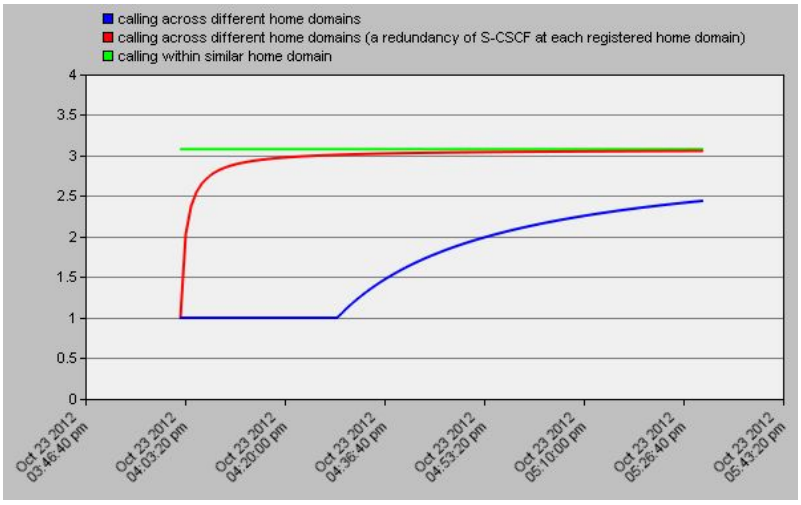


Figure 5.20: The MOS of different IMS setup scenarios (communication within and across communication domains)

5.4 The Effect of Network Parameters When the Traffic is Increased

In order to observe the behaviors of the network parameters when communication traffic is increased, three more callers are added into the communication scenarios. Each caller can generate repeatedly and randomly a calling traffic approximately 24 Kbps, which is totally 94 Kbps for four users. The simulation for both communication within similar IMS home domain (calling from visited domain case) and across IMS communication home domain scenarios are examined and compared. Moreover, the effect when having a 1:1 redundancy of S-CSCF is also simulated for both scenarios.

5.4.1 Calling from visited network (communication of users with similar registered home domain) in case of four callers

The network topology per figure 5.14,5.15 and 5.5 are evaluated with three additional users (totally four callers). Similarly to the previous simulations, three main reliability network parameters: Delay, Jitter, and MOS are examined and compared.

End-to-end delay

As shown by the figure 5.6, the previous results of the packet end-to-end delay of one caller cases, the results of calling from visited domain and with additional redundancy are similar and equal to ≈ 0.06 second. However, for the results of the four callers cases, per 5.21, the delay of no redundancy case (in red) has increased to be $\approx 33.3\%$ at the beginning of simulation period and is slowly reduced to be $\approx 12.8\%$ and remain steady near the end of the simulation period. For the redundancy case (in light blue), unlike one caller case, the delay is decreased and dropped and lower than the without redundancy case for $\approx 10\%$ and $\approx 3.7\%$ at the beginning and near the end of the simulation period.

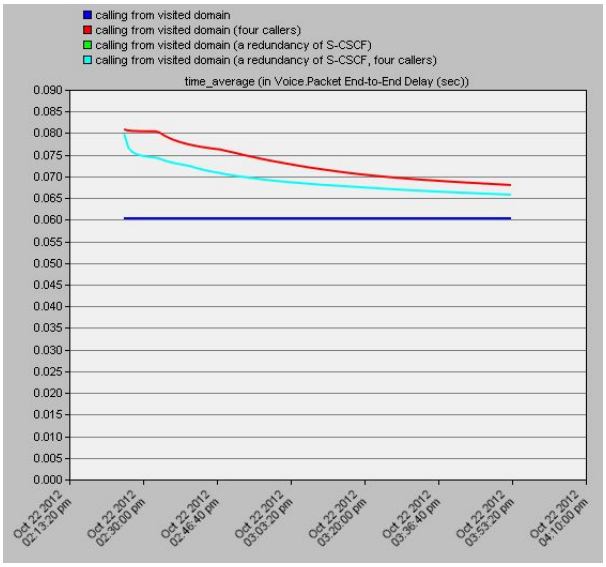


Figure 5.21: Packet end-to-end delays of different IMS setup scenarios (similar home domains)

Packet delay variation

The packet delay variation results of calling within similar domain scenarios are presented by figure 5.22. The delay variation of four caller cases has a bit higher level than one caller cases. For the one caller case with a redundancy of S-CSCF (in yellow), the results have a lower level of delay variation than without the redundancy case. However, the delay variation of four callers with redundancy (in light blue) presents a higher level than without redundancy case (in red).

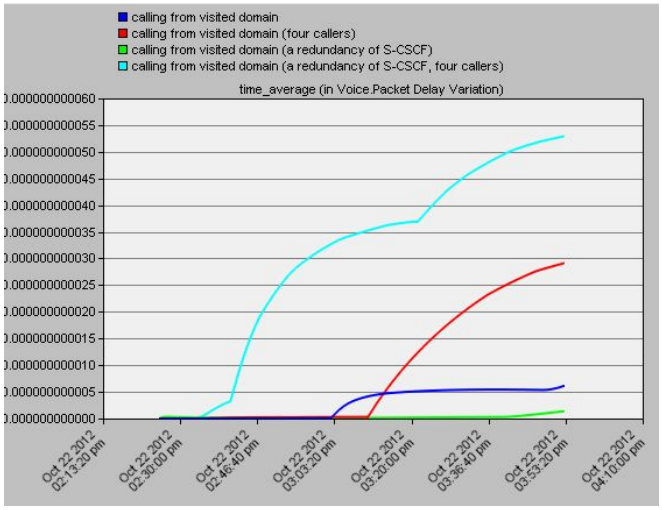


Figure 5.22: Packet delay variation of different IMS setup scenarios (similar home domains)

Jitter

From the results of figure 5.23, the jitter of one caller cases with redundancy (in yellow) seems to have less and lower jitter comparing with no redundancy case (in dark blue). However, the four callers with no redundancy case (in red) does not show the significant difference of jitter comparing with the one caller case (dark blue) except that its jitter is observed quicker than one caller case. The jitter of no redundancy case can be observed at the same time with the four callers with redundancy case (light blue). Besides, the frequency and jitter level of the four callers case is highest among other scenarios. The rapid changing of the jitter level also means that only a few packages experienced a delay. Moreover, the levels of jitter for all cases tend to be reduced and reached the same level near the end of the simulation period.

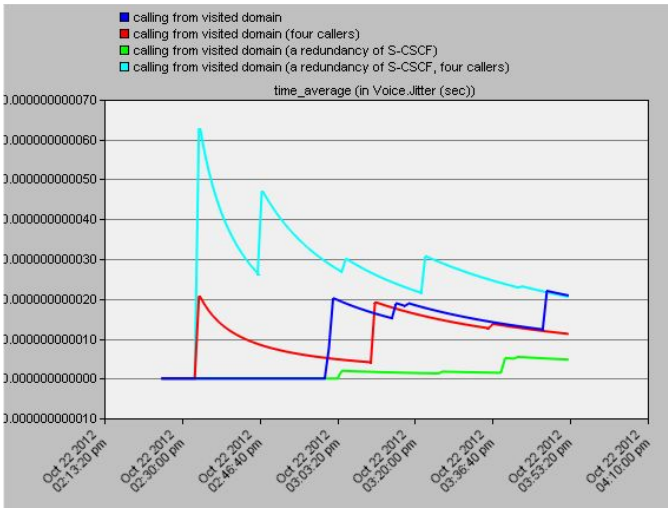


Figure 5.23: Jitter of different IMS setup scenarios (similar home domains).

MOS

When having four callers for both communication scenarios when having redundancy and without redundancy provides almost similar MOS score to the one caller cases and is ≈ 3.07. The three-dimensional plot of the MOS results is given by figure 5.24.

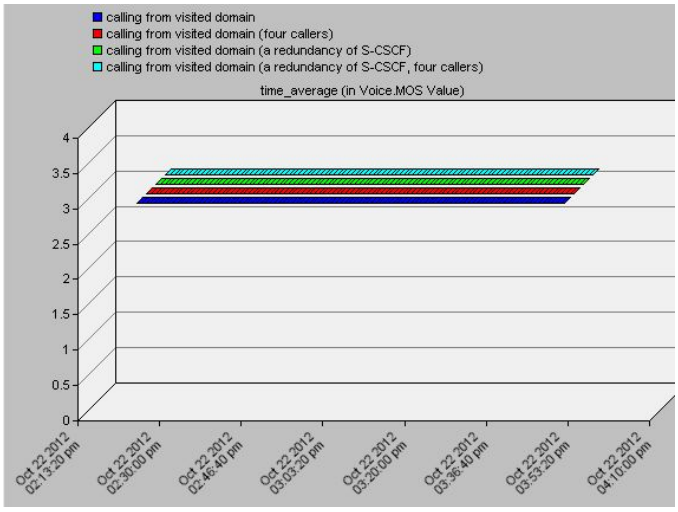


Figure 5.24: The MOS of different IMS setup scenarios (similar home domains)

5.4.2 Calling across different home domain in case of four callers

The network topology of calling across different home domain per figure 5.14 and 5.15 will be simulated with three additional users (totally four callers). Similarly to the previous simulations, three main reliability network parameters: Delay, Jitter, and the MOS will be examined and compared.

End-to-end delay

From results of figure 5.25, one caller case, the packet end-to-end delay when having redundancy (in yellow) is significantly reduced comparing with no redundancy case (in dark blue). The difference percentage of the delay at steady state is $\approx 18.60\%$. For the four callers without redundancy case (in red), the observed delay is lower than one caller case 0.05 sec at the beginning of the simulation period, then the average delay is increased and higher than the one caller case (dark blue) until the end of the simulation period. At the so-called steady state like region of the delay characteristics, the delay of one caller case (in dark blue) is $\approx 6.89\%$ higher than the four caller case. Moreover, for the four callers case with redundancy, the average delay (in light blue) is rapidly reduced, when comparing with the without redundancy case, at the beginning of the operating period and remain stable. The difference percentage is $\approx 15.21\%$.

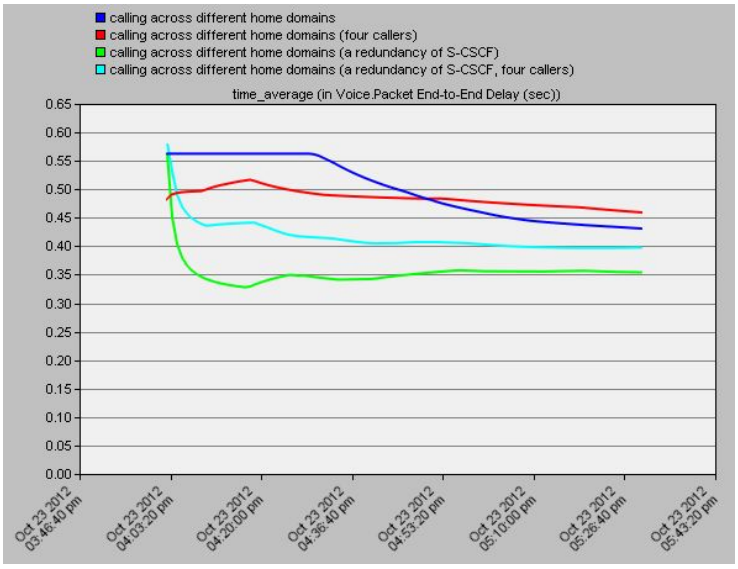


Figure 5.25: Delay of different IMS setup scenarios (across home domains)

Packet delay variation

For the four callers for both redundancy and without redundancy cases, the packet delay variation are less than the one caller cases as present by figure 5.26. Moreover, for the redundancy cases, the delay variation results (in light blue and yellow) are higher than the without redundancy cases. However, for one caller case, with redundancy (in yellow) experienced a lower delay variation than without redundancy case. Besides, the delay variation of the four caller cases is a bit higher than one caller cases.

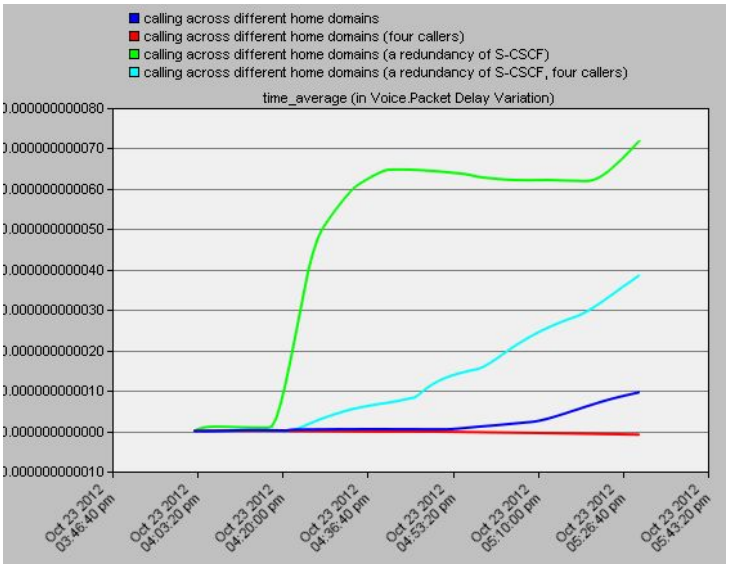


Figure 5.26: Packet delay variation of different IMS setup scenarios (across home domains)

Jitter

For both one caller and four callers cases, with redundancy cases showed to have higher jitter than without redundancy cases. However, the trend of jitter level for all communication scenarios does not significantly different in term of the degree of value ($10^{-10} sec$) as shown by figure 5.27. Moreover, at the steady state liked region, the jitter level of all scenarios is very close to each other.

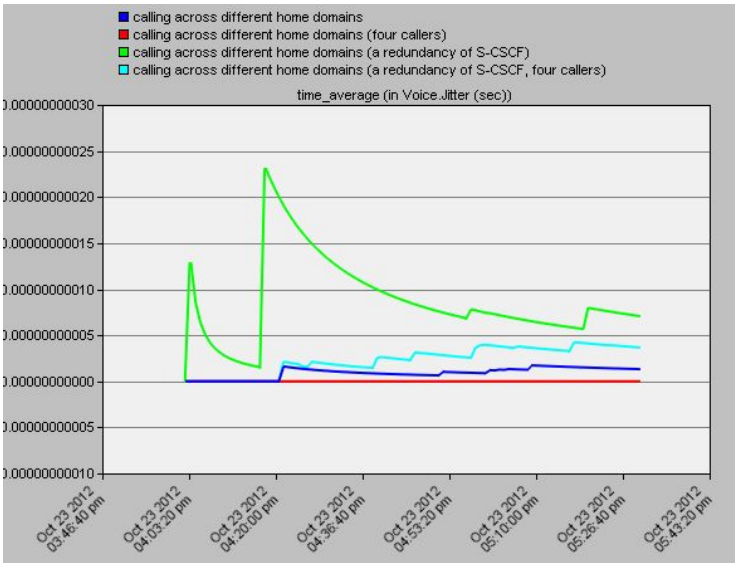


Figure 5.27: Jitter of different IMS setup scenarios (across home domains)

MOS

The MOS results of figure showed that the MOS score for both one caller and four caller cases are increased when having redundancy into the system. Moreover, the MOS score of the four callers with no redundancy case is higher than the one caller case. In particular, the increasing rate of the MOS score of the one caller case (in dark blue) is faster than the four caller case. Besides, with redundancy, the MOS score of both four callers and one caller cases is quickly raised at the beginning of the operating period and remain at ≈ 3.0 level until the end of the simulation period.

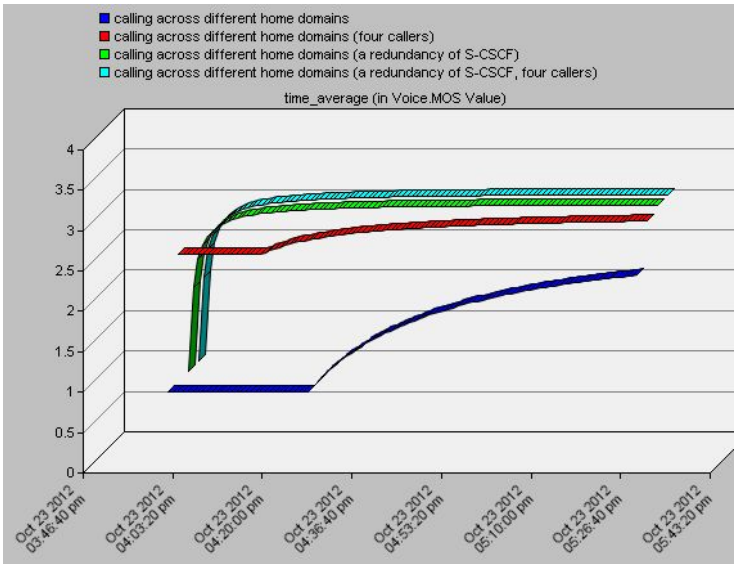


Figure 5.28: MOS of different IMS setup scenarios (across home domains)

5.5 Analysis and Conclusion of the Simulation Results

For the comparison propose, the plots of all calling scenarios (calling within and across home domain communication) are presented by the figure 5.29 and 5.30 The packet end-to-end delay and MOS are the two reliability parameters which were chosen due to these two parameters can significantly demonstrate resilience behaviors of the studied system.

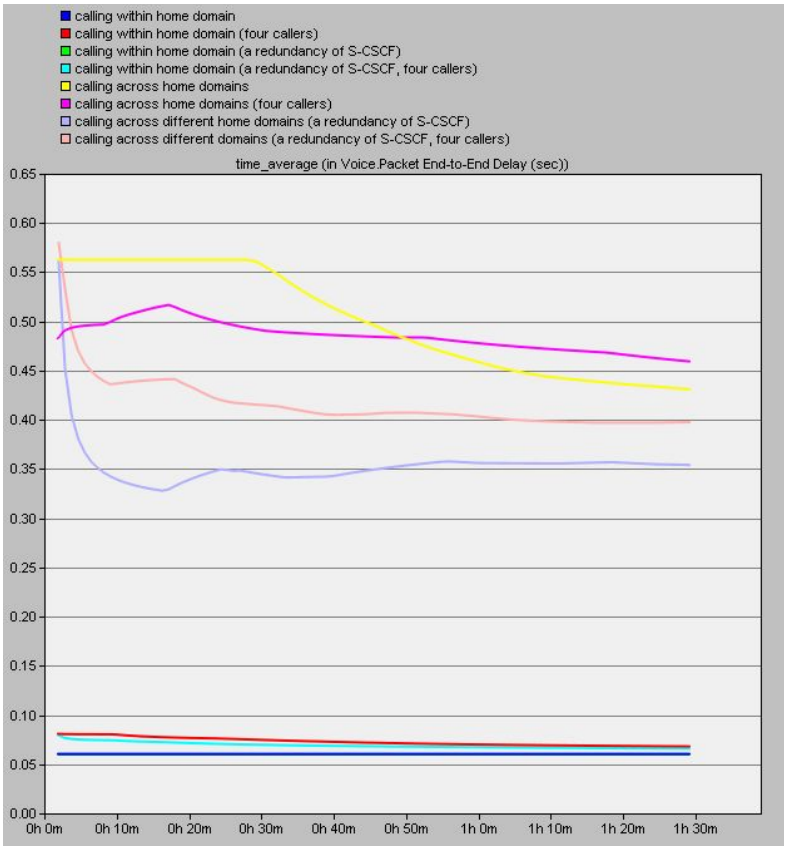


Figure 5.29: Packet end-to-end delays of different IMS scenarios (within and across home domains)

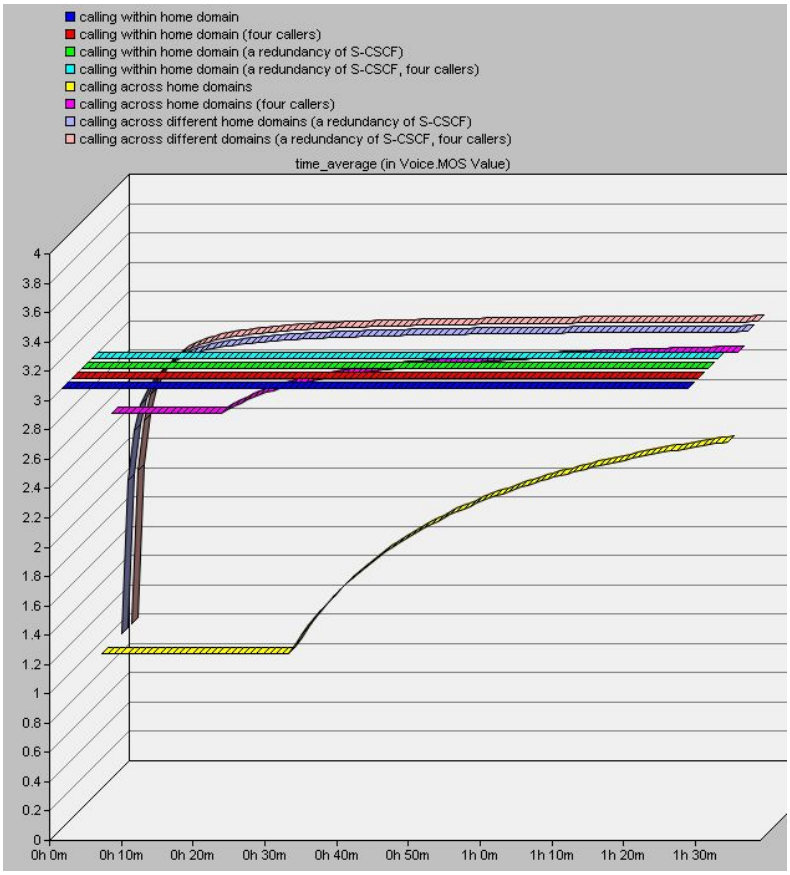


Figure 5.30: MOS of different IMS setup scenarios (within and across home domains)

5.5.1 Analysis of the calling within home domain cases

The packet end-to-end delay for all scenarios is quite stable along the simulation periods. And the average delay level is very close to each other. There is only $\approx 0.33\%$ of the difference percentage between calling inside the home domain and calling from the visited network cases. Then, different calling quality from these scenarios can be monitored from the packet delay variation and jitter plots results.The results also demonstrate the delay variation of the communication scenarios along the simulation period. For calling from visited home domain cases (with and without redundancy) will have higher delay variation than calling from within home domain cases. Moreover, with redundancy, less frequent and smaller delay variations are observed than without redundancy cases. In conclusion, communication within the similar registered

domain will not be affected by end-to-end packet delay. This includes communication between users that are located at the visited network areas. The similar conclusion can be drawn from the similar calling quality from the MOS plot. However, calling from the visited networks may experience frequent and high delay variation. In particular, both delay variation and jitter effects are quite low in terms of the degree of value. With redundancy of the S-CSCF, calling from the visited network can reduce the effect of delay variation.

5.5.2 Analysis of the calling within home domain cases of four callers

With additional callers, the packet end-to-end delay is increased for both with and without redundancy cases. Moreover, the average end-to-end delay will be significantly decreased when having redundancy. The results can be clearly observed with the one caller case. This due to the system performance or capacity that can support some certain level of traffic loads or voice traffics. Therefore, even with additional redundancy, the similar end-to-end delay is still observed. Moreover, four caller cases will experience high delay variations and jitter than one caller cases. Accordingly, four callers will generate more signaling traffic and delay. From the MOS results, the MOS score of all scenarios are quite stable and are at similar level regardless of the number of callers or redundancy. This implies that the overall voice quality is not affected by the number of callers. Therefore, calling within home domain cases, the voice quality is rarely affected even though end-to-end delay and variation of the signaling delay are increased.

5.5.3 Analysis of calling across home domain cases

The overall average end-to-end delay is quite high when comparing with calling within home domain scenarios. The increasing percentage for calling across home domains without and with redundancy are $\approx 1040\%$ and $\approx 580\%$ at the beginning simulation period and $\approx 684\%$ and $\approx 483\%$ near the end of simulation period respectively. Moreover, with redundancy, significant decreasing of the end-to-end delay can be clearly observed for all cases.The fluctuation of the delay level may be affected by the simulation traffic and configurations such as routing protocol (the RIP is assigned for this case). However, the average delay is of calling across home domain cases is much higher than calling within home domain cases. Besides, the interesting results of delay variation and jitter is found between calling across home domain with redundancy and without redundancy at some simulation period where redundancy case has higher delay variation and jitter than the without redundancy case. In conclusion, delay variation, and jitter may provide a detailed quality picture of the signal transmission at some specific simulation period. Obviously, the total calling quality

of the communication needs to consider based on overall results of the monitoring network parameters. Therefore, in this case, the average end-to-end delay and MOS should be used for representing an overall network reliability.

5.5.4 Analysis of calling across home domain cases of four callers.

Based on the results of 5.4.2, when the number of callers is increased or the calling traffic is increased, the average end-to-end delay is increased. Moreover, the average delay is clearly observed to be decreased when having redundancy for both one and four caller cases. For calling across home domain cases, the effects of a number of callers and redundancy are much higher when comparing with calling within home domain cases. Besides, the delay variation is quite sensitive and can not be solely predicted the overall characteristic of the network. Based on the MOS results, redundancy can significantly affect and improve reliability or voice quality of the system especially for calling across home domain cases or long distance communication. In conclusion, for calling within similar home domain case, the average end-to-end delay is quite stable even when increasing the signaling traffic, the average delay is a little bit increased from the beginning of the operating period. Moreover, with the redundancy of the S-CSCF unit, the average delay is decreased. Besides, increasing the traffics will increase delay variation of the system. The redundancy can also reduce the delay variation effect. For calling across home domain cases, the effect due to the average packet end-to-end delay is higher than calling within home domain cases: nine times or $\approx 450\%$ in difference percentage from the simulation results. Similar to the calling within home domain cases, the delay is increased when increasing the signaling traffics. Moreover, the percentage difference of reduction of the delay is much higher than calling within home domain cases. Therefore, redundancy or redundancy of the S-CSCF unit will highly affect long distance communication cases. Besides, similar to calling within home domain, the average delay variation behaviors are difficult to predict and quite high at some operating periods. Moreover, the signaling (voice) qualities based on the MOS score showed to improve with redundancy cases. This implies that redundancy is definitely needed for improving an overall quality of long distance communication system especially for calling across home domain cases. In brief, The simulation results agree well and support the previous analysis results of the transient and steady state availability and reliability of the network by using the proposed five state Markov model for both calling within and across different home domains. Accordingly, the proposed analysis model and the results are verified by the OPNET software which can simulate real IP network behaviors [5].

5.6 Conclusions

The end-to-end reliability features of different IMS-based communication scenarios were simulated and compared with various principal network parameters. Moreover, the overall reliability characteristics of the system were examined when increasing the network traffics and parallel redundancy of the key core IMS unit: S-CSCF unit. To conclude, from the simulation results, intra-domain communication will not likely be affected by end-to-end packet delay regardless of the user locations: calling from within the user registered home domain or the visited IMS network. On the contrary, inter-domain communication is profoundly affected by the delay especially when signaling traffics are increased. The overall reliability qualities are enhanced when adding a redundancy into the system. this implies that resilience of the system can be significantly improved by having redundancy at the important core IMS unit especially for inter-domain or long distance communication.

5.7 References

[1] Lawrence M Leemis and Stephen Keith Park. *Discrete-event simulation: A first course*. Pearson Prentice Hall Upper Saddle River, NJ, 2006.

[2] OPNET or Riverbed Modeler. Opnet or riverbed technologies inc, 2015.

[3] AH Enrique Vazquez and Jose Ignacio Fernandez. Sip-ims model for opnet modeler. *OPNET University Program Contributed Models*, 2005.

[4] James D McCabe. *Practical computer network analysis and design*. Morgan Kaufmann Publishers Inc., 1998.

[5] Mohd Nazri Ismail and Abdullah Mohd Zin. Network analyzer development: Independent data, opnet simulation tool and real network comparison. *Management and Tech-nology*, 1(1):97–105, 2010.

[6] Jan A Bergstra and CA Middelburg. Itu-t recommendation g. 107: The e-model, a computational model for use in transmission planning. 2003.

Chapter 6

A Novel Estimation Framework for Quality of Resilience

6.1 Introduction

Based on the results of chapter 2, chapter 4, and chapter 5, the reliability of the IMS-based network is influenced by many factors. In order to support the convergence of different access telecommunication technologies and providing end-to-end quality of service (QoS) and quality of resilience (QoR), a modern or practical end-to-end evaluation method is needed to satisfy modern services and user happiness. Therefore, this chapter provides significant limitations and possible solutions for modern reliability evaluation. A modern reliability evaluation is proposed by using Bayesian statistics and taking into account both objective and subjective network parameters. Additionally, the selective resilience measurement parameter algorithm is proposed for modern reliability evaluation method. The proposed evaluation framework in this chapter covers the system reliability evaluation challenges as stated in the QoR and challenges of this thesis (Section 1.3).

6.2 Related Works

Reliability of the system or service can be defined as an ability of the system or service that provide its intended function for a given period and environment [1]. The reliability defines a successful operation of the unit or system without a failure during the operating period. For reliability evaluation, availability may be used or evaluated interchangeably with reliability. Instead, availability

defines the readiness for operation of the unit or system. Nevertheless, if there are no recovery schemes, availability is similar to reliability as any failures leading to unavailability of the unit or system. Hence, the failure is the key parameter for reliability evaluation, especially the failure rate parameters.

The Bathtub curve is normally used to represent the lifetime or failure rate (FR) of the unit as shown by figure 6.1. It is widely used to present physical failure behaviors of the unit. Even though, different units have different failure characteristics, a system or interconnection of these units may also have a different failure features, the failure rate can be represented as a generalized form by using the bathtub curve as represented by figure 6.2.

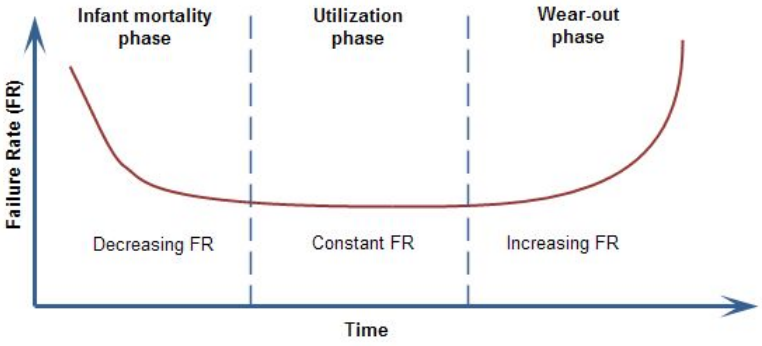


Figure 6.1: The typical FR or Bathtub curve of the unit versus operation time

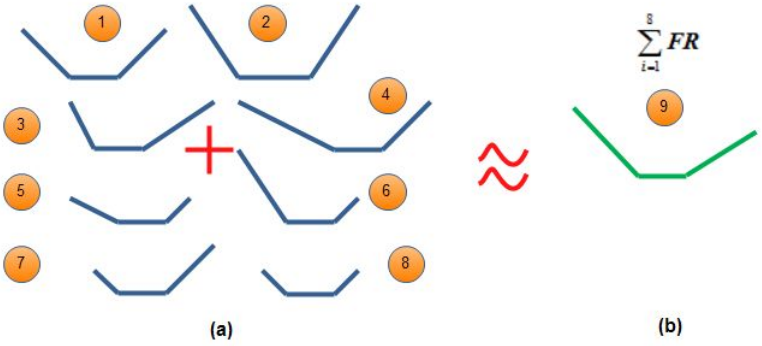


Figure 6.2: Representation of possible FR characteristic of a system of eight units: (a) the FR characteristic of each unit, (b) the FR of the system (combination of each unit)

In particular, most of the traditional evaluation methods focus only on the constant FR (utilization phase); thus, the reliability parameters are only evaluated with a constant failure rate assumption. The whole variation of FR characteristics would better represent the reliability of modern systems and services. For instance, a sensitive system or service (medical, energy or financial systems) is strongly influenced by the reliability than ordinary systems. In other words, the system needs to maintain high reliability even at the beginning of the operating period. In addition, a complex system such as a transportation or communication system needs to maintain reliability even after a long operating period due to it is impossible or extremely costly to replace the whole system for every failure. In addition, there is no rule or standard for what parameters should be monitored at what operating periods. For instance, the mean time to failure (MTTF) or availability of the system is used as measurement parameters to quantify reliability or performance of the system or vice versa. Therefore, to deliver a complexity of the traditional quality and reliability terms and evaluation methods, the related works are categorized into two main following topics: quality of experience (QoE) and user satisfaction.

6.2.1 XoX and QoE

Various methods have been proposed and applied for reliability evaluation of the system [2]. These techniques can be categorized into two main methods: qualitative and quantitative methods [3]. The quantitative methods provide the evaluation results in term of a countable or measurable quantity or the quantity that can be precisely defined. In contrast, the qualitative method provides the results in a descriptive term concerning feelings, or experiences.

Recently, there are many research works on how service quality terms affect users via multiple quality terms: Quality of Service (QoS), Grad of Service (GoS), Class of Service (CoS), Quality of Resilience (QoR), and Quality of Experience (QoE) [4, 5].

The main objective of these terms is to define the level of service quality which will properly assist user demands. These quality terms can be referred to as "XoX" where the first acronym "X" represents the first syllable of those terms such as Quality, Class or Grade. The last acronym "X" represents the last syllable of the quality terms such as Service, Resilience, and Experience. The acronym "XoX" stands for these different quality terms. The terms have gained more recognition due to the complexity of various services, and their features were advertised as a marketing plan to a customer. As a result, the QoE has also gained more interest because it directly directs on the end-user satisfaction of the service. However, apart from traditional QoS terms, various definitions of the quality terms when applying these terms based on different network architectures can cause confusion to users or service providers [4]. Besides, the authors provide a non specific definition and introduced likeness

among these terms based on the service class definitions provided by different standardization:internet engineering task force (IETF), international telecommunication union (ITU), 3rd generation partnership project (3GPP), The European Telecommunications Standards Institute ETSI, and The Institute of Electrical and Electronics Engineers IEEE.

Despite the fact that these terms are recognized in terminology; in any case, they are identified with one another. The authors likewise demonstrated how QoE can be influenced by other quality terms and factors (environmental, psychological and sociological aspects, user profile, or application features). Therefore, it is quite a challenge to efficiently evaluate the QoE. Although, the Mean Opinion Score (mean opinion score (MOS)) is used as one of the popular QoE measurement methods for voice and video applications based on the ITU recommendations. However, difficulties and challenges still exist for evaluating end-to-end QoE. The QoE measurement for the converged network and services has been investigated [5, 6, 7]. The user satisfaction has been shown as the most significant quality evaluation measure. Moreover, the measurement of QoE with respect to the convergence requirements (any service, anywhere, anytime, any user device, any media and networking requirements) has been observed [5]. Moreover, challenges of provisioning QoE over the converged networks have been discussed [8]. The work focused on QoE estimation of media services over an internet protocal (IP) network with three main QoE evaluations: objective, subjective and hybrid approaches. The definition of subjective and objective can be considered as qualitative and quantitative respectively. Accordingly, the objective measurement involves only technical related parameters, which involve either network or application QoS parameters as presented in [9]. The subjective assessments essentially apply user opinion metrics, for instance, questionnaire or the MOS. If utilizing both objective and subjective metrics for an evaluation, the method is called the hybrid approach [10]. In this work, the authors classified two additional quality terms called Quality of Delivery (QoD) and Quality of Presentation (QoP) which might affect the QoE of voice and video services. The QoD is applied for delivery reliability while the QoP focuses on perceived quality of the media. Therefore, the measurement parameters of the QoP and QoD relate to the application layer and the layers below the application layer respectively.

However, providing QoE over the converged network is arranged in term of how various quality measurement parameters, network technology, and standardization can affect the quality terms that literally affect the QoE. Therefore, to efficiently achieve end-to-end QoE or QoS, cooperativeness across different converging networks and synchronization of quality parameters between different standards are needed. Moreover, the authors [4, 5] recommended that QoR assessment should be considered as an independent factor besides QoS. Nevertheless, the term and its assessment parameters were derived from the QoS. In addition, degradation of these parameters can directly affect the user

perceived quality as stated by [11]. The authors pointed that the frequency of service interruption of voice and video applications can significantly impact the QoE. Based on the above-related works and discussions, the QoE is proven to be affected by QoS and QoR. In other words, these parameters can be influenced by each other and can directly or indirectly affect the user perceived quality.

6.2.2 XoX and user satisfaction

The user satisfaction can be regarded as the main goal of implementing the previously mentioned service quality terms. Therefore, the quality evaluation has been proposed and created to user demands and pleasures. The developing of client interest and advancement is firmly joined with one another. A good example in response to this significant bond and preparing for successful global competition is the national innovation strategy of Finland that has recently arranged the user-driven innovation framework and policy [12]. Moreover, a user-centric approach to evaluating QoS for future networks has been covered by the ITU [13]. Therefore, there is doubtlessly user satisfaction has gotten to be one of the essential necessities for both service providers and clients. Consequently, the network parameters that might impact user satisfaction, directly and indirectly, are needed to be considered and carefully evaluated.

Based on Subsection 6.2.1, considering the goal and measurement parameters of each quality term, the complicated relationship between each term can be represented toward the user satisfaction. Therefore, the relationship between various quality terms and parameters that affect user satisfaction can be expressed by figure 6.3. The figure represents two main aspects that can affect user satisfaction: quality-related factors and user-related factors. The quality terms are typical technical quality terms which have been defined as a set of standards for ensuring a quality level of the networks or services. These terms include other quality terms that have been specifically defined by a group of researchers (QoR or QoD). The user related terms can be defined as a term or factor which directly or indirectly affect user emotions or opinion of the perceived services. The quality terms and user-related aspects are considered as objective and subjective quality assessment domains respectively.

The QoE is classified as the largest set beyond QoS because it includes both subjective (user-related factors) and objective factors. the user-related factors are uncertain and are classified separately. Estimation of the subjective factors is difficult. In particular, even when the same user or tested conditions are being set up for quality assessment, the tested results can vary over time. In another word, the same user may provide different results at different period by using the same estimation method. The services type which is considered as a controllable factor is somehow randomly selected based on the user aspired.

Therefore, user expectation of the service types is then considered as one of the user-related factors.

In contrast, the QoS parameters can be observed by using estimation tools: both network monitoring software and hardware. The quality can be statistically analyzed and expressed in short-term or long-term quality parameters for a critical analysis [14]. Accordingly, the equivalent measurement methods can be applied to the other quality terms. These terms can be viewed as a subset under QoS as most of them are derived or specifically classified from the typical QoS definition.

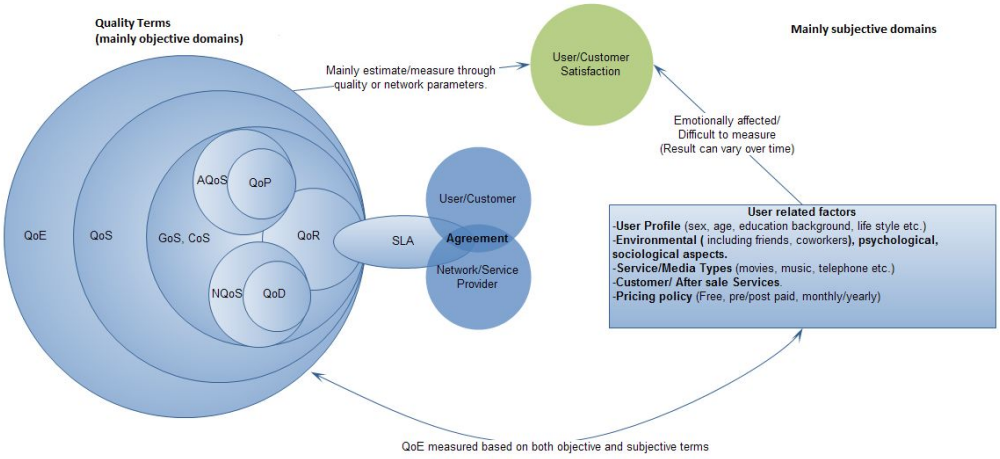


Figure 6.3: Association between quality terms and influenced factors of user satisfaction

To summarise, the QoE can be influenced by any quality terms. Therefore, evaluating QoE as well as user satisfaction of the service and system is quite challenges. Therefore, in this chapter, reliability is classified into two main terms based on the affected factors: subjective and objective reliability. The objective reliability is defined as typical reliability quantity that can be evaluated through technical or objective parameters such as network parameters. Furthermore, subjective reliability is the reliability quantity that can be derived from subjective factors. The graphical relationship between these terms is represented by figure 6.4.

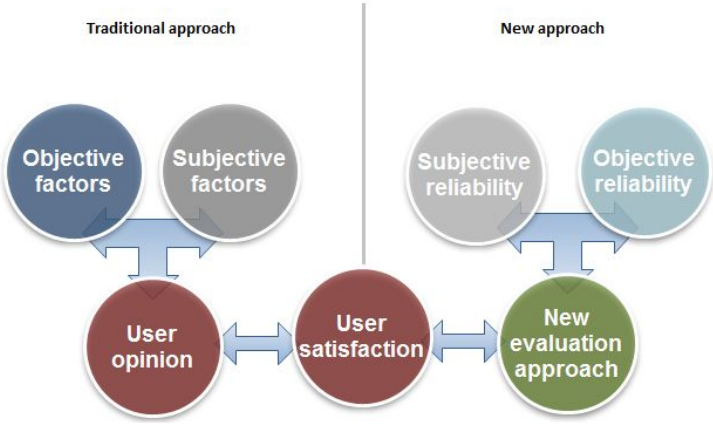


Figure 6.4: Relationship of affect factors and new reliability evaluation approach

6.3 The Resilience Assessment and Its Limitations

The QoR estimation can be dome by classification of the resilience measurement metrics into two main categories: short-term and long-term quality metrics [14].The short-term quality metrics are measurement parameters that affect the instantaneous availability of the system such as packet loss ratio, and packet delay. On the contrary, the long-term quality metrics are measurement parameters of overall service quality for the entire operating period. These parameters are traditional measures such as steady-state availability and mean time to failure. The short-term parameters such as instantaneous service availability and unavailability (downtime) can be estimated by dividing the operating period into a smaller period, Δt. Then the service availability and unavailability for those intervals can be represented by binary functions of user satisfaction: 1= fully satisfaction/available, 0 = unsatisfied/unavailable. The Δt period is chosen to be long enough to determine service downtime, which depends on the considering services or applications. The service state model per figure 6.5 was used to demonstrate the relationship between service availability and unavailability. However, the short term, instantaneous availability, was considered only at the degraded quality region where some of the QoS qualities are not compliant. The service availability and unavailability is determined by the short-term quality threshold based on the user-perceived service quality. The statistical data of the short-term service unavailability can be plotted and evaluated as downtime density histogram. The method, however, cannot evaluate dynamic failure behaviors between networking layers.

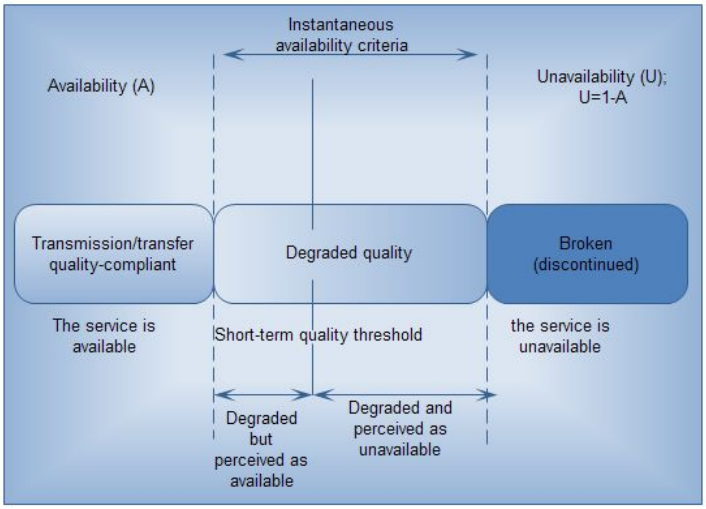


Figure 6.5: The service state model [14]

Moreover, the proposed method only consider the service degradation region, not the whole operating period. Further, assigning the proper threshold level for detecting service availability is not an easy task due to it involves many factors such as service types, and also it is difficult to determine whether or not the user is fully satisfied with perceived service quality during the short period (ΔtD. Therefore, overall parameters that affect the user's perceived quality is recommended for the evaluation. A modern evaluation approach of QoR is then proposed in next section. The method incorporates both objective and subjective parameters to estimate user satisfaction.

6.4 The Proposed Reliability Evaluation Framework

6.4.1 Selective resilience measurement parameter algorithm

According to the previous section and the ITU-T recommendation E.800, various quality measurement parameters have been proposed for different networking technologies such as Mean Down Time (mean down time (MDT)), Mean Time to Failure (MTTF), availability. Therefore, there is no standard or framework in how and what parameters should be selected to predict the system resilience. Therefore, this section proposes a selective resilience measurement

algorithm and modern reliability evaluation method. The proposed idea is inspired by the concept of QoR [14] where reliability parameters are classified into short and long term. However, the method has some disadvantages as mentioned in the section 6.3 and there is no administration or standard about how the measurement parameters should be collected for reliability assessments during what operating period.

Initial assumptions

To improve the drawbacks of the current method as mentioned in Section 6.3, the service state model per figure 6.6 is used for an evaluation process. The instantaneous availability needs to be evaluated for the whole operating period. Therefore, the observed results are obtained from any service quality regions: availability, degraded and unavailability of quality.

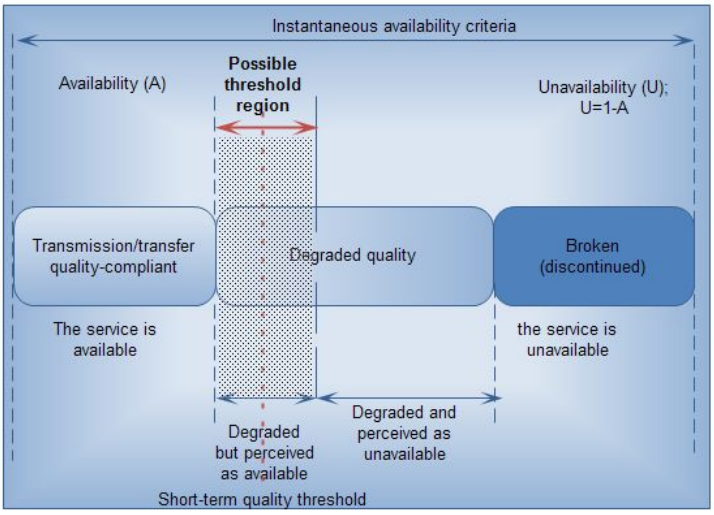


Figure 6.6: The service state model for the proposed selective parameter framework

In practice, determining if the service is fully satisfied or not during a short observation period. A number of estimation errors can easily convey the total results or failure distribution function into another direction. Therefore, a new threshold concept is proposed. The concept is developed from an evolution of failure terms: *Error*, *Fault* and *Failure*. An *Error* is a variance from accuracy. *Fault* is the imperfection of hardware or software, and a *Failure* is the incapability to perform a required function based on a given performance. Normally, either a number of error or fault events will easily lead to the failure

event. Therefore, a frequency of an error event is essential and used as an indicator of a high possibility of the first failure event. As mentioned earlier, it is difficult to precisely detect a failure during a short period due to most of the important hardware and software of contemporary systems are already designed and equipped with redundancy or automatic recovery mechanisms. Therefore, the best way to prepare for an uncertain event or a major failure is to prevent it from occurring. Accordingly, the threshold region per figure 6.6 is proposed. The region is defined as the area starting from the degraded service quality but perceived as the available region before reaching the degraded but perceived as the unavailable region. This region is before the threshold value of the traditional method. The proposed threshold, T_Q, is employed as a monitoring threshold if the undesired service quality, UQ, degraded below T_Q. The threshold can be fine-tuned or updated depending on the required sensitivity of the required service or system. The frequency of occurring UQ for the whole observation period can be calculated and represented as a percentage of having undesired signal quality by equation 6.1.

$$P_{f_i} = \frac{\sum UQ}{T_{observe}} \tag{6.1}$$

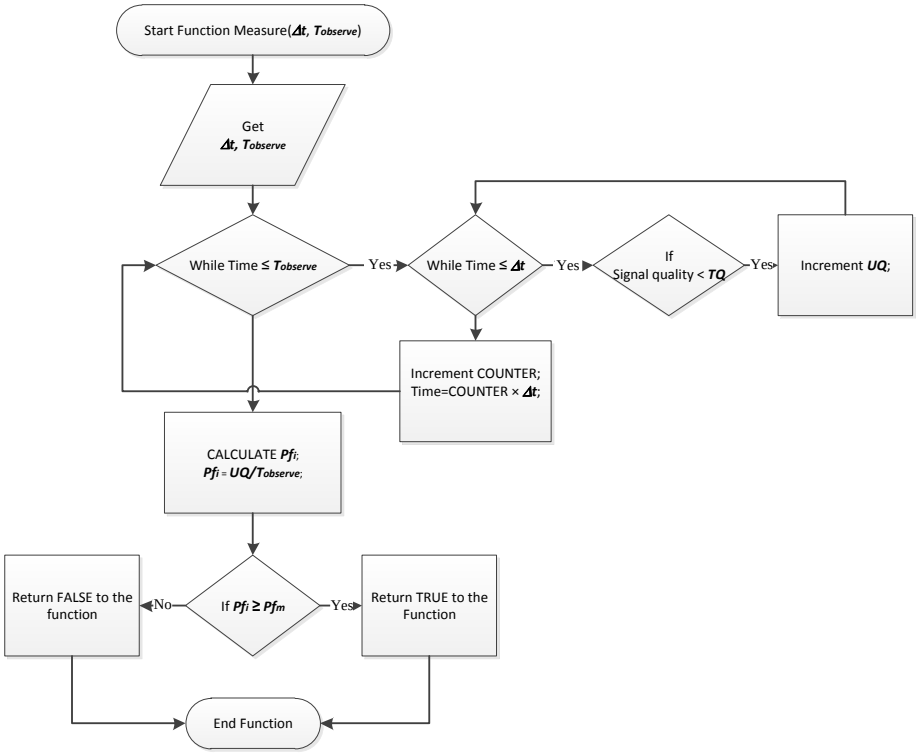
Where $\sum UQ$ is the summation of UQ during the short observation interval and $T_{observe}$ is the observation period. The period is less than or equal to the whole monitoring period, T_{total}. In order to evaluate P_{f_i}, the maximum allowed percentage, P_{fm}, is defined. The P_{fm} is used to indicate the maximum percentage of the allowance of a number of UQs for a given period before having or detecting the first failure event.

$$P_{fm} = \frac{Max \;\; no. \;\; of \;\; UQ}{T_{required \;\; operating \;\; period}} \tag{6.2}$$

Therefore, at the condition $P_{f_i} \geq P_{fm}$, the service checking schemes or system maintenance plan can be triggered or updated. The value is estimated to fit different service types or the need for the network operator to prevent the first failure event. Accordingly, the $T_{observe}$ is needed to be long enough to observe UQ per P_{fm} condition. Therefore, the relation of different periods can be given by equation 6.3.

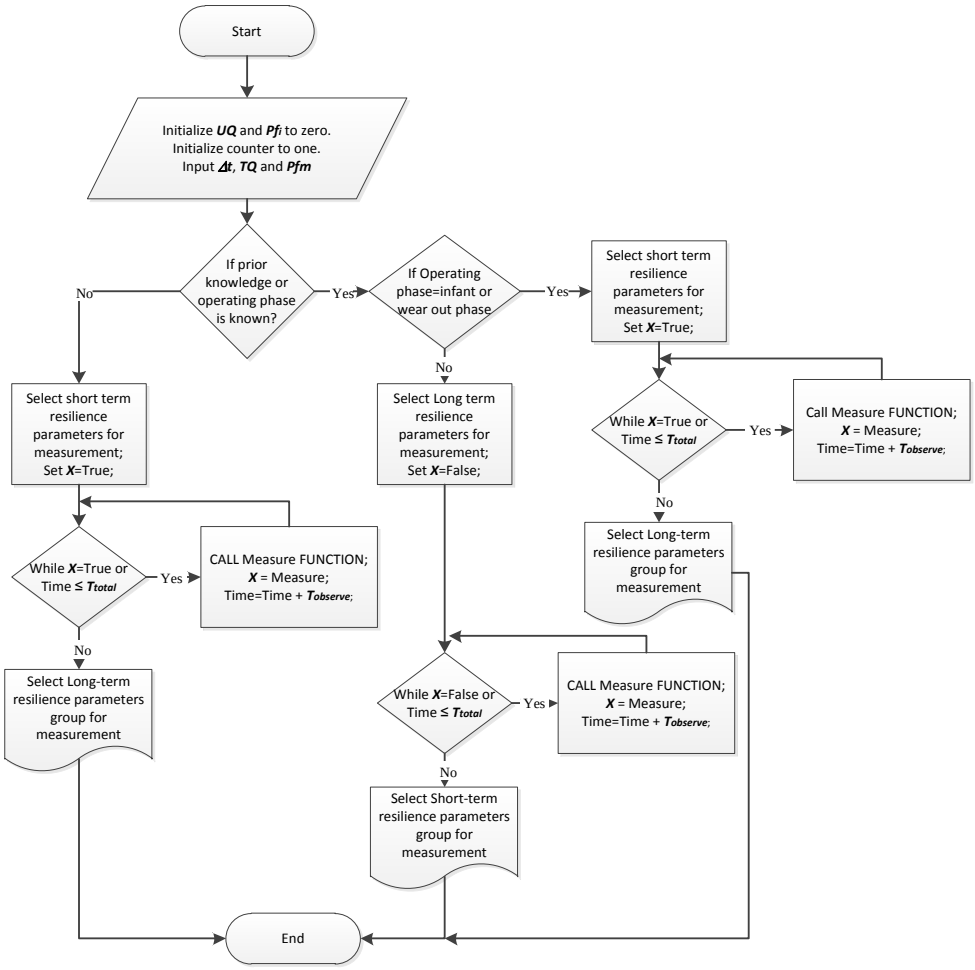
$$T_{observe} \leq T_{required \;\; operating \;\; period} \leq T_{total} \tag{6.3}$$

The selective algorithm



(a)

Figure 6.7: (a). The measurement function.



(b)

Figure 6.7: (continued) (b). The flowchart of the proposed selective parameter framework.

Based on subsection 6.4.1, the selective resilience measurement parameters algorithm is proposed and represented by figure 6.7. The algorithm proposes at presenting a suitable selective reliability measurement parameters framework at different operating periods. Moreover, the algorithm can be applied in the case of no prior information about the system. The algorithm can also be applied to any system regardless of the operating period. The main concept is employing short interval observation per [14] for the monitoring processes, so the detailed

characteristics of the service or system can be observed by using the proposed service state model of figure 6.6. The proposed framework is also applied as a major failure prevention method to trigger or update the maintenance plan based on the appropriate P_{fm} value.

6.4.2 The proposed reliability evaluation method

As mention in the chapter 2, different stochastic models have been developed and applied for reliability and performance analysis of system components. This section proposed one of the stochastic models called Bayesian statistic. The Bayesian approach has some advantages that can lead to the solution of the reliability estimation challenges as mentioned in the section 6.3. The Bayesian method has been developed and chosen as the consistency method for uncertainty evaluation in many research areas [15, 16]. The approach can take into account both objective (quantitative) and subjective (qualitative) factors. The Bayesian network modeling framework has been proposed where the quantitative Bayesian inference is logically predicted based on qualitative information [17]. Moreover, the state-of-the-art of Bayesian reliability models with Weibull failure distribution were exhibited [18]. The method can evaluates a system with unknown reliability structure at all three phases of the FR distribution. There are many varieties of Bayesian analysis, but the fundamental principle is based on the statistical expression of uncertainty of unknown parameters through the Bayes′s theorem. Therefore, the Bayesian analysis approach is flexible and applied to fit various system conditions. In addition, both expertise-based marketing decisions and technical data are possible to be combined and utilized for an efficient analysis. The approach also performs well in case of no prior system information. The practical use of the Bayesian theory in reliability estimation is reviewed as follows. Let X be the observed continuous random variable of the system's component, for instance, a life of the component. The distribution of X, $f(X)$, is varied by the unknown parameter, θ which represents the mean lifetime of the component. Inferences of the θ is estimated by the so-called Prior, $f(\theta)$, or distribution function of the θ. The possible range of the θ value is normally obtained by prior information or belief of an estimator. From the definition of conditional probability and Bayes theorem, the distribution of θ, given X is given by equation 6.4 and 6.5.

$$f(\theta \mid X) = \frac{f(X \mid \theta) f(\theta)}{f(X)} \tag{6.4}$$

Equation 6.4 , can be rewritten as

$$Posterior = \frac{Joint \;\; Distribution}{Marginal \;\; Distribution} \tag{6.5}$$

where $f(\theta \mid X)$ is the posterior distribution or the distribution of θ, given the parameter X. Alternatively, equation 6.4 is represented as

$$f(\theta \mid X) = \frac{1}{f(X)} \times f(X \mid \theta) \times f(\theta) \tag{6.6a}$$

$$f(\theta \mid X) = (normalized \;\; constant) \times f(X \mid \theta) \times f(X\theta) \tag{6.6b}$$

$$f(\theta \mid X) = (normalized \;\; constant) \times Likelihood \times Prior \tag{6.6c}$$

The $f(X \mid \theta)$ is equal to $L(\theta \mid X)$ or the likelihood of θ, given X which is equal to the function of X, given θ. Therefore, the posterior is evaluated by the relationship between the Joint and Marginal distribution or the relationship between the Likelihood and Prior data of the system. In case of n devices (denoted $X_1, X_2, \cdots, X_n$), the Posterior distribution which is the distribution of θ, given the parameters $X_1, \cdots, X_n$ can be given by equation 6.7.

$$f(\theta \mid X_1, \cdots, X_n) = \frac{Joint \;\; Distribution}{Marginal \;\; Distribution} \tag{6.7a}$$

$$f(\theta \mid X_1, \cdots, X_n) = \frac{f(X_1, \ldots, X_n \mid \theta) \times f(\theta)}{\int f(X_1, \ldots, X_n, \theta) d(\theta)} \tag{6.7b}$$

With the assumption of the Weibull failure distribution, $f_\theta(X)$ is given by equation 6.8.

$$f(X \mid a, b, \tau) = \frac{a}{b} \left(\frac{X - \tau}{b} \right)^{a-1} exp\left[- \left(\tfrac{X-\tau}{b} \right)^a \right] \tag{6.8}$$

where a is called shape parameter, b is called scale parameter, and τ is called the delay or location parameter. The distribution of the function is varied through these parameters; hence, different phases of the bathtub curve can be resemble according to this variation. Please be noted that the θ in this case refers to parameters a, b or τ. The failure or hazard rate function that corresponds to equation 6.8 is given by equation 6.9.

$$h(X) = \frac{a}{b} \left(\frac{X - \tau}{b} \right)^{a-1} \tag{6.9}$$

Then, the reliability function can be simply calculated and given as equation 6.10.

$$R(X) = exp\left[- \int_0^x h(X) dx \right] \tag{6.10a}$$

$$R(X) = exp\left[-\left(\frac{X-\tau}{b}\right)^{a}\right] \tag{6.10b}$$

Although, the Bayesian can incorporate qualitative information, the user-related factors and sociological aspects are not included via the above Bayesian analysis. Nonetheless, considering all uncertainty factors for quantitative evaluation is complicated. As mentioned by the subsection 6.2.2, the important of user-related factors or user satisfaction is increased and strongly affect the reliability of modern service or system. Therefore, the user-related factors are advised to take into account for modern reliability evaluation approach. From the equation 6.6c, the Posterior depends on the Likelihood and Prior information of the system. The Prior is typically acquired from pre-existing data about the system; as well, it can be acquired based on the belief of an experienced expert. For instance, the belief to have a similar distribution with the Posterior is called conjugate Prior [16]. Nevertheless, the belief is mainly related to the belief of the system components and behaviors. The Likelihood represents the relationship between the unknown parameter θ and the observed parameter X or how likely of X, given θ or vice versa. Accordingly, the modern reliability evaluation approach could be proposed and given by equation 6.11.

$$Posterior = (normalized\ \ constant) \times Likelihood \times LoSF \times Prior \tag{6.11}$$

Where $LoSF$ is the likelihood of the subjective factors; the term represents the likelihood of θ, given subjective factors. As a result, the proposed Posterior does not limit only for the objective factors, but also the subjective factors. In Particular, the subjective factors that influence user satisfaction of the service or system can take into account for an evaluation. As far as the characteristic of the subjective factors is concerned, their effect on user satisfaction is random and time-dependent by nature. Besides, these factors do not depend on system performance or objective factors. Thus, the $LoSF$ value is varied with time regardless of the performance of the system. For giving an idea about how the $LoSF$ affects the new posterior approach, the results trend per equation 6.11 are simulated and presented by table 6.1.

Table 6.1: The sample trend of new Posterior approach per the proposed equation 6.11

Likelihood		**Posterior Trend**	
Typical likelihood (Objective)	*LoSF* **(Subjective)**	**Short-term**	**Long-term**
H	**H**	**H**	**H**
H	**L**	**L**	**H**
L	**H**	**H**	**L**
L	**L**	**L**	**L**

The table 6.1 represents the sample simulation trend of the Posterior results when taken into account both typical Likelihood and $LoSF$. The $LoSF$ mainly focuses on subjective factors or the user's satisfaction while the typical likelihood term is mainly related to the objective factors. The results present two possible Likelihood values ($1 = Maximum/High(H)$ *or* $0 = Minimum/Low(L)$) and two main periods: short and long-term due to an uncertainty of user satisfaction or the $LoSF$. The short-term results are mainly affected by the subjective factors while the objective factors are mainly affected by the long-term quality results. In brief, the modern reliability evaluation approach incorporates both short-term and long-term posterior results. The results can also be continuously updated or estimated by applying various Bayesian inference techniques.

6.5 Conclusions

This section reveals a relationship and definition of several quality terms through their measurement parameters. An important of user satisfaction is pointed out as the main objective for modern services and systems. Moreover, some limitations of reliability evaluation approach have been exhibited. Therefore, the selective reliability parameters algorithm and modern Bayesian reliability evaluation approach for a system with Weibull distribution are proposed in this section. The proposed method can well incorporate both traditional objective quality and subjective quality factors for future reliability evaluation framework.

6.6 References

[1] Djamal-Eddine Meddour, Usman Javaid, N. Bihannic, T. Rasheed, and R. Boutaba. Completing the convergence puzzle: a survey and a roadmap. *Wireless Communications, IEEE*, 16(3):86–96, 2009.

[2] Kishor S. Trivedi. *Probability and statistics with reliability, queuing and computer science applications*. John Wiley and Sons Ltd., Chichester, UK, 2nd edition edition, 2002.

[3] Robin A Sahner, Kishor Trivedi, and Antonio Puliafito. *Performance and reliability analysis of computer systems: an example-based approach using the SHARPE software package*. Springer Science & Business Media, 2012.

[4] R. Stankiewicz, P. Cholda, and A. Jajszczyk. Qox: What is it really? *Communications Magazine, IEEE*, 49(4):148–158, 2011.

[5] R. Stankiewicz and A. Jajszczyk. A survey of qoe assurance in converged networks. *Comput. Netw.*, 55(7):1459–1473, May 2011.

[6] Fernando Kuipers, Robert Kooij, Danny De Vleeschauwer, and Kjell Brunnström. Techniques for measuring quality of experience. In *Proceedings of the 8th international conference on Wired/Wireless Internet Communications*, WWIC'10, pages 216–227, Berlin, Heidelberg, 2010. Springer-Verlag.

[7] Qin Dai. A survey of quality of experience. In Ralf Lehnert, editor, *Energy-Aware Communications*, volume 6955 of *Lecture Notes in Computer Science*, pages 146–156. Springer Berlin Heidelberg, 2011.

[8] H.G. Msakni and H. Youssef. Provisioning qoe over converged networks: Issues and challenges. In *High Performance Computing and Communication 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012 IEEE 14th International Conference on*, pages 891–896, 2012.

[9] A. Khan, Lingfen Sun, E. Ifeachor, Jose Oscar Fajardo, and F. Liberal. Video quality prediction model for h.264 video over umts networks and their application in mobile video streaming. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1–5, 2010.

[10] J. Valerdi, A. Gonzalez, and F.J. Garrido. Automatic testing and measurement of qoe in iptv using image and video comparison. In *Digital Telecommunications, 2009. ICDT '09. Fourth International Conference on*, pages 75–81, 2009.

[11] János Tapolcai, Dániel Máthé, András Zahemszky, Achim Autenrieth, Piotr Chołda, Tibor Cinkler, Didier Colle, and Krzysztof Wajda. Quantification of resilience for voice-over-ip applications. *Proc. ISBAT 2006*.

[12] Reinhilde Veugelers, editor. *The Evaluation of the Finnish National Innovation System - Full Report*. The Research Institute of the Finnish Economy, 2009.

[13] E. Ibarrola, Jin Xiao, F. Liberal, and A. Ferro. Internet qos regulation in future networks: a user-centric approach. *Communications Magazine, IEEE*, 49(10):148–155, 2011.

[14] Piotr Cholda, János Tapolcai, Tibor Cinkler, Krzysztof Wajda, and Andrzej Jajszczyk. Quality of resilience as a network reliability characterization tool. *Network, IEEE*, 23(2):11–19, 2009.

[15] William M. Bolstad. *Introduction to Bayesian Statistics*. Wiley-Interscience, 2nd edition, 2007.

[16] C. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, 2007.

[17] Rui Chang and M. Stetter. Quantitative bayesian inference by qualitative knowledge modeling. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, pages 2563–2568, 2007.

[18] Abdelaziz Zaidi, Belkacem Ould Bouamama, and Moncef Tagina. Bayesian reliability models of weibull systems: State of the art. *Applied Mathematics and Computer Science*, 22(3):585–600, 2012.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

This thesis has shown different aspects of reliability and quality terms definitions and contemporary end-to-end reliability analysis trend due to its requirement for future services and applications such as financial, medical and disaster warning and broadcast system. The study has sought the relation of the state-of-the-art of resilience parameters in term of quality of service (QoS) and quality of resilience (QoR). The study has also explored the key network reliability parameters and examined for describing the reliability of different IMS-based network topologies. Moreover, different reliability and performance analysis and methodologies of the IP multimedia subsystem (IMS) system has been evaluated and studied. The thesis designates that there is no standard or unique reliability or performance method and model that can meet all requirements of the services and system, especially the IMS system.

The thesis offered the proposed continuous-time Markov chain (CTMC) model for both simplex and redundancy system are measured and weighed with state-of-the-art models as presented in the chapter 2. The numerical analysis and the simulation results of the proposed models as shown in the chapter 3, chapter 4, and chapter 5 correspond well with the system reliability hypothesis and exhibit better failure and recovery characteristics of the system. Therefore, the proposed models with detailed failure and recovery rates can interpret sufficient reliability and availability characteristics for both simplex and redundancy system.

In addition, the reliability impact due to different redundancy conditions is exposed. The simulation results contribute interesting reliability features of

two important IMS-based communication scenarios: intra-domain and inter-domain (similar and different home domain) communication. The end-to-end reliability behaviors of inter-domain or long distance communication scenarios clearly are magnified with just a single redundancy. Moreover, an optimization of the number of parallel redundancy as shown in the chapter 4 determine an opportunity to formulate high availability and reliability system.

Finally, the modern reliability evaluation framework is exhibited in the chapter 6. The evaluation framework gives a novel idea to incorporate both objective and subjective parameters which include user-related factors for an evaluation. The method can weight overall reliability factors for estimation of future services and systems.

To conclude, the thesis addressed potential solutions to the end-to-end reliability challenges as stated in section 1.3 of the chapter 1. However, the proposed evaluation methods and models did not fully investigate all related factors such as quality terms, network parameters, and possible communication scenarios. However, the thesis reveals significant end-to-end quality and reliability limitations and realistically evaluation design for further extending toward future service and system.

7.2 Scope for the future work

Based on the conclusion, there are several approaches which can be further developed to assess end-to-end quality and reliability of modern services and systems. For instance, there are many stochastic models which can represent different characteristics of a different system. Therefore, a combination of these models can be reasonably implemented. Not only the hybrid model can link advantages of each model but also can increase the complexity of an evaluation. Therefore, the models and method are needed to be carefully chosen, designed and optimized to satisfy the service and system.

Moreover, with the internet technology, failure behaviors of important services and systems can be remotely observed and stored in the database. These data can be shared and cooperate among network administrators and service providers. This implies that a better quality and reliability evaluation can be predicted and achieved through an online or central high-performance computing. Besides, the failure or maintenance plan can be immediately informed and assisted through the modern communication system technologies.

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-721-8

AALBORG UNIVERSITY PRESS