DESIGN AND ANALYSIS FOR THE 3G IP MULTIMEDIA SUBSYSTEM

Presented

by

Muhammad Tanvir Alam

Bachelor of Science, North South University, Bangladesh

Master of Science, Oklahoma State University, USA

Graduate Student Member, IEEE

A dissertation submitted in fulfilment of the requirements of the degree of Doctor of Philosophy for the School of Information Technology, Bond University

Supervised

by

Dr. Zheng da Wu

Associate Professor of Computer Science

August 2007

DESIGN AND ANALYSIS FOR THE 3G IP MULTIMEDIA SUBSYSTEM

Statement of Originality

This thesis represents my own work and the material in this thesis has not been previously submitted for a degree or diploma in any university. To the best of my knowledge this thesis contains no material previously published or written by another person except where due acknowledgement is made in the thesis itself.

Signature:

Date:

Muhammad T. Alam

SID: 12793059

E-mail: muhamta123@gmail.com

School of Information Technology

Bond University

Gold Coast, QLD 4229

Australia

Thesis Abstract

The IP Multimedia Subsystem (IMS) is the technology that will merge the Internet (packet switching) with the cellular world (circuit switching). It will make Internet technologies, such as the web, email, instant messaging, presence, and videoconferencing available nearly everywhere. Presence is one of the basic services that is likely to become omnipresent in IMS. It is the service that allows a user to be informed about the reachability, availability, and willingness of communication of another user. Push to talk over Cellular (PoC) is another service in IMS that is intended to provide rapid communications for business and consumer customers of mobile networks. In order to become a truly successful mass-market service for the consumer segment, the only realistic alternative is a standardized Push-to-talk solution providing full interoperability between terminals and operators. Instant Messaging (IM) is the service that allows an IMS user to send some content to another user in near real-time. This service works under IETF's Message Session Relay protocol (MSRP) to overcome the congestion control problem. We believe the efficiency of these services along with the mobility management in IMS session establishment has not been sufficiently investigated.

In this research work, we identify the key issues to improve the existing protocols in IMS for better system behaviour. The work is centred on the three services of IMS: (1) Presence Service, (2) Push-to-Talk over cellular and, (3) Instant Messaging and over the issue of (4) IMS session set up. The existing session establishment scenario of IP Multimedia Subsystem (IMS) suffers from triangular routing for a certain period of time when an end IMS user or terminal is mobile. In this thesis, the performance of three possible session establishment scenarios in a mobile environment is compared by using an analytical model. The model is developed based on the expressions of cost functions, which represents system delay and overhead involved in sessions'

establishment. The other problem areas in optimizing presence service, dimensioning a PoC service and analysing service rates of IM relay extensions in IMS are identified. A presence server becomes overloaded when massive number of IMS terminals joins a network to request presence facility. Performance models are developed in this research to mitigate such load during heavy traffic for the presence service. Queuing analyses for different cases are provided while instant messaging chunks go through two consecutive relay nodes. The specific factors such as blocking probability, stability conditions, optimized subscription lifetime etc. in IMS environment parameters have been investigated. We have also elaborated models to dimension a PoC service for service providers with regards to controlling PoC session access, optimal PoC session timer, path optimization and number of allowable simultaneous PoC sessions for given network grade of service.

In a nutshell, the contribution of this dissertation are: (a) a proposed robust scheduler to improve performance of the IMS presence service, (b) several derived models to dimension IMS Push-to-talk over cellular service, (c) a new mechanism to reduce cost for the IMS session set ups in mobile environment and (d) evaluation of message blocking and stability in IMS Instant Messaging (IM) service by applying queuing theories. All of these analyses have resulted in recommendations for the performance enhancements with optimal resource utilization in IMS framework.

Acknowledgements

This PhD thesis would not have been possible without the support of many people and organizations to who I owe a great deal. I would like to thank particularly:

Dr. Zheng da Wu, for his belief in my abilities, his advice, his encouragement, his wisdom, his intellectual brilliance, and his endless kindness. Dr. Wu provided me the support I needed at every step. It is his careful supervision that brought me wherever I stand today in my research career.

The School of Information Technology, Bond University, for providing me with a scholarship, resources and the opportunity to pursue with a PhD, as well as introducing me to start my teaching career. Special thanks to Dr. Michael Rees and Stephanie Patching, who continually supported me and made all the difficult paperwork that much easier.

The Australian Government, for providing me with an International Postgraduate Research Scholarship Award during my PhD candidature.

The Australian Academy of Technological Sciences and Engineering (ATSE) for providing me with the Early Career Symposium Fellowship (ECEF) award that inspired me a great deal when I needed.

I would also wish to thank my fellow research students and friends who have ‘come and gone’ during my stay in the research room, room 5323 of IT School, Bond University.

Finally, I would like to thank my parents, who supported and encouraged me throughout this journey of pursuing my doctoral studies. I am ever grateful to their unselfish blessings and support. Every page of this thesis is indeed dedicated to them.

To

My Parents

and

Grandfather,

Late Shafiul A. Chowdhury

Publication Arising from this Thesis

Book Articles

1. M. T. Alam (2007). An Optimal Timer for Push-To-Talk Controller, In D. Taniar (Ed.), *Encyclopaedia of Mobile Computing and Commerce,* Vol: 2, pp: 724-728, Hershey, PA: Information Science Reference.

2. M. T. Alam (2007). Protocol Analysis Over 3G IP Multimedia Subsystem, In D. Taniar (Ed.), *Encyclopaedia of Mobile Computing and Commerc*e, Vol: 2, pp: 778-784, Hershey, PA: Information Science Reference.

Journal Papers

1. M. T. Alam, Z. D. Wu (2007). "Dimensioning and Optimization of Push-to-Talk over Cellular Server", *International Journal of Network Management,* John Wiley & Sons, Ltd. (In press, www.interscience.wiley.com), DOI: 10.1002/nem.625.

2. M. T. Alam, Z. D. Wu (2007). "Optimal Routing for SIP-based Session Set up over IMS in Mobile Environment", *International Journal of Internet Protocol Technology,* (In Press) Inderscience (www.inderscience.com).

3. M. T. Alam, Z. D. Wu (2007). "End-to-End Delay Measurement for Instant Messaging Relay Nodes," *Ubiquitous Computing and Communication Journal,* Vol 2 (2), pp: 1-11, ISSN: 1992-8424.

4. M. T. Alam (2006). "On Analysing Cost for Optimizing the Watcher Subscription Time in the IMS Presence Service", *Engineering Letters,* Vol 13(1), pp: 1-10, ISSN: 1816-093X.

5. M. T. Alam, Z. D. Wu (2006). "Admission Control Approaches in the IMS Presence Service", *International Journal of Computer Science,* WASET, Vol 1(4), pp: 299-314.

Additional Journal Publication Relevant but not Forming Part of the Thesis

6. M. T. Alam, Z. D. Wu (2007), "Asymptotic Analysis of Instant Messaging Service with Relay Nodes", *International Journal of Computer, Information, and Systems Science and Engineering,* Vol 1(1), pp: 1-9, ISSN: 1307-2331.

Conference Papers (fully refereed)

1. M. T. Alam, Z. D. Wu (2007). "Proposed Techniques to Dimension a Push-To-Talk over Cellular Server", *IEEE Consumer Communications & Networking Conference,* CCNC'07, 11-13 Jan, Las Vegas, Nevada.

2. M. T. Alam, Z. D. Wu (2006). "Efficient Scheduling for Reducing Load in the IMS Presence Service" *The IASTED International Conference on Wireless*

Networks and Emerging Technologies, July 3 – 5, Banff, Alberta, pp: 459-464, ACTA press, ISBN: 0-88986-563-9.

3. M. T. Alam, Z. D. Wu (2006). "Cost Analysis of the IMS Presence Service," *First IEEE International Conference on Wireless Broadband and Ultra Wideband Communications* AusWireless'06 Conference, 13-16 March 2006, Sydney, Australia, In CD, available at http://epress.lib.uts.edu.au/dspace/handle/2100/165.

4. M. T. Alam, Z. D. Wu (2005). "Comparison of Session Establishment Schemes Over IMS in Mobile Environment." *Fifth IEEE International Conference on Information, Communications and Signal Processing (ICICS 2005), ISBN:* 0-7803-9283-3, IEEE Catalogue Number: 05EX1118C, December 6-9, Bangkok, Thailand, pp: 638-642, Paper ID: 0581 ***(Registration fee waiver award)*** .

5. Muhammad T. Alam (2005). "An Optimal Method for SIP-Based Session Establishment Over IMS." *2005 International Symposium on Performance Evaluation of Computer And Telecommunication Systems (SCS 2005),* July 24-28, Hilton Cherry Hill/Philadelphia, Philadelphia, Pennsylvania, Sim Series., Vol 37, No. 3, pp: 692-698.

6. M. T. Alam, Z. D. Wu (2005). "Performance Analysis of SIP-Based Session Establishments Over IMS." *The IASTED International Conference on Wireless Networks and Emerging Technologies,* July 19 – 21, 2005, Banff, Alberta, Canada, pp: 178-183, ACTA press, ISBN: 0-88986-499-3, Paper ID: 474-022.

Additional Publication Relevant but not Forming Part of the Thesis

7. M. T. Alam, J. P. Thomas, I. Jonyer (2004). "Reducing Latency in Mobile Ad Hoc Networks by pre-fetching." *The 2004 International Conference on Pervasive Computing and Communications,* (ICWN'04, PCC'04) Las Vegas, Nevada, Vol 2, pp: 952-958, CSREA press

Contents

Chapter 1 Introduction 1

Chapter 2 Background 4
2.1 Overview of IMS Architecture 4
2.2 SIP in IMS 7
2.3 Session Establishment Scenario for a Mobile Terminal 8
2.4 Registration Scenario in IMS 11
2.5 Presence Service in the IMS 13
2.6 Push-to-Talk Service in IMS 20
2.6.1 PoC Server 24
2.6.2 Controlling PoC Function 25
2.6.3 Participating PoC Function 26
2.7 Instant Messaging in IMS 26
2.7.1 Modes of IM 27

Chapter 3 Literature Review 29
3.1 Quality of Service Issues 30
3.1.1 Message formats in IMS presence service 31
3.1.2 Subscription / Registration time 34
3.1.3 Presence Optimizations by IETF 36
3.1.4 Related work on PoC service 39
3.1.5 Mobility management in IPv6 44
3.1.6 Mobility management in SIP 46
3.1.6.1 HMSIP architecture 49
3.1.7 MIP and SIP Interactions 51
3.1.8 Constraints of Instant Messaging 52
3.1.9 Solutions from the literature on IM 53
3.1.9.1 MSRP Relays 55
3.2 Discussion of Problems based on Lit Review 56
3.3 Objective & Methodology 61

Chapter 4 Admission Control for Presence Server 63
4.1 Introduction 63
4.2 Overview of Class Based Queuing 65
4.2.1 Dynamic Class-Based Queue Management (DCQM) 68
4.2.2 Adaptive Group Size 69
4.3 Proposed Queuing System 70
4.4 Admission Control Mechanisms 73
4.4.1 Blocking Probability 75
4.4.2 Efficient Dropping of Buffer Messages 79
4.4.3 Performance Analysis 83
4.4.4 Effective Bandwidth 91
4.4.5 Transition Probabilities 92
4.5 Cost Consumption for PS 95
4.6 Simulation for Cost Consumption 96
4.7 Method for Optimizing Subscription Time 103
4.8 Summary 108

Chapter 5 Dimensioning Push-To-Talk Over Cellular Service 111
5.1 Introduction 111

5.2 The Four Problem Overview 113
5.3 Model Assumptions 118
5.4 Controlling Session Access 120
5.5 Load Sharing at PoC BS 125
5.6 Timer Control 128
5.7 Optimization of Simultaneous Sessions 132
5.7.1 Estimating steady state probabilities 136
5.7.2 Optimal values 138
5.8 Summary 144

Chapter 6 Efficient IMS Session Set Up in Mobile Environment 145
6.1 Introduction 145
6.2 Scenario Description 146
6.2.1 Reasons of a Session failure 147
6.2.2 The Three Session Set up Scenarios 148
6.3 Modelling 149
6.4 Simulation Model 158
6.5 Simulation Results 163
6.6 Threshold from Simulation 177
6.7 Queuing Analysis for Nodes 180
6.8 Summary of Analysis 183

Chapter 7 Queuing Analysis for Instant Messages with Relay Nodes 185
7.1 Introduction 185
7.2 Chunking method of MSRP 185
7.3 The Special Scenario and Related Work 187
7.4 System Assumptions 189
7.5 Modelling 190
7.5.1 Service rate of the first relay node is infinite 194
7.5.2 First relay node is saturated 196
7.5.2.1 Condition for stability 199
7.6 Summary 201

Chapter 8 Conclusions and Future Work 203

Appendix A Steady State of BCMP Model 210

Appendix B Effective Bandwidth of a Flow in WCBQ 211

Appendix C M/M/m Queuing System 214

Appendix D Load Sharing Expression 216

Appendix E Poisson Inter-Arrival Time & Density Function 217

Bibliography 219

List of Figures

Figure 2-1: GPP IMS architecture overview 4
Figure 2-2: Serving to serving procedure - same operator [24] 9
Figure 2-3: Serving to PSTN procedure - different operator [24] 10
Figure 2-4: Registration – User not registered [24] 12
Figure 2-5: Re-registration - user currently registered [24] 13
Figure 2-6: SIP Presence system 14
Figure 2-7: SIP-based IMS presence architecture 16
Figure 2-8: The XCAP protocol stack 17
Figure 2-9: Watcher subscription to own list 18
Figure 2-10: The RLS subscription to a presentity 19
Figure 2-11: The IMS terminal publishing presence information 20
Figure 2-12: Example of a PoC 1-to-many Group Session (voice transmission) [112]. 21
Figure 2-13: Logical architecture of PoC [112] 22
Figure 2-14: PoC architecture [112] 23
Figure 2-15: Direct media flow between Controlling PoC Function and PoC Client [112] 25
Figure 2-16: Pager-mode instant messaging in the IMS 28
Figure 3-1: Example of the RPID 33
Figure 3-2: Publishing and notifying presence information [49] 34
Figure 3-3: Example of the timed status extension 35
Figure 3-4: Resource list through an exploder 37
Figure 3-5: Simulation architecture of Raktale [139] 41
Figure 3-6: Pre-established Session [112] 42
Figure 3-7: Hierarchical registration in SIP [29] 47
Figure 3-8: HMSIP architecture for intra-domain handoff 50
Figure 3-9: Typical MSRP session with relays [213] 55
Figure 4-1: PS notifying watchers of a presentity's state change 63
Figure 4-2: CBQ building blocks 66
Figure 4-3: CBQ to estimate the throughput uses the rate of the bytes sent to calculate the inter-departure time 67
Figure 4-4: WCBQ model 71
Figure 4-5: Flow chart for WCBQ 72
Figure 4-6: Comparison of Group 1 blocking performance for varying offered traffic. 85
Figure 4-7: Comparison of Group 2 blocking performance for varying offered traffic. 86
Figure 4-8: Comparison of Group 3 blocking performance for varying offered traffic. 86
Figure 4-9: Probability of servicing messages for the three traffic groups 87
Figure 4-10: Minimum sojourn time for group 3 88
Figure 4-11: Probability of more than 2 arrivals for given Sojourn time 89
Figure 4-12: Number of message generation saved under WCBQ and throttled WCBQ 90
Figure 4-13: Markov chain for a presentity's states 93
Figure 4-14: Number of terminals watching at the class rate 97
Figure 4-15: Messages dropped in a minute on average 98
Figure 4-16: Cumulative message drop during simulation period 99
Figure 4-17: Comparison of message generation cost 99
Figure 4-18: Network Topology of message streams for FCFS 100
Figure 4-19: Average number of nodes watching a node at each class rate 101

Figure 4-20: Cost comparison of WCBQ-20 vs WCBQ-30 102
Figure 4-21: Number of message generation saved by throttled WCBQ compared to FCFS 103
Figure 4-22: Number of message generation saved by throttled WCBQ compared to WCBQ 103
Figure 4-23: Method for optimizing subscription time 105
Figure 4-24: Optimal lifetime of a watcher 106
Figure 4-25: Subscription and Notification of Presence information 107
Figure 5-1: (a) PoC short session (b) PoC long session and (c) Normal phone call 115
Figure 5-2: PoC route optimization between two PoC clients: *i* and *j* 117
Figure 5-3: Behaviour of session inter-arrival rate in terms of probability density function 119
Figure 5-4: Behaviour of session inter-arrival rate in terms of Cumulative distribution function 119
Figure 5-5: Markov model for accessing session 121
Figure 5-6: Total blocking probability for different protection level with 5 TRUs 123
Figure 5-7: Blocking probability for protection level with 5 TRUs 124
Figure 5-8: Effect of T for multiple installed TRUs 131
Figure 5-9: Effect of timer for various length slots 132
Figure 5-10: Markov model for the PoC BS states 134
Figure 5-11: Session states of a PoC Client 134
Figure 5-12: Four state Markov chain for session set up 136
Figure 5-13: Number of allowable simultaneous sessions 140
Figure 5-14: Number of allowable simultaneous sessions 141
Figure 5-15: Number of allowable simultaneous sessions 143
Figure 6-1: Three options for IMS session set up 147
Figure 6-2: Mobility in IMS by SIP 148
Figure 6-3: Movement of an IMS terminal 154
Figure 6-4: SIP session set up over UDP 159
Figure 6-5: Experimental test-bed prototype 161
Figure 6-6: Cost comparison for 4.8 Kbps, arrival rate 50msg/s 164
Figure 6-7: Cost comparison for 4.8 Kbps, arrival rate 100msg/s 164
Figure 6-8: Cost comparison for 4.8 Kbps, arrival rate 200msg/s 165
Figure 6-9: Cost comparison for Option 3 being successful in 2nd trial with 4.8Kbps channel and arrival rate 50msg/s 165
Figure 6-10: Cost comparison for Option 3 being successful in 2nd trial with 4.8 Kbps channel and arrival rate 100msg/s 166
Figure 6-11: Cost comparison for Option 3 being successful in 2nd trial with 4.8 Kbps channel and arrival rate 200msg/s 166
Figure 6-12: Cost comparison for 9.6 Kbps, arrival rate 50msg/s 167
Figure 6-13: Cost comparison for 9.6 Kbps, arrival rate 100msg/s 168
Figure 6-14: Cost comparison for 9.6 Kbps, arrival rate 200msg/s 168
Figure 6-15: Cost comparison for Option 3 being successful in 2nd trial with 9.6Kbps channel and 50msg/s 169
Figure 6-16: Cost comparison for Option 3 being successful in 2nd trial with 9.6Kbps channel and 100msg/s 169
Figure 6-17: Cost comparison for Option 3 being successful in 2nd trial with 9.6Kbps channel and 200msg/s 170
Figure 6-18: Option 3 cost for varying Q with arrival rate 50msg/s and 9.6Kbps channel 173
Figure 6-19: Option 3 cost for varying Q with arrival rate 100msg/s and 9.6Kbps channel 173

Figure 6-20: Option 3 cost for varying Q with arrival rate 200msg/s and 9.6Kbps channel 174
Figure 6-21: Q for various p_f 174
Figure 6-22: Cost for increased arrival rate with 4.8 Kbps channel 175
Figure 6-23: Cost for increased arrival rate with 9.6Kbps channel 176
Figure 6-24: Packet loss rate in 4.8Kbps channel 176
Figure 6-25: Packet loss rate in 9.6Kbps channel 177
Figure 6-26: P Vs Packet loss rate for 4.8Kbps channel with arrival rate 50msg/s 179
Figure 6-27: P Vs Packet loss rate for 9.6Kbps channel with arrival rate 50msg/s 180
Figure 6-28: *n* M/M/1 queues in series 182
Figure 7-1: Breaking a Message into Chunks [91] 186
Figure 7-2: Two-stage tandem network 188
Figure 7-3: SEND system with blocking for 2 relays open queuing 190
Figure 7-4: State changes of the relay nodes: (a) for state (0, 0), (b) for state (1, 0), (c) for state (2, 0), (d) for state (0, 1) 192
Figure 7-5: Plot of Eq. (7-20) 199
Figure 7-6: Plot for stability condition 201
Figure C-1: M/M/m queuing system [162] 214
Figure C-2: The state diagram of the M/M/m queue [162] 215

List of Tables

Table 3-1: PoC Call setup performance [139] 41
Table 4-1: Parameters for blocking performance with varying load 84
Table 5-1: Cost model for introduction of Push-to-Talk service [138] 112
Table 5-2: Blocking probabilities for N=10 125
Table 5-3: Number of allowable simultaneous sessions for a PoC client 139
Table 5-4: Number of allowable simultaneous sessions for a PoC client 140
Table 5-5: Number of allowable simultaneous sessions for a PoC client 141
Table 5-6: Number of allowable simultaneous sessions for a PoC client 142
Table 5-7: Number of allowable simultaneous sessions for a PoC client 143
Table 6-1: Message size for SIP over UDP/IPv6 158
Table 6-2: Percentage gain for doubling bandwidth with arrival rate 50msg/s 171
Table 6-3: Percentage gain for doubling bandwidth with arrival rate 100msg/s 171
Table 6-4: Percentage gain for doubling bandwidth with arrival rate 200msg/s 172

Abbreviations

ACK	Acknowledgement
AS	Application Server
BS	Base Station
BU	Binding Update
CBQ	Class Based Queuing
CDMA	Code Division Multiple Access
CN	Corresponding Node
Diffserv	Differentiated Services
FER	Frame Error Rate
GGSN	Gateway GPRS Support node
GoS	Grade of Service
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
HA	Home Agent
I-CSCF	Interrogating-Call/Session Control Function
IETF	Internet Engineering Task Force
IM	Instant Messaging
IMS	IP Multimedia Subsystem
IP	Internet Protocol
MIME	Multipurpose Internet Mail Extension
MIP	Mobile IP
MMD	Multimedia Domain
MN	Mobile Node
MSRP	Message Session Relay Protocol
MTU	Maximum Transmit Unit
NAT	Network Address Translator
OMA	Open Mobile Alliance
PA	Presence Agent
P-CSCF	Proxy-Call Session Control Function
PDP	Policy Decision Point
PoC	Push-to-Talk over Cellular
PRACK	Provisional Acknowledgement

PS	Presence Server
PTT	Push-to-Talk
PUA	Presence User Agent
QoS	Quality of Service
RAN	Radio Access Network
RLS	Resource List Server
RPID	Rich Presence Information Data Format
S-CSCF	Serving- Call/Session Control Function
SCTP	Stream Control Transmission Protocol
SDP	Session Description Protocol
SGSN	Serving GPRS Support node
SIP	Session Initiation Protocol
SMS	Short Messaging Service
TCP	Transmission Control Protocol
TRU	Transmit/Receive Unit
UA	User Agent
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
URI	Uniform Resource Identifier
UTRAN	Umts Terrestrial Radio Access Network
VoIP	Voice over IP
WCBQ	Weighted Class Based Queuing
XDMC	XML Document Management Client
XDMS	XML Document Management Server
XML	Extensible Mark up Language