Performance Analysis of Signalling System No. 7 Networks during Signalling Transfer Point Congestion

Amish Harkisan Chana

Submitted in fulfilment of the academic requirements for the degree of

Doctor of Philosophy

in the School of Electrical and Electronic Engineering at the
University of Natal
Durban, South Africa

December 2002

Supervisor: Prof. Fambirai Takawira

Preface

The research work presented in this thesis was performed by Mr. Amish Chana, under the supervision of Prof. Fambirai Takawira, at the University of Natal's School of Electrical and Electronic Engineering. This research was initially sponsored by Telkom S.A. Ltd. as part of the Telkom Teletraffic Initiative Programme (TTIP), and later became part of the research programme at the University of Natal's Centre for Radio Access Technologies, which is sponsored by Telkom S.A. Ltd. and Alcatel Altech Telecoms.

Parts of this thesis have been presented by the author at the Teletraffic `96 conference in Durban, South Africa, at the Teletraffic `97 conference in Grahamstown, South Africa and at the SATCAM 2000 conference in Somerset-West, South Africa. A paper has also been accepted for presentation at the WCNC 2003 conference in New Orleans, Louisiana, USA. Three journal papers have also been submitted for review to the journals, "IEEE Transactions on Vehicular Technology," and "Wireless Networks."

The whole thesis unless specifically indicated to the contrary in the text, is the author's own work, and has not been submitted in part, or in whole to any other university.

As the candidate's supervisor, I have approved this thesis for submission.

Name: .....

Date: .....

Signature: .....

Acknowledgements

Firstly, I wish to thank my supervisor, Prof. Fambirai Takawira for his invaluable advice, encouragement and guidance during the period of my postgraduate studies. His patience and support is also greatly appreciated. A great thanks is also owed to my mother for her patience, encouragement and support during the entire duration of my tertiary education.

Thanks is also owed to Telkom S.A. Ltd. and Alcatel Altech Telecoms for their generous financial support, and sponsorship towards the purchase of the necessary computing hardware and software for the completion of this degree. I would also like to acknowledge Vodacom South Africa for their contribution towards my SATCAM 2000 conference expenses.

Abstract

The growth of mobile networks and the imminent deployment of third generation networks and services will require signalling networks to maintain their integrity during increased unanticipated traffic volumes. As signalling networks become larger and more complex, an analysis of protocol operation is necessary to determine the effectiveness of the current protocol implementation and to evaluate the applicability of the proposed enhancements.

The objective of this study is to develop analytical models to analyse the impact of Signalling Transfer Point congestion on network performance when simple message discard schemes are used as the primary flow control mechanism, and to investigate suitable congestion and flow control mechanisms to help alleviate the congestion. Unlike previous studies, that are localised and only concentrate on the nodes around the congested entity, the models presented here examine the impact of network wide and focused overloads on the entire network. The study considers both the fixed-line and mobile network environments, and analyses the performance of the ISDN User Part and Mobile Application Part protocols. The call completion rate and location update success rate are used to measure performance, instead of message throughput, since these parameters provide a more appropriate measure of the grade-of-service and more accurately reflect the level of service provided to a customer.

The steady state equilibrium models, derived here, can be used to quickly estimate the safe operating regions of a signalling network, while the transient models provide a more intuitive perspective of the traffic processes that lead to congestion. Furthermore, these models can be used to examine the network performance for different message priority schemes, routing algorithms, overload scenarios and network configurations. The performance of various congestion control mechanisms that incorporate non-linear throttling schemes is also evaluated, together with an examination of the impact of congestion on multiple user parts in a mobile network environment.

Message priority schemes are found to offer little or no advantage in a fixed network environment, but in a mobile network they can be used to maintain the network's performance at an optimum level during periods of overload. Network performance is also improved if congestion controls block load-generating traffic at the initial onset of congestion and then gradually restore traffic as the performance improves.

Table of Contents

1. Introduction 1-1

1.1 General 1-1
1.2 Focus of this Study 1-1
1.3 Background on SS7 1-2
1.4 SS7 Network Architecture 1-3
1.5 SS7 Protocol Architecture 1-4
1.5.1 The Message Transfer Part 1-5
1.5.1.1 MTP level 1 (Signalling data link functions) 1-5
1.5.1.2 MTP level 2 (Signalling link functions) 1-5
1.5.1.3 MTP level 3 (Signalling network functions) 1-6
1.5.2 The User Parts 1-9
1.5.3 Evolution of Signalling Transport 1-10
1.6 ISDN User Part Call Set-up Procedure 1-11
1.7 Signalling in GSM Networks 1-12
1.8 Mobility Management Signalling 1-13
1.8.1 GSM Location Management Procedures 1-13
1.8.2 GSM Call Delivery Procedures 1-15
1.8.3 Mobility Management in Future GSM, GPRS and Third Generation Mobile Networks 1-15
1.9 Congestion and Flow Control in SS7 1-16
1.9.1 MTP Flow Control 1-17
1.9.1.1 International Option (IO) 1-17
1.9.1.2 National Option with congestion priorities (NOCP) 1-18
1.9.1.3 National Option without congestion priorities (NOWP) 1-19
1.9.2 ISDN User Part Congestion Control 1-20
1.9.3 SCCP Congestion Control 1-21
1.9.4 Comments on SS7 congestion and flow controls 1-22
1.9.5 STP congestion 1-22
1.10 Thesis Outline 1-23
1.11 Original Contributions of this Thesis 1-25

2. Performance Modelling and Evaluation of Signalling Protocols 2-1

2.1 Introduction 2-1
2.2 Congestion and Flow Control in Packet Switched Networks 2-2
2.2.1 Flow Control Levels 2-2
2.2.2 Congestion and Flow Control Schemes 2-3
2.2.2.1 Sliding Window Flow Control 2-3
2.2.2.2 Buffer Allocation Algorithms 2-3
2.2.2.3 Isarithmic Flow Control 2-4
2.2.2.4 Choke Packet Scheme 2-4
2.2.2.5 Other Congestion and Flow Control Methods 2-5
2.3 Performance Modelling of Multilayered Protocol Architectures 2-5
2.3.1 Hierarchical Decomposition and Aggregation 2-6
2.3.2 Iterative Decomposition of a Generic Model 2-8
2.4 Performance Modelling and Analysis of SS7 Networks 2-9
2.4.1 Link Level Congestion 2-9
2.4.2 Processor Congestion 2-11

2.4.3 Sustained Overloads *2-13*
2.4.4 Effectiveness of the TFC Procedure *2-13*
2.4.5 Multiple User Part Congestion Control Interactions *2-14*
2.4.6 Combined Control *2-15*
2.4.7 Link Level Queuing Delays in Signalling Networks *2-15*
2.4.8 Clustered Arrival Processes *2-17*
2.4.9 Traffic Characteristics *2-18*
2.4.10 Impact of New Services *2-18*
2.5 The Effect of Message Reattempts on Network Performance 2-19
2.6 Mobility Management Research 2-19
2.6.1 Signalling Load in Mobile Networks *2-19*
2.6.2 Memoryless Mobility Management Methods *2-20*
2.6.2.1 Hierarchical Database Architecture 2-20
2.6.2.2 Multiple Paging Areas per Location Area 2-22
2.6.2.3 Pointer Forwarding 2-22
2.6.2.4 Lightweight Call Delivery Protocols 2-22
2.6.3 Memory-Based Mobility Management Methods *2-23*
2.6.3.1 Caching Strategies 2-23
2.6.3.2 Long-term Profile Storage 2-23
2.6.3.3 Storage of User Mobility Patterns 2-25
2.7 Summary 2-25

3. Steady State Equilibrium Analysis of SS7 Network Performance in the PSTN 3-1
3.1 Introduction 3-1
3.2 The Prioritised Queuing Discipline 3-2
3.3 Equilibrium Analysis 3-5
3.4 Numerical Results and Discussion 3-13
3.4.1 Network Wide Overloads *3-14*
3.4.1.1 Single Discard Threshold 3-14
3.4.1.2 Multiple Discard Thresholds 3-18
3.4.1.3 Selection of the STP discard thresholds 3-22
3.4.1.4 Different Network sizes 3-23
3.4.1.5 The impact of caller reattempts and REL reattempts 3-24
3.4.2 Network failures *3-25*
3.4.3 Focused Overloads *3-27*
3.5 Summary 3-28

4. Transient Analysis of SS7 Network Performance in the PSTN 4-1
4.1 Introduction 4-1
4.2 Transient Mathematical Analysis 4-1
4.3 Numerical Results and Discussion 4-5
4.3.1 Network Wide Overloads *4-5*
4.3.1.1 Single Discard Threshold 4-6
4.3.1.2 Multiple Discard Thresholds 4-9
4.3.1.3 Different message priority schemes 4-11
4.3.1.4 Influence of threshold settings on network performance 4-12
4.3.1.5 The influence of call-holding times on network performance 4-14
4.3.2 Network Failures *4-15*
4.3.3 Focused Overloads *4-16*
4.3.4 Sensitivity of the meta-stable region to changes in the system parameters *4-17*
4.4 Summary 4-18

5. Steady State Equilibrium Analysis of SS7 Network Performance in the PLMN 5-1
5.1 Introduction 5-1
5.2 Steady State Equilibrium Analysis 5-1
5.2.1 GSM Location Management 5-3
5.2.2 GSM Call Delivery 5-6
5.2.3 Simple Location Update Protocol 5-8
5.2.4 Super-Charged Location Update Protocol 5-9
5.2.5 Lightweight Location Lookup Protocol 5-10
5.3 Numerical Results and Discussion 5-10
5.3.1 Single Discard Threshold 5-11
5.3.1.1 Location Management Performance 5-12
5.3.1.2 Call Delivery and Call Set-up Performance 5-14
5.3.1.3 Mobility Management and Call Set-up interaction 5-14
5.3.1.4 The impact of location update reattempts 5-15
5.3.2 Multiple Discard Thresholds 5-15
5.3.2.1 Location Management Performance 5-16
5.3.2.2 Call Delivery and Call Set-up Performance 5-17
5.3.3 Performance of the Simple Location Update and Super-Charged Location Update Protocols 5-18
5.3.4 The Lightweight Location Lookup Protocol (LiLLP) 5-19
5.4 Summary 5-20

6. Transient Analysis of SS7 Network Performance in the PLMN 6-1
6.1 Introduction 6-1
6.2 Transient Mathematical Analysis 6-1
6.2.1 GSM Location Management 6-2
6.2.2 GSM Call Delivery 6-5
6.2.3 Simple Location Update Protocol 6-8
6.2.4 Super-Charged Location Update Protocol 6-9
6.2.5 Lightweight Location Lookup Protocol 6-10
6.3 Numerical Results and Discussion 6-11
6.3.1 Single Discard Threshold 6-12
6.3.1.1 Location Management Performance 6-12
6.3.1.2 Call Delivery and Call Set-up Performance 6-14
6.3.1.3 The influence of call holding times on network performance 6-15
6.3.2 Multiple Discard Thresholds and Multiple User Part Interactions 6-16
6.3.3 Influence of Subscriber Profile Transfers on Super-Charged Networks 6-18
6.4 Summary 6-19

7. Congestion and Flow Control in SS7 7-1
7.1 Introduction 7-1
7.2 Implementation of SS7 Congestion and Flow Control Mechanisms 7-2
7.3 Comparison between the different Congestion and Flow control mechanisms 7-4
7.3.1 Network Wide Overload 7-5
7.3.2 Focused Overload with a single STP failure 7-8
7.4 Congestion Control in Multiple MTP User Parts 7-14
7.4.1 Implementation of SCCP Congestion Control 7-14
7.4.2 Effectiveness of Congestion Controls in GSM Networks 7-15
7.5 Optimisation of the Congestion and Flow Controls 7-18
7.5.1 Optimising the performance of the flow control procedure 7-18
7.5.1.1 The Impact of TFC Messages in a Network Wide Overload Scenario 7-18
7.5.1.2 The Impact of TFC Messages in a Focused Overload Scenario 7-19

7.5.2 The Influence of Flow control timers on network performance in the National Option without Congestion Priorities *7-20*
7.5.3 Optimisation of the ISUP congestion controls *7-21*
7.6 Summary 7-23

8. Conclusion 8-1

A. Appendix A-1
A.1 Algorithm for the Determination of the Steady-State Equilibrium Solution A-1
A.2 Routing Tables A-3
A.3 Simulation Modelling Tools A-4
A.3.1 Optimised Network Engineering Tools *A-5*
A.3.2 The SS7 Simulator *A-7*
A.3.2.1 Operational Description A-7
A.3.2.2 Model Scope and Limitations A-8
A.3.2.3 Signalling Point Architecture A-8
A.3.2.4 Signalling Transfer Point Architecture A-12

References R-1

Bibliography B-1

List of Figures

Chapter 1

Figure 1-1. Structure of a typical SS7 network in relation to a circuit switched network..... 1-3
Figure 1-2. Topology of the Singapore National SS7 network. 1-4
Figure 1-3. Architecture of Signalling System No. 7. 1-5
Figure 1-4. Basic format of a message signal unit 1-6
Figure 1-5. MTP level 3 – Signalling Network Functions 1-7
Figure 1-6. Routing label structure [ITU-T Recommendation Q.704]..... 1-7
Figure 1-7. Example of routing in a SS7 network..... 1-8
Figure 1-8. Message flow for a busy call. 1-11
Figure 1-9. Message flow for an answered call..... 1-11
Figure 1-10. Basic GSM network entities and their functional relationship with each other.. 1-12
Figure 1-11. Signalling Protocol Architecture used within a MSC..... 1-13
Figure 1-12. Message flow for a typical GSM location update procedure..... 1-14
Figure 1-13. Message flow for the retrieval of call routing information..... 1-15
Figure 1-14. Buffer threshold scheme implemented in the International Option..... 1-18
Figure 1-15. Buffer threshold scheme implemented in the National Option with congestion priorities..... 1-18
Figure 1-16. Example of the SRSCT procedure [Zepf & Rufa, 1994]..... 1-19
Figure 1-17. Example of signalling link congestion status transitions [ITU-T Recommendation Q.704]..... 1-20
Figure 1-18. ISUP congestion control mechanism [Zepf & Rufa, 1994]..... 1-21
Figure 1-19. SCCP step-wise congestion control mechanism..... 1-21
Figure 1-20. Idealised central processor based STP [Rumsewicz, 1993]..... 1-23

Chapter 2

Figure 2-1. Flow control levels [Gerla & Kleinrock, 1980]..... 2-3
Figure 2-2. Generic submodel and functional block diagram of MTP level 3 2-7
Figure 2-3. Interconnection of queuing submodels (adapted from [Conway, 1991]). 2-8
Figure 2-4. Network composed of open systems, entities and sources 2-9
Figure 2-5. Expected buffer occupancy during congestion control..... 2-10
Figure 2-6. Distributed hierarchical tree based database architecture..... 2-21
Figure 2-7. Location Lookup and Connection Set-up Message Flows for various protocols [Cui et al., 1998]..... 2-23
Figure 2-8. Graph showing the typical VLR Utilisation Characteristics (in terms of the number of subscribers registered on the VLR) for VLRs with city and residential coverage areas in a GSM network..... 2-24

Chapter 3

Figure 3-1. Priority thresholds of the M/M/1/K_1, K_2, K_3 queue. 3-3
Figure 3-2. State transition diagram of the M/M/1/K_1, K_2, K_3 queue..... 3-3
Figure 3-3. Example of routing in a simple network..... 3-7
Figure 3-4. SS7 network of fully connected STPs. 3-13
Figure 3-5. Total call completion rate at each SP region 3-14
Figure 3-6. Total message loads leaving each SP region and arriving at each STP..... 3-15

Figure 3-7. Individual message loads from each SP region. 3-16
Figure 3-8. Call completion rate for calls to the adjacent and non-adjacent regions. 3-17
Figure 3-9. Message throughput for traffic to the adjacent and non-adjacent regions. 3-18
Figure 3-10. Total end-to-end message transfer delay. 3-18
Figure 3-11. Expected number of messages in a STP queue 3-18
Figure 3-12. Total call completion rate at each SP region for the 01212_{ISUP} priority scheme ($K_1 = 100, K_2 = 120, K_3 = 130$). 3-19
Figure 3-13. Total call completion rate at each SP region for different multiple discard threshold schemes ($K_1 = 100, K_2 = 120, K_3 = 130$) 3-20
Figure 3-14. Expected number of messages in a STP queue 3-21
Figure 3-15. Message discard probabilities at a STP (for 01212_{ISUP} priority scheme with $K_1 = 100, K_2 = 120, K_3 = 130$). 3-21
Figure 3-16. Probability of discard, p_1, in a single priority scheme 3-22
Figure 3-17. Probability of discard, p_2, in a two priority scheme ($K_1 = 100$ messages and $K_2 = K_3$) 3-22
Figure 3-18. Probability of discard, p_3, in a three priority scheme ($K_1 = 100$ messages and $K_2 = 120$ messages) 3-22
Figure 3-19. Call completion rate at each SP region for various network sizes ($K_1 = K_2 = K_3 = 100$) 3-24
Figure 3-20. Call completion rates for various customer reattempt probabilities ($K_1 = K_2 = K_3 = 100$, r = 12/13). 3-25
Figure 3-21. Call completion rates for various REL reattempt probabilities ($K_1 = K_2 = K_3 = 100$, c = 0.7). 3-25
Figure 3-22. Call completion rate of each region when STP-1 has failed. 3-25
Figure 3-23. Call completion rate when STP-1 and STP-3 have failed (01212_{ISUP} priority scheme with $K_1 = 100, K_2 = 120, K_3 = 130$). 3-26
Figure 3-24. Call completion rate when STP-1, STP-2 and STP-3 have failed ($K_1 = K_2 = K_3 = 100$). 3-26
Figure 3-25. Call completion rate for a focused overload to SP region-1 (01212_{ISUP} priority scheme with $K_1 = 100, K_2 = 120, K_3 = 130$). 3-27
Figure 3-26. Call completion rate for a focused overload to SP region-1 (10212_{ISUP} priority scheme with $K_1 = 100, K_2 = 120, K_3 = 130$). 3-27

Chapter 4

Figure 4-1. Call arrival rate at each SP region (excluding reattempts). 4-6
Figure 4-2. Call completion rate for each SP region ($K_1 = K_2 = K_3 = 100$). 4-7
Figure 4-3. IAM message load from each SP region ($K_1 = K_2 = K_3 = 100$). 4-7
Figure 4-4. REL message load from each SP region ($K_1 = K_2 = K_3 = 100$). 4-7
Figure 4-5. STP buffer occupancy ($K_1 = K_2 = K_3 = 100$). 4-8
Figure 4-6. End-to-end message transfer delay ($K_1 = K_2 = K_3 = 100$). 4-8
Figure 4-7. Call completion rate for each SP region ($K_1 = 100, K_2 = 120, K_3 = 130$). 4-9
Figure 4-8. Occupancy of a STP's processing buffer 4-9
Figure 4-9. The number of REL and RLC messages in a STP processing buffer 4-10
Figure 4-10. Call completion rate for small first offered loads 4-11
Figure 4-11. Call completion rates for the 01212_{ISUP} and 00212_{ISUP} message priority schemes 4-11
Figure 4-12. ACM message occupancy in a STP for the 01212_{ISUP} and 00212_{ISUP} message priority schemes. 4-11
Figure 4-13. Call completion rate for the 10212_{ISUP} message priority scheme. 4-12
Figure 4-14. STP buffer occupancy for the 10212_{ISUP} message priority scheme. 4-12
Figure 4-15. Call completion rate for various sizes of K_1, in a single priority scheme. 4-13
Figure 4-16. Call completion rate for various sizes of K_2, in a two priority scheme ($K_1 = 100$) 4-13

Figure 4-17. Call completion rate for various sizes of K_3, in a three priority scheme (K_1 = 100, K_2 = 120)..... 4-13
Figure 4-18. REL message load for various sizes of K_3, in a three priority scheme (K_1 = 100, K_2 = 120)..... 4-13
Figure 4-19. Call completion rate for different mean call holding times..... 4-14
Figure 4-20. Call completion rate for an overload of 20 calls/sec with a duration of 500 seconds when STPs 1 and 3 have failed..... 4-15
Figure 4-21. Call completion rate for a focused load to SP region-1 when STPs 1 and 3 have failed (01212_{ISUP} priority scheme with K_1 = 100, K_2 = 120, K_3 = 130)..... 4-16
Figure 4-22. Call completion rate for a single STP network, with K_1 = 100, commencing from the meta-stable region. Different call completion rates are obtained depending on the type of variables used in the program code..... 4-17

Chapter 5

Figure 5-1. GSM signalling network architecture..... 5-11
Figure 5-2. Location update success rate for each region..... 5-12
Figure 5-3. Accuracy of each HLR for various offered loads..... 5-13
Figure 5-4. Location cancellation success rate for each region..... 5-13
Figure 5-5. Individual message loads from each region..... 5-13
Figure 5-6. Total message load leaving each region..... 5-13
Figure 5-7. Call success rate for each region..... 5-14
Figure 5-8. Message loads from each region..... 5-14
Figure 5-9. Location update success rate for various location update offered loads and call arrival rates..... 5-15
Figure 5-10. Call completion rate and roaming number retrieval success rate for various location update offered loads..... 5-15
Figure 5-11. Location update success rate for various attempt counter values..... 5-15
Figure 5-12. Location update success rate for each region (K_1 = 100, K_2 = 120, K_3 = 130)..... 5-16
Figure 5-13. Location cancellation success rate for each region..... 5-17
Figure 5-14. HLR accuracy in each region (K_1 = 100, K_2 = 120, K_3 = 130)..... 5-17
Figure 5-15. Roaming number retrieval success rate at each region for different priority schemes 5-17
Figure 5-16. Call completion rate at each region for different priority schemes (K_1 = 100, K_2 = 120, K_3 = 130)..... 5-17
Figure 5-17. Location update success rate comparison between GSM, the Simple Location Update and Super-Charged Location Update Protocols..... 5-19
Figure 5-18. Comparison between the GSM and LiLLP protocols..... 5-20

Chapter 6

Figure 6-1. Location update success rate for each SP region ($K_1 = K_2 = K_3$ = 100)..... 6-13
Figure 6-2. Location cancellation success rate for each SP region ($K_1 = K_2 = K_3$ = 100)..... 6-13
Figure 6-3. HLR update rate for each SP region ($K_1 = K_2 = K_3$ = 100)..... 6-14
Figure 6-4. Roaming number retrieval success rate and call completion rate for each SP region ($K_1 = K_2 = K_3$ = 100)..... 6-14
Figure 6-5. Occupancy of a STP's processing buffer ($K_1 = K_2 = K_3$ = 100)..... 6-15
Figure 6-6. Call completion rate for different mean call holding times..... 6-15
Figure 6-7. REL message load for different mean call holding times..... 6-16
Figure 6-8. Location update success rate (K_1 = 100, K_2 = 120, K_3 = 130)..... 6-17
Figure 6-9. Roaming number retrieval success rate and call completion rate..... 6-18
Figure 6-10. Influence of a momentary increase in δ on a Super-Charged network..... 6-18

Chapter 7

Figure 7-1. Buffer occupancy at STP-1 for IO_2 7-6
Figure 7-2. Buffer occupancy at STP-1 for IO_3 7-6
Figure 7-3. Number of TFC messages in a STP buffer (for IO_2 and IO_3). 7-7
Figure 7-4. Buffer occupancy at STP-2 for IO_2, during a focused overload. 7-10
Figure 7-5. Call completion rate at region-1 for various congestion and flow control schemes, during a focused overload to this region. 7-11
Figure 7-6. Congestion level in STP-2 for the NOCP and NOWP control schemes 7-13
Figure 7-7. Congestion level in STP-2 for the IO_2 and IO_5 control schemes. 7-13
Figure 7-8. Number of REL messages generated by various congestion and flow control schemes, during a focused overload. 7-13
Figure 7-9. Comparison between the location update success rates of different congestion control schemes in region-1. 7-16
Figure 7-10. Comparison between the call completion rates obtained for TFC transmission intervals of $n = 1$ and $n = 8$ in the IO_2 control scheme. 7-18
Figure 7-11. Comparison between the call completion rates obtained for TFC transmission intervals of $n = 1$ and $n = 8$ in the NOCP control scheme. 7-18
Figure 7-12. Comparison between the call completion rates obtained for TFC transmission intervals of $n = 1$ and $n = 8$ in the NOCP control scheme. 7-20

Appendix

Figure A-1. Traffic distribution for the network examined. A-3
Figure A-2. A simulation model in the network domain. A-5
Figure A-3. An example of the layered structure of process modules in a signalling point A-6
Figure A-4. Example of a state transition diagram and the code associated with one of the states A-6
Figure A-5. Node model for the Signalling Points. A-9
Figure A-6. Process model for the MTP level 3 message discrimination and distribution functions in a Signalling Point. A-9
Figure A-7. Process model for the MTP level 3 message routing function in a Signalling Point. A-10
Figure A-8. Process models for the new call traffic generators. A-11
Figure A-9. User part process models. A-11
Figure A-10. Node model for the Signalling Transfer Points. A-12
Figure A-11. Process model for the MTP level 3 functions in a Signalling Transfer Point. A-13

List of Acronyms

3G	Third Generation
3GPP	Third Generation Partnership Project
ACC	Automatic Congestion Control
ACM	Address Complete Message
ANM	Answer Message
ANSI	American National Standards Institute
ATM	Asynchronous Transfer Mode
BCMP	Baskett, Chandy, Muntz & Palacois
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CC	Combined Control
CCS	Common Channel Signalling
CI	Congestion Indication
CL	MAP Cancel Location Message
CLA	MAP Cancel Location Acknowledgement Message
CLM	Congested Link Method
DPC	Destination Point Code
DSS1	Digital Subscriber Signalling System No. 1
ETSI	European Telecommunications Standards Institute
FCFS	First-come First-serve
FISU	Fill-in Signal Unit
GLR	Gateway Location Register
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HLR	Home Location Register
IAM	Initial Address Message
IBL	Input Buffer Limit
IETF	Internet Engineering Task Force
IN	Intelligent Network
IO	International Option
IP	Internet Protocol
ISD	MAP Insert Subscriber Data Message
ISDA	MAP Insert Subscriber Data Acknowledgement Message
ISDN	Integrated Services Digital Network
ISUP	ISDN User Part
ITU-T	International Telecommunications Union – Telecommunications Sector
LiLLP	Lightweight Location Lookup Protocol
LSSU	Link Status Signal Unit
LU	Location Update
MAP	Mobile Application Part
MSC	Mobile Switching Centre
MSU	Message Signal Unit
MT	Mobile Terminal
MTP	Message Transfer Part
NOCP	National Option with congestion priorities
NOWP	National Option without congestion priorities
NSS	Network and Switching Subsystem

OPC	Originating Point Code
OSI	Open Systems Interconnection
PLMN	Public Land Mobile Network
PRN	MAP Provide Roaming Number Message
PRNA	MAP Provide Roaming Number Acknowledgement Message
PSTN	Public Switched Telephone Network
PVLR	Previous VLR
RCT	Route set Congestion Test
REL	Release Message
RL	Restriction Level
RLC	Release Complete Message
RSL	Restriction Sublevel
SCCP	Signalling Connection Control Part
SCP	Service Control Point
SGSN	Serving GPRS Support Node
SIB	Status Indication Busy
SIF	Signalling Information Field
SIGTRAN	IETF Signalling Transport
SIO	Service Information Octet
SLS	Signalling Link Selection
SLU	Simple Location Update
SP	Signalling Point
SRI	MAP Send Routing Information Message
SRIA	MAP Send Routing Information Acknowledgement Message
SRSCT	Signalling-route-set-congestion-test
SS7	Signalling System No. 7
SSP	Service Switching Point
STP	Signalling Transfer Point
TC	Transaction Capabilities
TCP	Transmission Control Protocol
TFC	Transfer Controlled
TFP	Transfer Prohibited
TFA	Transfer Allowed
UL	MAP Update Location Message
ULA	MAP Update Location Acknowledgement Message
UUI	User-to-user Information
VLR	Visitor Location Register

1. Introduction

1.1 General

Signalling System No. 7 (SS7) is the life-blood of modern telecommunications networks. Common Channel Signalling (CCS) was conceived in the late 1970s with the intention to provide faster call set-up times and enhanced security. Since then signalling protocols have evolved to support integrated services, intelligent network services, mobility management and messaging services. The next decade will see SS7 support the widespread deployment of bearer-independent and transport-independent signalling protocols in next generation telecommunications networks, only to be rivalled by emerging Internet protocols.

The Signalling System No. 7 protocol allows for the exchange of messages related to call control, database transactions and management information between network elements in a telecommunications network. Unlike traditional packet switched networks, SS7 is designed to provide a high level of reliability and performance under large traffic volumes, and particularly during failure and overload conditions. Congestion and flow controls therefore play a pivotal role in maintaining the integrity of the signalling network under adverse conditions.

The growth of mobile networks and the imminent deployment of third generation (3G) networks and services will require signalling networks to maintain their integrity during increased unanticipated traffic volumes. The distributed processing architecture of mobile networks and emerging 3G services creates traffic patterns that were previously not observed in the traditional telecommunications environment. Coupled with signalling network outages over the past decade, future signalling protocol enhancements will have to focus on network design procedures, network reliability in the context of emerging services, improved congestion and flow control strategies, and a more thorough analysis of protocol operation under a wide range of implementation scenarios [Bolotin et al., 1994].

The current SS7 congestion control schemes implicitly assume that transmission bandwidth, rather than processing speed, is the primary bottleneck [Kant & Ong, 1997]. But, the growing complexity of new services and the imminent widespread deployment of broadband networks, to support the target 3G services, has shifted attention towards the effect of processing overheads on system performance, and the necessity for more efficient flow control mechanisms to prevent processor overloads.

As the world strives towards a ubiquitous integrated services high-speed digital network, the network signalling protocols will have to evolve to satisfy the new service requirements.

1.2 Focus of this Study

There exists limited literature on the performance evaluation of large SS7 networks. Previous studies have either analysed the network elements in isolation or concentrated on congestion of a single link or node, typically in a scenario where one node of a mated signalling transfer point (STP) pair has failed [Skoog, 1988]. These studies do not consider the overall network structure and model all the sources as being directly connected to a single STP, which is the intermediate node during a focused overload [Rumsewicz, 1994]. In addition, previous studies [Smith, 1994] and protocol standards [ITU-T Recommendation Q.704] have also concentrated on transmission bottlenecks that result in link level congestion. However, the trend towards the integration of broadband signalling links [ITU-T Recommendation Q.2210] and high-speed Internet Protocol (IP) based signalling links [IETF RFC 2719] into current and future networks aims to address

the bandwidth bottleneck and shortcomings of narrowband signalling links. This may consequently lead to processor capacity becoming the primary bottleneck during periods of unexpected traffic peaks. Furthermore, previous studies have focused on SS7 congestion scenarios in a Public Switched Telephone Network (PSTN) environment and no work exists on the impact of STP congestion on mobile networks, where signalling traffic is dominated by mobility management messages.

The focus of this study is to develop analytical models to analyse the impact of STP congestion on the performance of a multiple STP signalling network when no feedback control mechanisms are present, and to investigate suitable congestion and flow control mechanisms to help alleviate STP processor congestion.

The mathematical analysis assumes that the feedback control mechanisms are either not implemented or are not invoked during congestion, instead the excess signalling traffic is simply discarded when the buffer resources are exhausted. The steady state equilibrium models, derived here, principally provide a means of quickly estimating the safe operating regions of a signalling network, while the transient models provide a more intuitive perspective of the traffic processes that eventually lead to network congestion. These models consider the impact of application level processes on network performance by explicitly modelling the transfer of individual signalling messages. In addition, one is also able to calculate various performance measures, such as the call completion rate, location update success rate, message throughput, and the delay for various first offered loads. The network models developed here can also be used to:

a) determine the maximum number of calls sustainable by a signalling network,
b) determine the maximum number of location updates sustainable by a signalling network,
c) define criteria for the selection of buffer thresholds,
d) examine network performance when different message priority schemes are adopted,
e) analyse network performance during normal and failure conditions, and
f) examine the impact of different overload scenarios on network performance.

An investigation of various congestion and flow control mechanisms is also performed, in order to ascertain which control mechanisms are effective and robust over a wide range of implementation scenarios. Previous studies on the efficacy of SS7 congestion and flow controls show that they perform poorly in some scenarios and are completely ineffective in a multiple user part environment ([Mayer, 1997] and [Zepf & Rufa, 1994]). The user part congestion control mechanism used in most studies is based on a simple step-wise throttling scheme. However, ITU-T Recommendation Q.764 states that this aspect of the user part congestion controls is considered as implementation dependent. But, researchers have overlooked this point, in favour of the well-known linear step-wise throttling scheme. This study examines various congestion control strategies, some of which are found to outperform the traditional control schemes. The national option without congestion priorities, which to date has not been examined by other researchers, is also considered. Some researchers have also noted that the probe packets sent by the signalling route set congestion test procedure might not necessarily travel on the route at which congestion was detected [Willmann & Kühn, 1990]. These unexplored areas require further research, and are therefore also investigated.

1.3 Background on SS7

Signalling System No. 7 is a common channel signalling (CCS) protocol, which is primarily used in modern telecommunications networks to carry signalling messages on a packet-switched network that is logically independent of the underlying circuit-switched trunk network. The International Telecommunications Union – Telecommunications Sector (ITU-T) formulates the international protocol specifications for SS7. These standards are meant to be a guide for the development of national signalling protocols and define end-to-end interoperability within the

international network. For example, the American National Standards Institute (ANSI) is responsible for the development of SS7 standards for the North American telecommunications networks. Similarly, a number of European countries have also defined country specific variants. The national variants are similar to the ITU-T Recommendations in most respects, but they also address country specific requirements and procedures that are not covered or are referred to as implementation dependent in the ITU-T Recommendations.

1.4 SS7 Network Architecture

A simple SS7 network consists of two types of nodes called Signalling Points (SPs) and Signalling Transfer Points (STPs), which are connected together with signalling links.

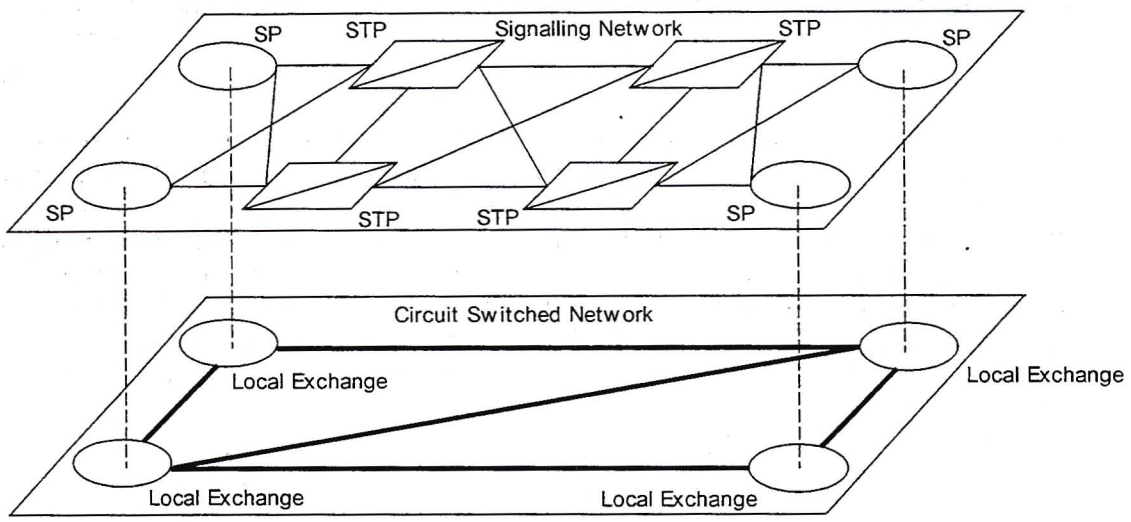


Figure 1-1. Structure of a typical SS7 network in relation to a circuit switched network.

Signalling points act as traffic sources and sinks for call related signalling messages and are connected to each other through the underlying trunk network. Signalling points with additional functionality, beyond those of simple call control, are often referred to as Service Switching Points (SSPs) and Service Control Points (SCPs). A SSP is a SP that is capable of launching database queries to SCPs and then interpreting the received responses, while a SCP is responsible for the retrieval of information from databases, such as credit information required for pre-paid billing.

Signalling Transfer Points act as packet switches, which determine the next link and node to which a signalling message is to be routed. Interconnectivity between international networks and networks operated by different network operators is usually through gateway STPs. Gateway STPs provide protocol conversion procedures, usually from the national implementation of SS7 to the ITU-T compliant protocol standard. They also maintain network security and provide message-screening capabilities.

Any two signalling points that are capable of communicating, by the exchange of signalling information, are said to have a *signalling relation*. The path taken by these messages and how they are routed determines the signalling mode of operation between the source and destination SPs. There are three possible signalling modes:

- Associated signalling – This is the simplest mode of operation. Here signalling messages exchanged between two SPs are transferred over a link set that directly connects the two nodes.
- Non-associated signalling – Here messages are transferred via one or more signalling transfer points to the destination node. There is no preferred or predetermined route taken by the messages.

- Quasi-associated signalling – This mode is a specific type of the non-associated mode where the path taken by signalling messages is predetermined and fixed, except during failures. In most implementations, messages usually traverse the least number of STPs between the source and destination nodes.

Signalling networks generally have a mesh structure. The signalling network structure depicted in Figure 1-1 is referred to as a STP quad structure and is usually the basis of more complex network structures. The STPs in this structure are mated (or paired) in each physical region, to allow for a 100% redundancy. The paired configuration allows for traffic to be diverted on to an alternate route, during single link set or node failures, without increasing the number of STPs traversed. To accommodate for traffic loads that can be experienced during failures the mated STPs are engineered to handle twice the peak load expected under normal conditions. In order to meet availability and diversity requirements, each STP has to be located at different physical locations and the link sets to each STP should be allocated physically diverse transmission paths.

Other possible configurations generalise the mesh structure in Figure 1-1 to form a backbone of fully connected STPs in a larger network environment. The Singapore National network, shown in Figure 1-2, is a typical example. The network consists of six fully connected STPs and six SP regions (or network clusters). The SPs in each region are connected to the two neighbouring STPs [Lazar et al., 1994]. Unlike the quad STP structure the STPs in this network only require a 50% reserve capacity during single STP failures (assuming that the load from each region is more or less equal). For example, if STP 1 fails, half of its load would be transferred to STP 2 and the other half would be transferred to STP 6.

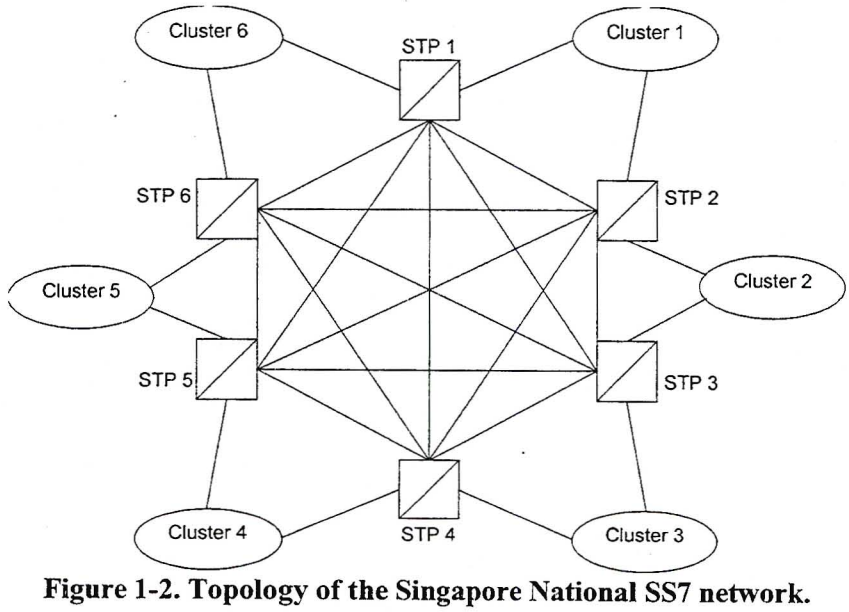


Figure 1-2. Topology of the Singapore National SS7 network.

Signalling networks may also consist of a mixture of signalling relations in associated and quasi-associated modes, where the associated mode is used as the primary route and the quasi-associated mode is used as a backup in case the primary route fails. Additional examples of alternate signalling network structures are given in Goldberg & Shrader [1990] and Modarressi & Skoog [1990].

1.5 SS7 Protocol Architecture

Developed around the same time as the Open Systems Interconnection (OSI) reference model, SS7 is based on a layered protocol architecture. The SS7 architecture was originally developed

for a telecommunications environment, whereas the OSI reference model was developed for a data communications environment, and thus lacked explicit definitions for the Transport, Session and Presentation layers. The trend towards integrated services in modern telecommunications networks has thus made it necessary to align SS7 more closely with the OSI model. Figure 1-3 illustrates the functional relationship between the various SS7 levels and how their associations relate to the OSI reference model.

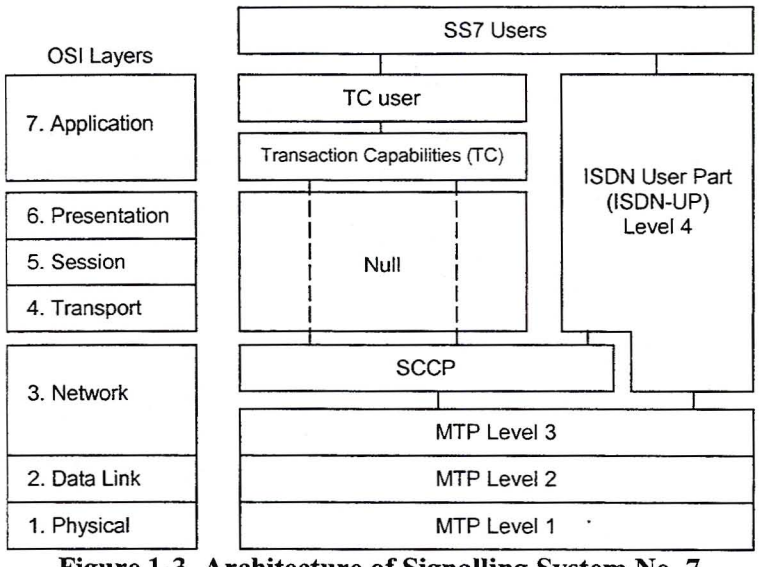


Figure 1-3. Architecture of Signalling System No. 7.

1.5.1 The Message Transfer Part

The message transfer part (MTP) is responsible for the reliable transfer of signalling messages from source to destination. In addition, the MTP has to take appropriate actions in response to network failures in order to ensure that message throughput is maintained according to the required performance specifications defined in ITU-T Recommendation Q.706. The MTP can be separated into three functional levels (shown in Figure 1-3):

1.5.1.1 MTP level 1 (Signalling data link functions)

Level 1 [ITU-T Recommendation Q.702] is the physical level and defines the physical, electrical and access characteristics of the signalling data link. A digital 64 kb/s (or 56 kb/s in North America) bi-directional data link derived from a timeslot in a 2.048 Mb/s E1 (or 1.544 Mb/s T1 in North America) is commonly used.

1.5.1.2 MTP level 2 (Signalling link functions)

Level 2 [ITU-T Recommendation Q.703] defines the functional procedures used to transfer signalling messages across a data link, connecting two nodes. Signalling messages are transferred over the signalling links in variable length packets known as *signal units*. There are three types of signal units:

- The message signal unit (MSU) is used to transfer signalling messages from the level 4 user parts or management information from the level 3 signalling network management part.
- The link status signal unit (LSSU) is used to synchronise and indicate the status of a signalling link.
- The fill-in signal unit (FISU) is sent continuously when there is no other signalling traffic to transfer.

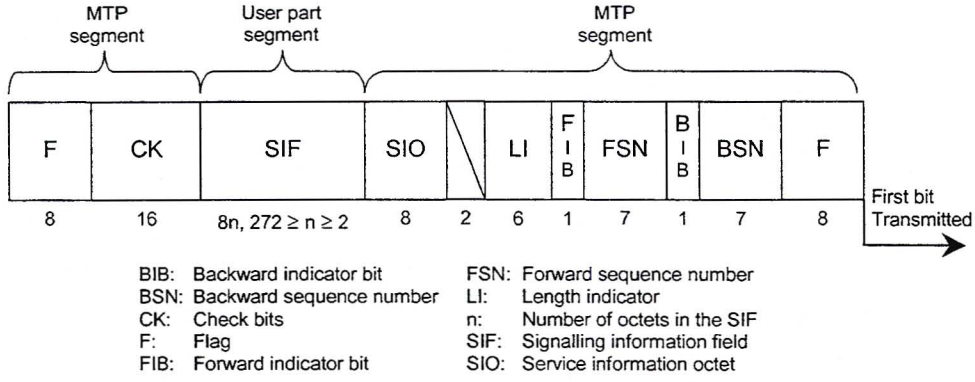


Figure 1-4. Basic format of a message signal unit [adapted from ITU-T Recommendation Q.703].

To ensure the reliable transfer of signalling messages between directly connected nodes, the following functions are defined for level 2:

- Signal unit delimitation and alignment
- Error detection/correction
- Initial alignment
- Signalling link error monitoring
- Flow control

1.5.1.3 MTP level 3 (Signalling network functions)

The MTP level 3 functions [ITU-T Recommendation Q.704] are equivalent to the lower half of the OSI network layer. MTP level 3 is responsible for network management and has to ensure the reliable transfer of signalling messages from source to destination, as well as during link and node failures. The signalling network functions can be grouped into two distinctive categories (Figure 1-5), namely:

- *Signalling message handling functions* – which handle the routing of messages through the signalling network, and
- *Signalling network management functions* – which control management and reconfiguration of the signalling network.

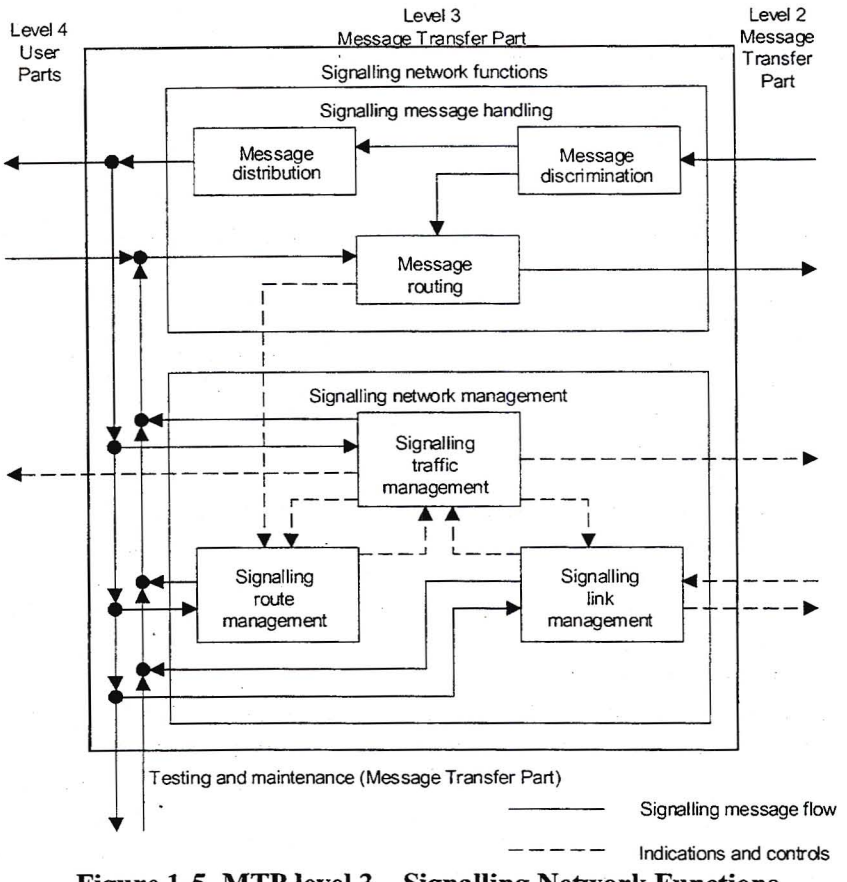


Figure 1-5. MTP level 3 – Signalling Network Functions [ITU-T Recommendation Q.704].

a) Signalling Message Handling: Signalling message handling includes three functions – message discrimination, message distribution and message routing. These functions deliver signalling messages to the appropriate user part, management function or outgoing signalling link. Routing is accomplished by examining the routing label (Figure 1-6) and service information octet (SIO) assigned to each signalling message. Signalling points and signalling transfer points are identified by a unique 14-bit address know as a *point code*. Each signalling message is assigned an originating point code (OPC) and a destination point code (DPC), which identify the source and destination nodes of the message, respectively.

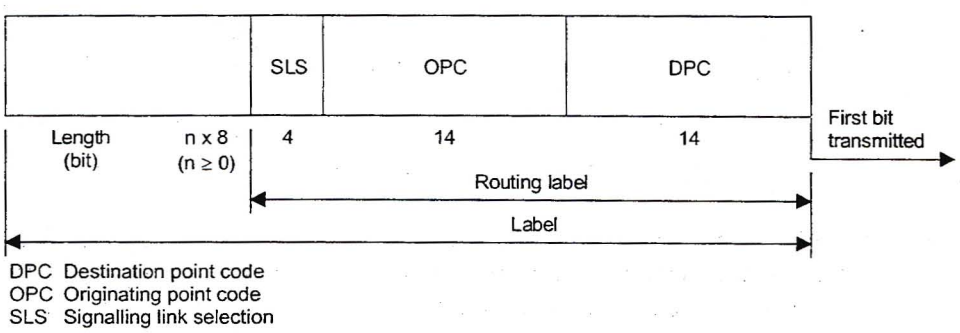


Figure 1-6. Routing label structure [ITU-T Recommendation Q.704].

The message discrimination function uses the DPC to determine if messages received from level 2 are addressed to itself or another node. If the message is destined to another node and the signalling point has transfer capabilities, then the message is sent to the message routing function. If the receiving node is also the destination of the message, then the message is forwarded to the message distribution function. The message distribution function uses the SIO to deliver the message to the appropriate user part or level 3 management function. The message

routing function uses the DPC of a message to select the appropriate outgoing link. If more than one signalling link can route the message to its destination, then the 4-bit SLS field is used to load share traffic across link sets and across all the links within a link set. Load sharing attempts to distribute traffic evenly across all possible routes and links.

Link No.	SLS Code
0	0000, 0100, 1000, 1100
1	0001, 0101, 1001, 1101
2	0010, 0110, 1010, 1110
3	0011, 0111, 1011, 1111

Table 1-1. Example of SLS code assignments for four signalling links.

Link No.	SLS Code
0	0000, 0101, 1010, 1111
1	0001, 0110, 1011
2	0010, 0111, 1100
3	0011, 1000, 1101
4	0100, 1001, 1110

Table 1-2. Example of SLS code assignments for five signalling links.

A link set is a collection of signalling links between two signalling points, and a combined link set is a collection of one or more link sets. A signalling route set refers to a collection of routes that exist between the originating and destination point. In order to achieve an even traffic distribution across all the links within a link set, the number of links has to be a power of two (i.e. 1, 2, 4, 8 or 16). Tables 1-1 and 1-2 show examples of SLS code assignments for link sets of four and five signalling links, respectively. In the link set with five signalling links, link 0 carries 25% of the traffic while the other links carry 18.75% of the total traffic. An example of routing with SLS codes is illustrated in Figure 1-7. The links represented by the dashed lines are only used to route traffic when the primary signalling links fail.

All the messages that are transmitted for a particular circuit connection are assigned the same routing label. However, messages from the destination do not necessarily have to return on the same route as that used in the forward direction, as long as the same route is maintained for the entire duration of the transaction.

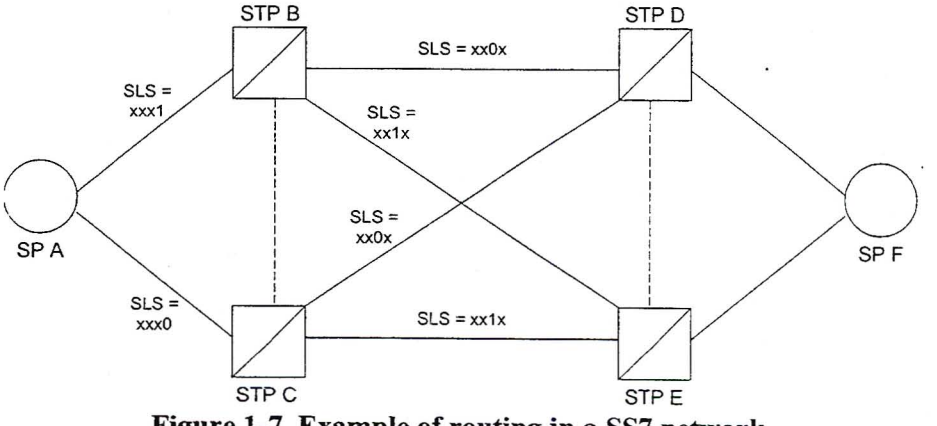


Figure 1-7. Example of routing in a SS7 network.

b) Signalling network management: Signalling network management includes three functions – signalling traffic management, signalling route management and signalling link management. These functions respond to link and node status changes, by reconfiguring the network during failures and controlling traffic flow during congestion.

Signalling link management provides procedures necessary for the activation of new signalling links or link sets, restoration of failed signalling links and the deactivation of a signalling link when the error rate is too high. The link management function also informs the traffic management and route management functions of link status changes, to allow them to re-route traffic.

The signalling route management function conveys the network status information necessary to control the accessibility of signalling routes. The following procedures are defined:

- The transfer-controlled (TFC) procedure indicates signalling link congestion at a STP.
- The transfer-prohibited (TFP) and transfer-restricted (TFR) procedures inform adjacent nodes that traffic to a particular destination should not be routed through that STP.
- The transfer-allowed (TFA) procedure informs adjacent nodes that routing through a STP (which previously sent out a TFP or TFR message) has returned to normal.
- The signalling-route-set-test procedure is used by SPs that previously received TFP and TFR messages, to retrieve updated information on signalling route availability.
- The signalling-route-set-congestion-test (SRSCT) procedure is used to update the congestion status information associated with a particular destination.

The signalling traffic management function diverts traffic depending on the availability or unavailability of a link or route. For example, if a signalling link becomes unavailable (possibly due to a failure) the changeover procedure redirects traffic to alternate signalling links, were possible. This function attempts to redirect traffic without causing message loss or the duplication of MSUs in the transmission and retransmission buffer of the failed link. When the signalling link becomes available, the changeback procedure is used to restore its signalling traffic load.

1.5.2 The User Parts

Level 4 is made up of several different protocols called User Parts or Application Parts, and is only found in signalling points. A STP is functionally equivalent to a message router and therefore does not have application level user parts. However, STPs may be equipped with SCCP to allow for global title address resolution and for routing to remote databases. The following is a brief description of some of the MTP user parts.

Signalling Connection Control Part (SCCP) [ITU-T Recommendations Q.711-Q.716]
The function of SCCP is to provide end-to-end routing since MTP is only capable of providing connectionless point-to-point routing. Although SCCP is a level 4 user part, it supplements the functions of MTP and together they provide full OSI layer 3 functionality. SCCP provides four classes of service:

- Class 0: Basic connectionless service
- Class 1: Sequenced connectionless service
- Class 2: Basic connection-orientated service
- Class 3: Flow control connection-orientated service

SCCP also allows for messages to be addressed in terms of global titles. Global title routing is used when a node does not have the point code of a destination (e.g. when a mobile subscriber's ISDN number is used to access his/her HLR). This is accomplished by using the global title translation function to map the address digits (e.g. a mobile subscriber's ISDN number) to a point code and/or subsystem number in order to be able to route a message to a subsequent STP for further translation in its SCCP or to the final destination via MTP routing [ITU-T Recommendations Q.714]. By separating the SCCP functions from MTP level 3; the routing overhead added to signalling messages that only require MTP transfer capabilities, is reduced.

ISDN User Part (ISUP) [ITU-T Recommendations Q.761-Q.769]
ISUP is used to establish and clear circuit switched connections associated with voice or data calls and supports the supplementary services provided by ISDN. ISUP is compatible with the ISDN protocol (Digital Subscriber Signalling System No. 1 or DSS1). DSS1 can be considered as being an extension of SS7 to the subscriber. Even though DSS1 and ISUP messages are not the same, there is a direct mapping between their respective message types. ISUP uses the services of both MTP and SCCP to send signalling messages to their destination. A description of the ISUP call set-up procedure is given in Section 1.6.

Transaction Capabilities (TC) [ITU-T Recommendations Q.771-Q.775]
This protocol is designed to support applications that require the exchange of non-call set-up related signalling messages such as database transactions and information transfer between application level entities. In mobile networks, the Mobile Application Part uses the TC services to exchange information between the Home Location Registers, Visitor Location Registers and the Mobile Switching Centres.

1.5.3 Evolution of Signalling Transport

The high volume of mobility management traffic in mobile networks and the increased use of intelligent network (IN) services have highlighted some of the limitations of traditional SS7 transport protocols. These include:

- Limited bandwidth: The 4-bit SLS code limits the maximum bandwidth available to a route-set to 1024 kb/s (since the 64kb/s timeslots of an E1 are used as signalling links). The bandwidth available is lower, if signalling links also carry traffic to more than one destination (e.g. in the quasi-associated mode).
- Restriction in the size of MSUs: The Signalling Information Field (SIF) of MSUs (Figure 1-4) has a maximum size of 272 octets. Large user data fields therefore need to be segmented and re-assembled by SCCP [ITU-T Recommendation Q.715], which imposes an additional processing overhead on SCCP. In GSM networks the Mobile Application Part in a Home Location Register segments the subscriber profile into multiple user data messages and waits for an equivalent number of acknowledgements from the Visitor Location Register before the profile transfer is considered to be successful. These segmentation operations also impose an additional processing overhead on MTP level 3, since a large number of small MSUs have to be routed instead of one large MSU (the size of the MSU has a negligible impact on the processing speed of MTP level 3 in current SP and STP implementations).
- Head-of-line blocking: Large SCCP messages increase the queuing delay experienced by short ISUP messages [Kosal & Skoog, 1994] on low-speed signalling links. Furthermore, if the receiver detects an error in a large MSU, the transmitting end retransmits the corrupted MSU and all the subsequent MSUs, even though the subsequent MSUs may have been delivered correctly the first time and they may also not be part of the same transaction as the affected MSU.

To address the bandwidth limitations of SS7, the ITU-T has completed a specification for the transport of SS7 messages over high-speed signalling links based on ATM, in ITU-T Recommendation Q.2210. This specification also allows for SIFs of up to 4095 octets to be transported over signalling links.

Work is also currently in progress within the IETF's Signalling Transport Working Group (SIGTRAN), to define protocols for the transport of SS7 messages over IP networks. In the SIGTRAN protocol suite, the SS7 transport layer protocols are replaced by equivalent adaptation layer protocols [IETF RFC 2719]. The adaptation layer provides the functions necessary to transport SS7 over the Stream Control Transmission Protocol and IP. Some of the features available, include [IETF RFC 2960]:

- no limitation in the size of messages,
- sequenced delivery of messages across multiple streams to overcome head-of-line blocking,
- unordered or unsequenced delivery of messages, and
- network fault tolerance through the support for multi-homed end-points.

1.6 ISDN User Part Call Set-up Procedure

When a call arrives from a customer, the originating exchange selects a voice circuit and then sends an initial address message (IAM) to the destination exchange via the SS7 network. The IAM provides the destination with information such as the calling and called party numbers, protocol/service requirements and it indicates whether further information will be available through subsequent address messages.

When the destination exchange receives the IAM the dialled digits are examined and the called party's line is interrogated to determine its status and to check if the requested service is available. If the called party is busy, a release message (REL) is sent to the originating exchange, where the circuit is immediately made available for other calls (Figure 1-8). If the called party is not busy and the call can be accepted, then an address complete message (ACM) is returned as an acknowledgement (Figure 1-9). At the same time the destination informs the called party that a call has arrived, by generating a ringing tone on the called subscriber's line. Upon receipt of the ACM, the originating exchange sends a ring-back tone to the calling party. This method of applying tones allows the telephone company to leave voice circuits disconnected (but reserved) until the call is answered. If the called telephone is not answered or is busy, the circuit can be immediately released and used for another call.

When the call is answered the destination exchange sends an answer message (ANM). The voice circuit is immediately cut through once the originating exchange receives the ANM. Once the voice circuit is established, the conversation can begin between both parties and no further messages are exchanged until either party terminates the call. However there are some features associated with ISDN supplementary services that may require exchanges to share information during the duration of the call. User-to-user information (UUI) messages and UUI fields in other signalling messages, for example, are used to transport the data and information transparently through the signalling network.

At call termination, when either party hangs-up, a REL is sent to the other party's exchange. When an exchange receives a REL message the voice circuit is returned to the idle condition and the release complete message (RLC) is returned as an acknowledgement. If any of the messages exchanged during this transaction exceed the maximum allowed size of 272 octets, a segmentation message is used to transfer the additional segment of the over-length message.

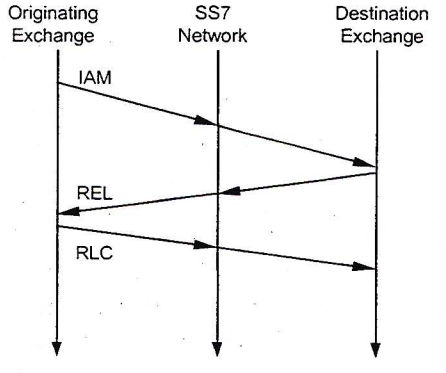


Figure 1-8. Message flow for a busy call.

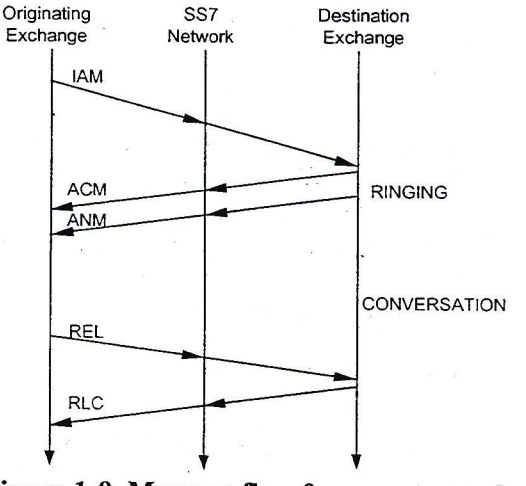


Figure 1-9. Message flow for an answered call.

Call set-up recovery actions are implemented in ISUP to ensure that messages are not lost or arrive out of sequence during a transaction, and to ensure that the expected responses arrive within a specified period of time. When an IAM is sent by the originating exchange, an

awaiting address complete timer (T7) is started. If this timer expires before receipt of either an ACM or ANM, the connection release procedure is started and a call-failed indication is returned to the calling party. This situation normally occurs when an IAM or ACM is discarded in a congested network node or link. If the ACM arrives before T7 expires, then T7 is stopped and the awaiting answer timer (T9) is started. If the ANM does not arrive before T9 expires, then the connection is released and a call-failed indication is sent to the calling party.

During call termination, when a REL message is sent, two timers – T1 and T5 (with T1 < T5) – are started to ensure that a RLC is received from the destination exchange. If a RLC response does not arrive before T1 expires, the REL is retransmitted and timer T1 is restarted. This procedure is repeated until T5 expires and then the reset circuit procedure is initiated. With this procedure, a reset circuit message is sent every T16 seconds until an acknowledging RLC is received.

1.7 Signalling in GSM Networks

The *Groupe Spécial Mobile* (now called *Global System for Mobile Communications* or GSM) standard for mobile networks was developed in Europe in the 1980's. Since 1989, the European Telecommunications Standards Institute (ETSI) took responsibility for the standardisation and development of the GSM specifications. While GSM is the dominant technology for second generation cellular networks, other incompatible technologies include North America's Interim Standard 95 (IS-95) and Interim Standard 54 (IS-54), and Japan's Personal Digital Cellular (PDC).

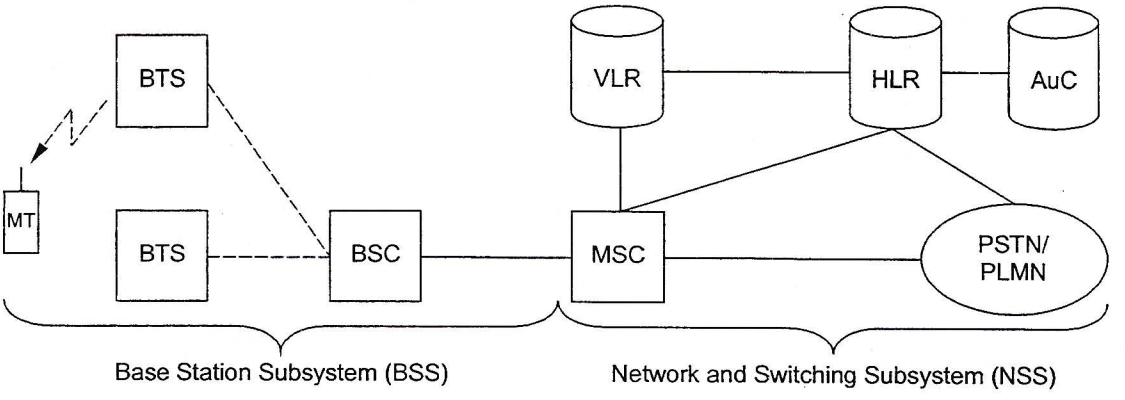


Figure 1-10. Basic GSM network entities and their functional relationship to each other.

Figure 1-10 shows the typical network elements in a simple GSM network and their functional relationship to each other in the signalling domain. The solid lines show the SS7 signalling relation between the respective network elements. Mobile network elements are grouped into two categories; namely the Base Station Subsystem (BSS) and the Network and Switching Subsystem (NSS). The purpose of these elements in a GSM network is as follows:

- Mobile Terminal (MT)[1]: The MT is usually a handheld device that is used by a subscriber to set-up and receive calls and other services.
- Base Transceiver Station (BTS): The BTS serves one or more cells and comprises the radio interface equipment (including the antenna and signal processing equipment).
- Base Station Controller (BSC): A BSC controls one or more BTS elements, in addition to radio resource management and handovers between cells.
- Home Location Register (HLR): The HLR is a database that stores subscriber profiles with subscription information (e.g. bearer services, roaming restrictions, etc.) and location information of a subscriber's current MSC and VLR.

[1] The text within this thesis uses the terms '*mobile terminal*' (MT) and '*subscriber*' interchangeably.

- Mobile Switching Centre (MSC): The MSC is an element that performs call processing and location management for MTs located in its coverage area.
- Visitor Location Register (VLR): The VLR is database that stores subscriber profiles while the subscribers roam on a MSC controlled by that VLR. A VLR may be responsible for one or more MSCs.
- Authentication Centre (AuC): The AuC is associated with a HLR and stores authentication keys for each subscriber. The keys are used to generate data that is used by the MSC and VLR for authentication and ciphering purposes.

In most implementations some of the functional entities described above are integrated into a single physical node, e.g. a MSC and VLR usually co-exist on a single platform and therefore each VLR is responsible for only one MSC. In addition to the above, most GSM networks also have other functional entities such as Equipment Identity Registers, Short Message Service Centres [ETSI GSM 03.02] and SCPs (e.g. for pre-paid billing).

Figure 1-11 shows the signalling protocol architecture within a MSC. The BSS Application Part (BSSAP) consists of following two functional parts [ETSI GSM 08.06 and GSM 08.08].

- Direct Transfer Application Part: This application is used for the transfer of call control and mobility management messages between the MSC and MT.
- BSS Management Application Part: This application supports procedures between the MSC and BSS, e.g. radio resource management and handover control.

Non-call control related signalling between GSM network entities in the NSS is performed by the Mobile Application Part (MAP) [ETSI GSM 09.02]. Some of functions performed by MAP include location management, inter-MSC handover control and the transfer of authentication information. In North American networks Interim Standard 41 (IS-41) [Lin & DeVries, 1995] provides mobility management functions similar to those provided by MAP.

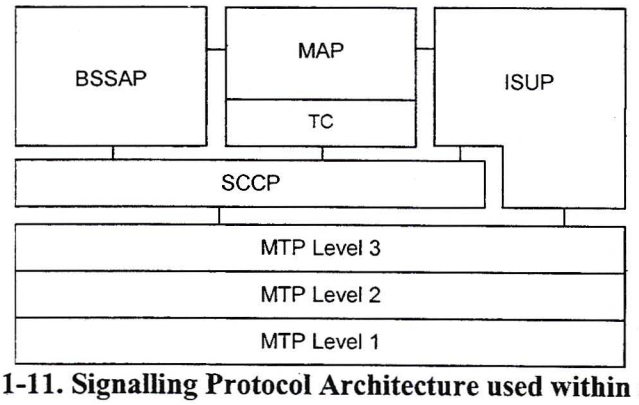


Figure 1-11. Signalling Protocol Architecture used within a MSC.

1.8 Mobility Management Signalling

The basic purpose of mobility management is to track the current location of mobile terminals, and to allow the network to deliver and maintain calls and services to those terminals. Mobility management within the NSS can thus be categorised into two areas:

- Location Management: This includes the methods used to track a mobile terminal, such as location registration, location updating and authentication.
- Call (or service) Delivery: This includes the methods used to search for a called mobile terminal in order to deliver a call or non-voice service (e.g. a short message).

1.8.1 GSM Location Management Procedures

In GSM, the location update procedure is activated in a mobile terminal when it enters a new

location area, the periodic location update timer expires or when the mobile terminal is switched on. When a mobile terminal initiates a location update, a timer T3210 is started. To limit the number of location update attempts, when location updates are unsuccessful, a location update attempt counter is incremented for each failed attempt. If a location update attempt fails due to a radio resource connection failure, timeout of T3210 or location update rejection; then the following procedure is performed [ETSI GSM 04.08]:

i) The radio resource connection is aborted if it is still present.
ii) Timer T3210 is stopped if it is still running.
iii) The attempt counter is incremented.
iv) If the attempt counter is less than 4, then timer T3211 is started else timer T3212 is started.
v) If the attempt counter is greater than or equal to 4, then the update status of the mobile terminal is set to NOT UPDATED, if it is not already in this state.

Timer T3211 is used to restart the location update procedure after a previous failed attempt, and Timer T3212 is the periodic location update timer. T3210 and T3211 are defined on the mobile terminal and have the following fixed values:

- T3210 = 20 seconds
- T3211 = 15 seconds

Timer T3212 is network implementation dependent and has a range from 6 minutes to 1530 minutes, with a granularity of 6 minutes. A more detailed description of the above procedure and timers, including detailed message flow diagrams is available in [ETSI GSM 04.08].

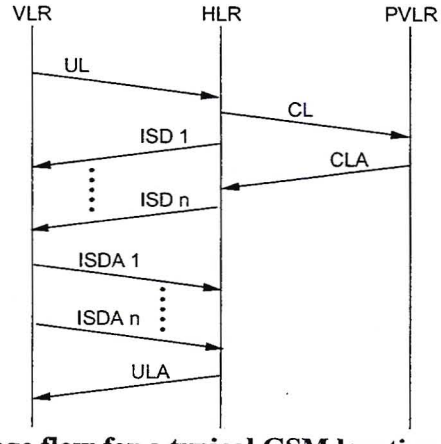


Figure 1-12. Message flow for a typical GSM location update procedure.

Assuming the mobile terminal has entered the coverage area of a new VLR then Figure 1-12 illustrates the signalling message flow between the new VLR, the subscriber's HLR and the subscriber's previous VLR (PVLR). The new VLR sends a MAP UPDATE LOCATION (UL) message to the subscriber's HLR. If the subscriber is allowed to roam in the new VLR, the HLR initiates the insert subscriber data procedure and the cancel location procedure. The cancel location procedure is used to cancel the subscriber's entry in the PVLR, by sending a MAP CANCEL LOCATION (CL) message from the HLR to the PVLR. After deletion of the entry, the PVLR responds with a MAP CANCEL LOCATION acknowledgement (CLA) message. The insert subscriber data procedure is used to transfer the subscriber's profile to the VLR. This is implemented by sending a MAP INSERT SUBSCRIBER DATA (ISD) message from the HLR to the new VLR. The VLR responds with a MAP INSERT SUBSCRIBER DATA acknowledgement (ISDA) message. Additional ISD and ISDA message flows may be required to transfer the subscriber's entire profile. After completion of the insert subscriber data procedure, the HLR responds with a MAP UPDATE LOCATION acknowledgement (ULA) message. Each one of these procedures has an associated timer with a range from 15 to 30 seconds. For example, if the insert subscriber data procedure fails, then the ULA message is still sent to the VLR when the corresponding timer expires, but with the *user error* parameter set to *System Failure*. The MAP level procedures and timers are described in greater detail in [ETSI

GSM 09.02].

1.8.2 GSM Call Delivery Procedures

Call delivery procedures are used to obtain routing information for mobile terminating calls. Figure 1-13 illustrates the signalling message flow for the retrieval of call routing information. On receipt of a call a MAP SEND ROUTING INFORMATION (SRI) message is sent from the originating MSC to the called subscriber's HLR. The HLR reacts by sending a MAP PROVIDE ROAMING NUMBER (PRN) message to the VLR that controls the area where the called subscriber is currently roaming. Upon receipt of this message, the VLR responds with a MAP PROVIDE ROAMING NUMBER acknowledgement (PRNA) message containing a roaming number, which is temporarily assigned to the called subscriber. The HLR forwards the roaming number in a MAP SEND ROUTING INFORMATION acknowledgement (SRIA) message to the originating MSC. Upon successful receipt of the roaming number the ISDN User Part in the originating MSC uses this information to establish a circuit switched path to the MSC of the called subscriber.

As described for the location management procedures, the timers used by the MAP call delivery procedures also have a range from 15 to 30 seconds. For example, if the provide roaming number procedure fails then the MAP SEND ROUTING INFORMATION acknowledgement (SRIA) message is sent to the VLR with the *user error* parameter, set to *System Failure*. The above procedures and timers are described in greater detail in [ETSI GSM 09.02]

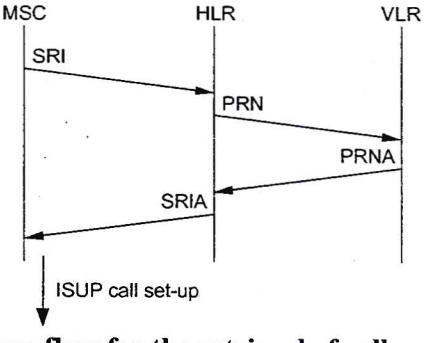


Figure 1-13. Message flow for the retrieval of call routing information.

1.8.3 Mobility Management in Future GSM, GPRS and Third Generation Mobile Networks

The Third Generation Partnership Project (3GPP) was formed in 1998. It a collaborative group made up of different standardisation bodies from around the world, with the task of producing Technical Specifications and Technical Reports for Third Generation Mobile Systems based on evolved GSM networks [3GPP Working Procedures, 2000].

The General Packet Radio Service's (GPRS) location management procedure is equivalent to that of GSM, while network-initiated data transfers (the equivalent of a circuit switched call delivery to a MT) also generate SRI and SRIA messages between the Gateway GPRS Support Node and the HLR [3GPP TS 23.060]. The HLR procedures differ slightly since a provide roaming number procedure is not required for GPRS. To ensure that future networks are compatible and inter-work with present-day systems the basic mobility management protocols, described above, remain unchanged within the circuit switched domain [3GPP TS 29.002]. The major consequence of the introduction of GPRS and new 3G services is the increase in the size of the subscribers' profiles and therefore more ISD messages will be required to transfer a subscriber's entire profile. However, the following methods have been specified by 3GPP to

help ease the signalling load (and the HLR transaction load) in future GSM, GPRS and 3G mobile networks:

- Super-Charging [3GPP TR 23.912]: Super-Charging reduces the signalling load generated by location updates from subscribers who register with the same VLRs on a daily basis. This is achieved by not deleting the subscriber's profile when he/she leaves the area controlled by a VLR and when the subscriber returns to the VLR the HLR does not perform the insert subscriber data procedure.
- Gateway Location Registers [3GPP TR 23.909]: A Gateway Location Register is simply a node that acts like a VLR towards a HLR and like a HLR towards a VLR. Its purpose is to cache subscriber profiles at a local database in order to reduce frequent location updates to the HLR.

Even though the mobility management aspects of this research focuses on GSM protocols it is also applicable to the mobility management signalling [3GPP TS 29.002] and call control signalling [3GPP TS 22.205] protocols defined for the circuit switched infrastructure of third generation networks. The location management procedures defined for GPRS are equivalent to those defined for GSM, with the exception that GPRS attempt counter allows for up to 5 attempts if previous attempts have failed [3GPP TS 24.008 and 29.002].

1.9 Congestion and Flow Control in SS7

The function of signalling traffic flow control is to limit input traffic during network failures and congestion, by sending status information to the traffic sources. User parts respond to congestion indications by reducing or blocking signalling traffic to the affected destination. Congestion situations may arise within each individual SS7 level, i.e., in level 2 (link receiver congestion or route set congestion), in level 3 (nodal congestion), and in level 4 (user part congestion) [Zepf et al., 1991].

Link level flow control throttles traffic when congestion is detected at the receiving end of a signalling link. During congestion the receiving end informs the transmitting end, of the congestion situation, by sending a status indication busy (SIB) and by withholding acknowledgements to any message signal units that are received from the transmitting end. The SIB indication is sent in the status field of a link status signal unit and will continue to be sent periodically until the congestion situation has abated. If the congestion persists for too long the transmitting end will take the link out of service and attempt to re-route traffic, if possible. The congestion detection mechanism used in the link receiver is implementation dependent [ITU-T Recommendation Q.703].

Route set congestion occurs when predetermined levels in the transmission and retransmission buffers of outgoing signalling links are exceeded. A route set is considered as being congested if at least one link within the route set is congested. This is based on assumption that load sharing is practised, and that all the links within a link set enter congestion uniformly. Signalling points respond to route set congestion by sending a MTP-STATUS primitive, with a congestion indication (CI) parameter, to the local user parts. Signalling transfer points respond to route set congestion by sending transfer controlled (TFC) messages to the traffic sources. Three types of signalling traffic flow control mechanisms have been defined, namely:

- The International Option (IO): In this option a route has two states; it is either congested or uncongested and MTP does not discard messages unless no further buffer storage resources are available. This option is used in the national signalling networks of most countries and on the international signalling links.
- The National Option with congestion priorities (NOCP): In the NOCP signalling messages are assigned congestion priorities, and multiple congestion states in MTP control the selective discard of messages during periods of congestion. This option is used in North American signalling networks.

- The National Option without congestion priorities (NOWP): This option is similar to the IO, except that multiple congestion states are supported. This option is used in the United Kingdom.

A detailed description of the MTP flow controls, and their corresponding ISUP and SCCP congestion controls is given in Sections 1.9.1, 1.9.2 and 1.9.3, respectively.

Signalling point and signalling transfer point congestion (or level 3 congestion) occurs when traffic arriving from the signalling links overwhelms the level 3 routing processor. Messages arriving from level 2 are queued in a level 3 message processing buffer. If the level 3 processor is overloaded, this buffer will become backlogged and eventually messages will be discarded when no further storage resources are available. The ITU-T Recommendation Q.704 states that the congestion and flow control mechanisms to be used for SP/STP congestion are implementation dependent, but should be compatible with procedures, messages and primitives specified for signalling route set congestion. The ANSI Standards [ANSI T1.111] have addressed this issue in more detail, and suggest the use of either [Rumsewicz, 1994]:

- a backpressure mechanism on the incoming links, which slows down the acceptance of messages by delaying the transmission of acknowledgements at the receiving end, or
- the application of flow controls defined for route set congestion at level 3, such that TFC messages are sent to the traffic sources during congestion.

With the first option, congestion at level 3 leads to a backlog of messages in the link receiver's buffers. These are messages waiting for buffer resources at level 3 to become available. The backlogged messages eventually lead to receiver congestion and consequently invoke the link level flow controls, thereby blocking messages at the transmitter and indirectly causing route set congestion. The traffic sources thus respond to the route set congestion by reducing their traffic to the affected destination. The second option uses the flow control schemes defined for route set congestion to detect and control congestion in the level 3 input buffers.

The automatic congestion control (ACC) mechanism is used to handle ISUP congestion. Two levels of congestion are defined: mild and severe. When congestion occurs, an optional ACC parameter indicating the current congestion level is added to release messages that are sent to the adjacent exchanges. The adjacent exchanges react by reducing their traffic to the affected exchange and resume normal transmission after a predetermined time period. Congestion detection at the affected exchange and the traffic reduction mechanism at the adjacent exchanges are implementation dependent.

Congestion in SCCP is handled by sending a subsystem-congested message, with a congestion level parameter, to the traffic sources. SCCP allows for eight levels of congestion to be reported. While the detection of congestion is implementation dependent the actions taken at the originating or relay nodes are equivalent to those used to throttle traffic during route set congestion (Section 1.9.3).

1.9.1 MTP Flow Control

1.9.1.1 International Option (IO)

In the international option congestion is resolved from three buffer occupancy thresholds, labelled abatement (A), onset (O) and discard (D), where $A < O < D$ (Figure 1-14). The international option does not discard messages, unless no further buffer storage space is available. The discard threshold represents the maximum buffer capacity available, while the hysteresis is provided to prevent oscillations, during recovery from congestion. If the link is in the uncongested state and the onset threshold is exceeded, the link status changes to the congested state and remains in this state until the abatement threshold is crossed from above [Zepf & Rufa, 1994].

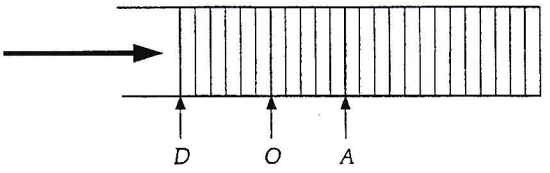


Figure 1-14. Buffer threshold scheme implemented in the International Option.

When a MSU is received from a local user part for any link of the congested route set, it is passed to level 2 for transmission and a congestion indication (CI) primitive is sent to each of the local user parts. Congestion indications are returned for the initial message that triggers the congestion and every n (default $n = 8$) messages, or alternatively for every N (default N ranges from 279 to 300) octets received thereafter to the affected link. The CI contains the DPC of the MSU received for transmission on the affected route set.

When route set congestion is detected at a STP, the route management function is triggered to send a transfer controlled (TFC) message back to the originating SP. Transfer controlled messages are generated for the initial message that triggers congestion and every n messages (or N octets) thereafter. The TFC contains the DPC of the MSU that triggered it, to indicate that route set congestion has occurred to the specified destination. On receipt of a TFC message, the traffic management function of MTP level 3 sends a CI primitive, with the address of the affected destination, to each local user part.

1.9.1.2 National Option with congestion priorities (NOCP)

In the national option with multiple congestion states and congestion priorities, MSUs are assigned discard priorities and the link level buffers are divided into multiple congestion states. The link level buffers are assigned M ($1 \leq M \leq 3$) separate congestion detection thresholds and '$M + 1$' congestion states, where state zero implies that the link is uncongested. For each congestion state $m > 0$ ($m = 1, \ldots, M$) three thresholds are defined; O_m, A_m and D_m denote the congestion onset, abatement and discard thresholds for congestion state m, respectively. When onset threshold O_m is exceeded, the link status changes to congestion state m. While the link is in congestion state m and the buffer size is below D_m, all incoming MSUs with priority '$m - 1$' or higher are accepted while MSUs with a lower priority are discarded. When D_m is exceeded, MSUs with priorities less than m are discarded. The link remains in congestion state m until the abatement threshold, A_m, is crossed from above – the congestion status is then decremented by one [Zepf & Rufa, 1994].

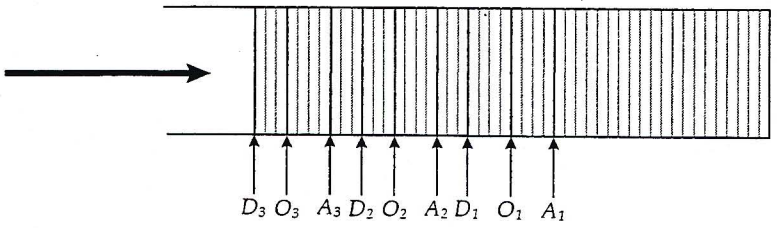


Figure 1-15. Buffer threshold scheme implemented in the National Option with congestion priorities.

When the congestion status of a signalling link at a SP changes, a CI message is sent to each of the local user parts, indicating the address of the affected destination (obtained from the DPC of the message that triggered the congestion) and the current congestion state. The user parts are expected to take appropriate action to stop generation of signalling messages with a priority less than the current congestion state. Low priority messages received from the local user parts for the affected destination are discarded by MTP level 3.

When a STP selects a congested signalling link for transmission of a message with a priority less than the current congestion state, the following actions are taken:

- A TFC is generated and the message is transmitted if its congestion priority is greater than or equal to the current discard threshold.
- A TFC is generated and the message is discarded if its congestion priority is less than the current discard threshold.

The receipt of a TFC message at a SP invokes the *signalling route set congestion test* (SRSCT) procedure in MTP level 3 and a CI primitive, containing the address of the affected destination and the current congestion state, is sent to the local user parts. The SRSCT procedure starts a timer, T15, for the affected destination. While the SRSCT procedure is active, messages received from the local user parts with a priority less than the current congestion state (m) are discarded. If additional TFCs with a congestion status greater than or equal to m are received, T15 is restarted and CIs are sent to the local user parts. If T15 expires without receipt of a TFC, a signalling route set congestion test (RCT) message with discard priority '$m - 1$' is sent to the affected destination and a second timer, T16, is started. If a TFC with a congestion status that is greater than or equal to m is received during this period then T15 will be restarted and T16 is stopped. If T16 expires, the congestion status is decremented and T15 is started for the new congestion status. This process continues, until the congestion status drops to zero. Figure 1-16 illustrates the SRSCT procedure.

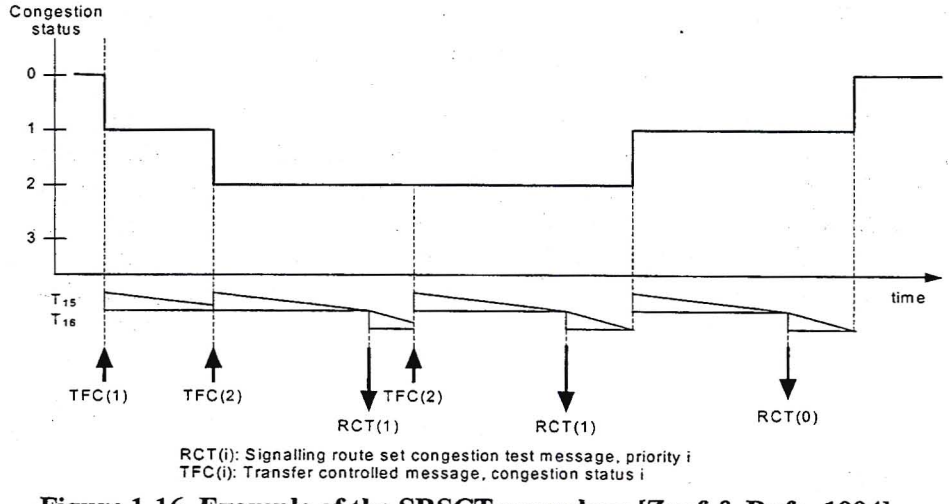


Figure 1-16. Example of the SRSCT procedure [Zepf & Rufa, 1994].

The assignment of congestion priorities to signalling messages is performed by the user parts, and is implementation dependent. Two bits in the SIO field are used to carry the congestion priority information. The IAM messages are usually assigned the lowest priority – 0, since it makes sense to discard messages from new calls rather than those from calls already in progress. ACMs and RELs are assigned a priority of 1, while ANMs, RLCs and network management messages (such as TFCs) are assigned the highest priority – 2. The highest priority messages will only be discarded when the buffer resources are exhausted and D_3 is reached.

1.9.1.3 National Option without congestion priorities (NOWP)

The national option without congestion priorities can be considered as a more sophisticated version of the International Option. The buffer threshold assignments are identical to those used in the IO, except that '$S + 1$' ($1 \leq S \leq 3$) congestion states are defined, where zero represents no congestion and S the highest level of congestion. The congestion state is determined from the buffer occupancy levels observed during congestion. At the onset of congestion the signalling link is assigned a predetermined congestion state s ($s = 1, \ldots, S$) and a timer, Tx, is started. If the congestion status is s and the buffer occupancy remains continuously above the onset

threshold during Tx, the congestion status is incremented to '$s + 1$', i.e. the transition (Tx, $\ell > \ell o$), in Figure 1-17, will only occur if '$\ell > \ell o$' for the entire duration of Tx. When the buffer occupancy drops below the abatement threshold, timer Ty is started. If the buffer occupancy remains continuously below the abatement threshold during Ty, then the congestion status is decremented, i.e. the transition (Ty, $\ell < \ell a$), in Figure 1-17, will only occur if '$\ell < \ell a$' for the entire duration of Ty. Otherwise the current congestion status remains unchanged.

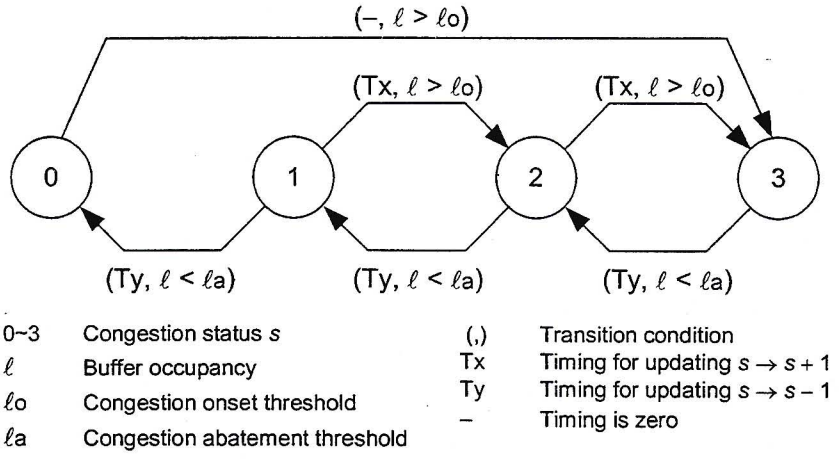


Figure 1-17. Example of signalling link congestion status transitions [ITU-T Recommendation Q.704].

As described in the IO above, a TFC is generated at a STP for the first message that triggers congestion and every n (default $n = 8$) messages or every N (default N ranges from 279 to 300) octets received thereafter for transmission on the affected link. Except that the TFCs also contain the current congestion status. On receipt of the TFC, MTP level 3 sends a CI with the DPC of the affected destination and the current congestion status to each local user part. Congestion in the local signalling links of a SP results in CIs being sent to each user part for every n messages or N octets received.

1.9.2 ISDN User Part Congestion Control

The MTP user parts are required to respond to a congestion indication by reducing their load to the affected destination. In the National Option with congestion priorities users are expected to stop all traffic to the affected destination with a priority less than the current congestion state. In the International Option and National Option without congestion priorities traffic to the affected destination is reduced in several steps. When the first CI is received, the traffic load to the affected destination is reduced by one step and two timers T29 and T30 are started. Any further CIs that are received for the affected destination while T29 is active are ignored, so as not to reduce the traffic too rapidly. If a CI is received after T29 has expired and during T30, the traffic load will be dropped by one more step, and T29 and T30 will be restarted. This step-wise reduction continues until the last level is reached, at which point all traffic to the affected destination is blocked. If T30 expires, the traffic is increased by one step and T30 is restarted unless transmission of the full traffic load has resumed. The number of traffic reduction steps and the amount of increase or decrease in traffic at each step transition are considered as implementation dependent issues. Figure 1-18 illustrates the operation of the ISUP congestion control mechanism.

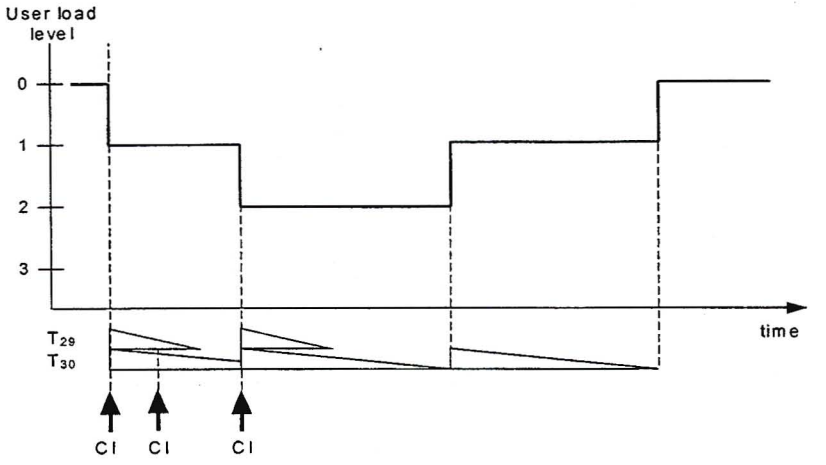


Figure 1-18. ISUP congestion control mechanism [Zepf & Rufa, 1994].

1.9.3 SCCP Congestion Control

The SCCP congestion control for the IO and NOWP is similar to the ISUP congestion control in some respects, but each SCCP message type is assigned an importance value to control the reduction of traffic to the affected destination on a message type basis. The load reduction mechanism consists of N+1 Restriction levels (RL) and M Restriction sublevels (RSL), where N = 8 and M = 4. At the lowest restriction level, zero, no messages are blocked. For higher restriction levels, messages with an importance value that is lower than the RL are discarded. If the importance value of a message equals the RL, then the 4 restriction sublevels are used to determine if 0%, 25%, 50% or 75% of these messages should be discarded.

When the first CI is received the RSL is incremented, thereby resulting in the blocking of 25% of the messages with an importance value of zero and two timers Ta and Td are started. Congestion indications received while Ta is active are ignored. If a CI is received after Ta has expired and while Td is active, Ta and Td are restarted and the RSL is incremented to block an additional 25% of messages with the corresponding restriction level. If the RSL is incremented to M then its value is reset to zero and the RL is incremented (Figure 1-19). This process can continue until RL = N, when all the outgoing messages are blocked. If Td expires, the RSL and RL are decremented in an equivalent manner and Td is restarted unless transmission of the full traffic load has resumed (the operation of Ta and Td is analogous to the operation of T29 and T30 in Figure 1-18). The ITU-T Recommendation Q.715, suggests that values selected for M, N, Ta and Td should allow for synchronisation with the ISUP congestion controls.

RL = 0	RSL = 0	non-blocking
RL = 0	RSL = 1	
RL = 0	RSL = 2	
RL = 0	RSL = 3	
RL = 1	RSL = 0	
RL = 1	RSL = 1	
...	...	
RL = 7	RSL = 2	
RL = 7	RSL = 3	
RL = 8		100% blocking

Figure 1-19. SCCP step-wise congestion control mechanism.

1.9.4 Comments on SS7 congestion and flow controls

The SS7 congestion control procedures are based on the affected destination concept, even though congestion detection is based on the affected route concept [Manfield et al., 1993]. A number of researchers have found the affected destination concept to be unnecessarily restrictive, since loads may not be equally balanced over all the links of a link set, especially if the signalling traffic includes IN and UUI messages. When a user part receives a congestion indication it throttles all the traffic to the affected destination, regardless of whether the traffic uses the affected route or link that triggered the congestion. In this situation it would be advantageous if only traffic on the congested route was throttled or alternatively, instead of throttling traffic on the congested route, the traffic could be rerouted on to an uncongested route. Another consequence of the affected destination concept is that RCT messages used in the NOCP may not travel on the same route or link as that on which congestion was detected. In this case the congestion information may not be reliably signalled to the source node.

SCCP is widely used in mobile networks to transfer mobility management traffic and IN transactions. However, the major shortcomings of SCCP is its inability to provide end-to-end congestion control information and signalling of route-set congestion information to higher layer applications, especially in scenarios where STPs may also perform global title translation and relay functions ([ITU-T Recommendation Q.715] and [Zepf & Rufa, 1994]). The ITU-T Recommendation Q.715 therefore recommends that SCCP take congestion control actions of behalf of higher layer applications.

1.9.5 STP congestion

Overload control in the STP's level 3 routing processors is particularly important in central processor based architectures (or any architecture where a variable number of incoming links can be assigned to a single level 3 processor). Figure 1-20 illustrates a simplified central processor based STP architecture. In STPs it is likely that the signalling links will be relatively lightly loaded, possibly as low as 5% to 20%, in a busy hour. The level 3 processor normally operates at between 30% to 40% utilisation, but it is possible for link loads to increase by many times over their normal levels – in fact, more than the capability of the level 3 processor to absorb the additional load [Rumsewicz, 1993].

Various factors can stimulate an increase in calls and signalling traffic to a STP; including a media stimulated events, software faults, configuration faults or the failure of a network element [Karmarkar, 1994]. In mobile networks an increase in mobility management traffic can occur for various reasons:

- The failure of a BTS or BSC that provides radio coverage on the edge of a location area can provoke a surge in the location update traffic due to mobile terminals registering on the adjacent VLR in areas where spill-over radio coverage is available from the adjacent location area.
- A surge in location updates can occur during the failure/recovery of a VLR or HLR.
- A large number of subscribers propagating to a particular area (e.g. a cricket match) can also lead to a increase in the number of location updates and calls in that area.

The transport of signalling traffic by asynchronous transfer mode (ATM) and IP signalling links is also likely to add to the congestion problem by creating greater resource mismatches between the existing narrowband network elements and the new broadband technologies. As a consequence, researchers envisage that the level 3 routing processors of a STP are the most likely candidates to experience congestion during signalling traffic overloads.

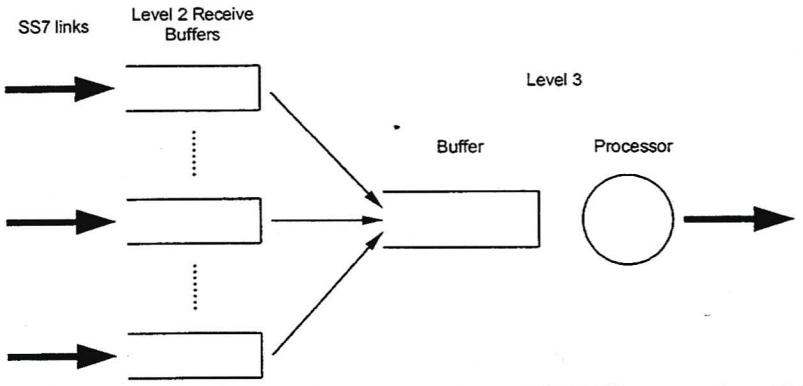


Figure 1-20. Idealised central processor based STP [Rumsewicz, 1993].

1.10 Thesis Outline

This thesis is organised as follows:

- Chapter 1 presented an introduction to SS7, together with an overview of the SS7 protocol architecture and a description of the typical call set-up and mobility management procedures. The SS7 congestion and flow control mechanisms were then discussed in detail and lastly, the chapter discussed congestion in STPs with central processor based architectures.

- Chapter 2 is a literature survey of publications on congestion and flow control in signalling networks and mobility management research. The first two sections discuss basic congestion control techniques and the application of modelling techniques for multilayered protocol architectures to the analysis of SS7 network performance. The subsequent section reviews previous research on the performance analysis of the SS7 networks and the various control schemes that have been proposed to overcome the shortcomings of the current protocol implementation. The next section discusses research studies that consider the impact of customer reattempts and application level recovery actions on network performance and the last section addresses mobility management research including the various techniques that have been proposed to reduce signalling load in mobile networks.

- Chapter 3 develops analytical models to analyse the steady-state performance of priority based queuing models and signalling networks with multiple STPs within a PSTN. The subsequent sections examine the results obtained from the analysis and these are compared to those obtained by simulation. The results are presented for various overload scenarios and message discard schemes.

- Chapter 4 develops an analytical model to examine the transient performance of PSTN signalling networks. The numerical results are then compared to those obtained by simulations for various overload scenarios and message discard schemes. The last section investigates whether equilibrium in the meta-stable region of the call completion rate curve (obtained in the steady state analysis) is realisable.

- Chapter 5 analyses the steady-state performance of SS7 in Public Land Mobile Networks (PLMNs). Analytical models are derived to analyse the GSM mobility management protocols and their interaction with ISUP call set-up procedures. In addition, this section also analyses mobility management protocols that have been proposed for third generation networks and their performance during STP congestion.

- Chapter 6 extends the analysis of mobility management protocols by examining the transient performance of PLMNs during STP congestion. The interaction between SCCP and ISUP messages and the processes that lead to sustained congestion are examined.

- Chapter 7 investigates the effectiveness of various congestion and flow control mechanisms during STP congestion in various overload scenarios. Here, a number of congestion control schemes are proposed and their performance is compared to control schemes that are often used by other researchers. This chapter also examines how the different congestion and flow control parameter settings influence network performance, and the interaction between congestion controls in a multiple user part environment is also considered.

- Finally, Chapter 8 summarises the thesis and lists the main insights and conclusions gained from this research investigation.

1.11 Original Contributions of this Thesis

The main contributions of this research study are:

a) A derivation of the equilibrium distribution of a priority based queuing discipline with three levels of message discarding. This queuing model is used to analyse the performance measures of the processing buffer in a STP and to evaluate the impact of different message priority schemes on network performance. While this research concentrates on STP processor congestion, this queuing model can also be used to analyse the performance of the link transmission queues when the feedback control mechanisms are ineffective.

b) The derivation of generalised analytical models to analyse the steady state and transient behaviour of SS7 networks with one or more STPs. Unlike previous studies that concentrate on the performance of the congested entity, the analysis presented here considers the performance of all the STPs, it evaluates all the link loads and determines the traffic loads to and from all the SP regions. Furthermore, different traffic loads can be defined for each source-destination pair and different routing algorithms can be used in the analysis.

c) Explicit modelling and analysis of the performance of the GSM mobility management signalling protocols and third generation mobility management protocols. Unlike pervious studies on mobility management, this work explicitly models the individual signalling messages and application level recovery procedures during normal and congestion scenarios. In addition the impact of the increase in subscriber profile sizes on network performance is also considered. The analysis also allows for an examination of the interaction between the mobility management procedures and the ISUP call set-up procedures.

d) An investigation of various congestion control schemes that are able to outperform the commonly used linear step-wise control schemes, in different overload scenarios is performed. The proposed congestion controls operate by throttling the IAM load sharply at the onset of congestion and then they gradually reintroduce the IAM load if no further TFC messages are received.

e) A performance evaluation of the National Option without congestion priorities, relative to the performance of the International Option and the National Option with congestion priorities, is performed. Recommended values for the NOWP flow control parameter settings are also determined.

Parts of the research included in this thesis have been presented (or will be presented) by the author at the following conferences:

1. Chana, A & Takawira, F, (1996), "Modelling Congestion Control in SS7 networks," Teletraffic Systems Engineering Seminar, Electronic Engineering, University of Natal, South Africa, September 1996.

2. Chana, A & Takawira, F, (1997), "Performance of the Transfer-Controlled Procedure in Controlling SS7 Congestion," Teletraffic '97, Rhodes University, South Africa, September 1997.

3. Chana, A & Takawira, F, (2000), "The Impact of Location Management Protocols on the Signalling Traffic Load in Mobile Networks," SATCAM 2000, September 2000.

4. Chana, A & Takawira, F, (2003), "An Examination of Signaling Traffic in a Super-Charged Mobile Network," accepted for presentation at WCNC 2003.

Parts of this research have also been submitted to the following journals for review:

1. Chana, A & Takawira, F, (2002), "An Analysis of Mobile Network Performance during Signalling Network Congestion," submitted to IEEE Transactions on Vehicular Technology.

2. Chana, A & Takawira, F, (2002), "SS7 Congestion Control Performance in Mobile Networks," submitted to IEEE Transactions on Vehicular Technology.

3. Chana, A & Takawira, F, (2002), "The impact of STP Congestion on the Performance of Mobility Management Procedures," submitted to Wireless Networks.

2. Performance Modelling and Evaluation of Signalling Protocols

2.1 Introduction

The goal of Signalling System No. 7 is to provide a packet switched signalling transport backbone that can support call set-up procedures and database transaction in the traditional telephone network, and also provide the flexibility necessary to support new and emerging telecommunications services. A packet switched network is principally a resource-sharing environment where each user is not assigned dedicated resources throughout the network during the period of the transaction, but instead specific resources are only utilised as required. But resource sharing has its disadvantages; transmission capacity is not always guaranteed for the duration of the transaction and congestion could lead to a decreased throughput, increased delays and possibly a network wide deadlock. Congestion usually occurs in a packet switched network when traffic from the source nodes overwhelms the available resources in a particular network entity. There has been a tremendous amount of research on congestion in various packet switched networks over the past three decades. Section 2.2 briefly examines various basic congestion and flow control schemes (including some that are analogous to SS7's congestion and flow controls).

To ensure optimal performance, network planners have to be able to predict the performance of a network and locate potential bottlenecks before they occur. The simplest performance models view packet switched networks as a collection of queues with exponential service times, where packets enter the network, visit a series of queues and eventually exit the network. Each queue within the network could thus represent the delay encountered by a packet within the transmission links or transit nodes of the analogous real-world network element. But, since the OSI model was adopted, a number of researchers became concerned about the processing overheads associated with the layered structure of OSI compliant protocols. They mooted that the complexity of multilayered protocols could give rise to processing bottlenecks and thus limit the achievable throughput of the higher layer protocols in emerging broadband networks [Conway, 1991]. Section 2.3 examines modelling methodologies that have been proposed to analyse complex multilayered protocol architectures, including SS7.

Congestion and flow control schemes are usually designed and optimised to operate efficiently in well known and quantifiable overload scenarios. However, the Signalling System No. 7 protocol was designed to operate in an evolving integrated services environment. The traffic patterns attributed to new services were therefore previously not observed in telecommunications networks. Research over the past few years has thus investigated the effectiveness (or ineffectiveness) of SS7 congestion and flow control mechanisms and their applicability to support the current and future network services. Section 2.4 examines previous studies in this area and discusses the various methods that have been proposed to overcome the shortcomings of the current protocol implementation.

Several factors can influence network performance during an overload; these include the duration and severity of the overload traffic, the robustness of the congestion and flow control procedures, and the reaction of customers and application level processes to the reduction in their perceived grade-of-service. Applications are often designed to respond to packet discarding, due to high error rates or congestion, by reattempting to send the packet until transfer is successful or until a reattempt timer expires. Customers, in a telecommunications network, respond similarly to call failures. While most research on network congestion has overlooked the impact of reattempts, Section 2.5 examines some of the studies that highlight the

importance of incorporating application level and customer reattempt behaviour in performance models.

Second generation mobile networks rely on SS7 for the transport of mobility management related signalling messages. Since mobility management procedures are the highest contributors of signalling traffic and database transaction load, various researchers have made several proposals that aimed to minimise the mobility management load. Section 2.6 examines research on mobility management in the Network Subsystem. Since most of the research in this area has been intended for use in third generation networks and is compared to GSM and IS-41, some of the shortcomings of these proposals are also discussed.

2.2 Congestion and Flow Control in Packet Switched Networks

Congestion occurs when the demand from sources exceeds the available network resources. Congestion is usually perceived as increased delays and throughput degradation. In most protocols this problem is further aggravated by retransmissions, when no response is received from the destination node before a transmission failure timer expires. Resources that are commonly shared in a packet switched network include memory buffers, link bandwidth and processing capacity. Increasing the capacity of the resources doesn't necessarily improve performance and consequently could lead to greater resource mismatches in some instances [Jain, 1990]. Thus, the function of congestion and flow controls is to continuously monitor resource utilisation (such as link load) and to either increase the scarce resource (by allocating additional memory or bandwidth), or to reduce the demand for resources (by sending choke packets to the sources), in order to ultimately improve network performance.

Flow controls attempt to prevent congestion at the affected resource by controlling the flow of traffic into the network. The basic objectives of flow control mechanisms are as follows:

- maximise the useful throughput,
- minimise the transfer delay,
- minimise the congestion and flow control overheads, and
- resource matching.

Flow controls can be implemented at the different protocol levels, where congestion is possible. Subsection 2.2.1 discusses the different levels of flow control that are often distinguished in literature.

Congestion control mechanisms are usually implemented as close as possible to the source of the traffic. The basic aim of congestion control is to alleviate congestion once it appears, and to suppress some (or all) of the offered traffic load, until the congestion situation has passed. Subsection 2.2.2 discusses various congestion and flow control schemes and some of their applications.

2.2.1 Flow Control Levels

Flow control can be exercised in various protocol levels. The following have been commonly identified [Gerla & Kleinrock, 1980]:

1. Hop level – This level of flow control exists between neighbouring nodes and is often implemented in the data link layer. The effects of hop level flow control can eventually propagate from the congested node to the traffic sources. This type of congestion propagation is often referred to as *backpressure*. Link level flow control in SS7 (Section 1.9) is an example of hop level flow control.
2. Transport level – This level of flow control is responsible for the reliable delivery of packets from source to destination and is required to prevent congestion of the higher application

layer processes. Window flow control schemes are often implemented at this level. The transmission control protocol (TCP) is an example of a transport level protocol that implements window flow control [Spragins, 1991].

3. Entry to exit level – This level is responsible for protecting exit switches from buffer congestion. Most of these controls are based on window flow control schemes. The IBM Systems Network Architecture implements this type of flow control [Gerla & Kleinrock, 1980].
4. Network access level – The network access flow control attempts to control input traffic based on the network's congestion status. This flow control is implemented in the network layer of source nodes and controls network access with the aid of buffer allocation schemes. The input buffer limit scheme, discussed in Section 2.2.2.2, is a typical example.

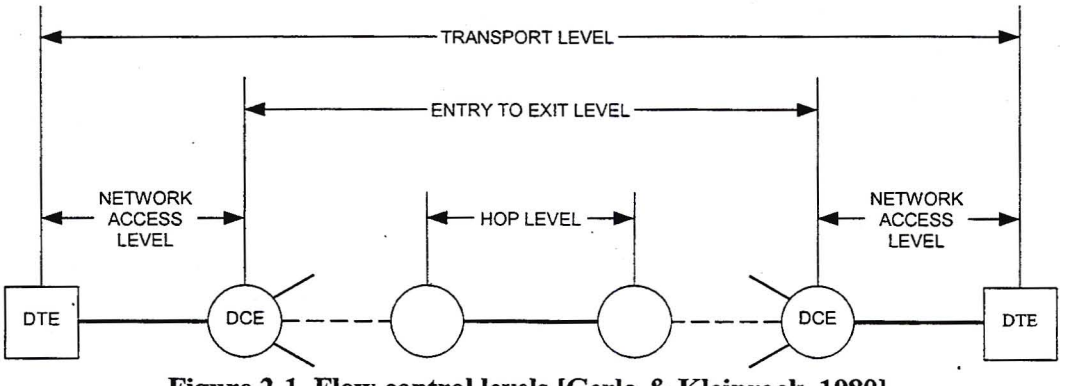


Figure 2-1. Flow control levels [Gerla & Kleinrock, 1980].

2.2.2 Congestion and Flow Control Schemes

2.2.2.1 Sliding Window Flow Control

The sliding window flow control scheme is one of the most commonly used flow control methods. Here the sender can only transmit a prescribed number of packets, often referred to as the transmitting window size. When a packet is transmitted the window size is decremented by one, and further packets may only be transmitted if the window size is greater than zero. Receipt of an acknowledgement from the destination, for each packet transmitted, increments the window size by one. In some implementations an acknowledgement is only received once the entire window has been transmitted. The advantage of this scheme, lies in the ability of the destination node to regulate the flow of incoming packets by varying the window size, and thus reduce traffic load during congestion. The Stream Control Transmission Protocol, which has been defined to allow for reliable SS7 transport over IP, uses a window flow control mechanism [IETF RFC 2960].

2.2.2.2 Buffer Allocation Algorithms

Buffer allocation algorithms are often implemented at the input switches of a network, with the intention to reduce incoming traffic during congestion and to prevent deadlock. In its simplest implementation, a switch has a finite pool of buffers available for the storage of packets that are waiting to be processed. Once the reserved capacity is depleted, surplus packets are discarded. Depending on the discard policy, either the first (oldest) packet in the queue, the last (newest) packet or a randomly selected packet from the queue will be discarded. The higher layer protocol recovery actions in the source nodes are therefore required to ensure that packets successfully arrive at their destinations.

The input buffer limit (IBL) scheme proposed by Lam & Luke Lien [1981] divides input packets into two classes – new (or incoming) and transit packets. Only a fixed number of new packets, which is less than the available buffer capacity, are accepted while transit packets are

able to fill the entire buffer. This method is effective at reducing congestion and increasing throughput, when compared to systems that do not distinguish between new and transit traffic. This method is favourable since transit traffic has already utilised some of the network resources and it therefore makes sense to assign a higher priority to these packets. Schwartz & Saad [1979] extended the IBL scheme to block input packets if the number of packets already in the buffer is greater than a specified threshold. The effect is a slightly improved throughput and delay performance. In other implementations packets are assigned a higher priority as they get closer to their destination, which effectively reduces the probability of them being discarded. In some protocols, packets are assigned priority classes based on their relative importance. Here, packets with a higher priority have a higher probability of being accepted at a congested node than packets with a lower priority.

Irland [1978] considered buffer management policies in packet switches, where a finite buffer capacity is shared between several output queues. Simulations and analysis confirmed that under unbalanced traffic conditions the busiest link's queues tend to monopolise most of the available buffer capacity during congestion. The analysis presented compares the performance between the unrestricted buffer sharing policy, the optimal restricted buffer sharing policy, fixed-storage (no sharing) policy and the square-root sharing policy. In the optimal sharing scheme, which performed the best, performance depends on the traffic load and therefore requires a recalculation of the optimum buffer limits for variations in incoming traffic load to minimise the probability of loss. This requires an adaptive controller, which may be difficult to implement in practice. However, the optimal scheme was found to be well approximated by the square-root scheme, which is easier to implement and does not require a recalculation of the buffer limits for different traffic loads.

2.2.2.3 Isarithmic Flow Control

Isarithmic flow control [Davies, 1972] limits the number of packets that are allowed in a network, through the use of *transmission permits*. Whenever a node needs to send a packet, it has to obtain a permit from the network before transmitting the packet. There are two types of permits, which may exist in a network, fixed and/or free permits [Schwartz & Saad, 1979]. Fixed permits are stored at a node, these are returned to the source once the accompanying packet reaches its destination. Free permits randomly circulate the network and have to be captured by a node before a packet can be transmitted.

Schwartz & Saad [1979] analytically analysed a number of isarithmic models, while Price [1977] presented results obtained through simulations. In both cases, the results indicate that the isarithmic flow control performs well with uniform traffic streams but poorly with non-uniform and time-varying traffic streams. Network design for the isarithmic scheme is also difficult, since the distribution of permits within the network could result in some nodes being unable to obtain enough permits. It is also possible for localised congestion to occur in some parts of a network, where permit concentration is high. Transmission errors and node failures are also potential problems, since they result in permits being lost.

2.2.2.4 Choke Packet Scheme

Choke packet feedback control was originally proposed by Majithia et al. [1979]. In this scheme, the resources in a node are constantly monitored and explicit feedback messages are sent to the source nodes if congestion were to occur. Congestion feedback packets are often called *choke packets*, *source quench messages* or TFC messages (in SS7). The congestion and flow control mechanisms discussed by Majithia et al. are similar to those used in the International Option of SS7 (Section 1.9.1.1). However in their scheme, packets that have already generated a choke packet are tagged, so that they do not to create further choke packets. Choke packets are also tagged, to prevent them from generating additional choke packets on other congested paths.

In Majithia et al. two types of load reduction algorithms are implemented at the source nodes. In the first, every *n*th packet to a particular destination is blocked, i.e. if the first reduction step discards every fourth packet; then when the next choke packet arrives every third packet is blocked and so on until every packet is blocked. In the second method an exponential reduction rate is used. Here every second packet is allowed in the network when the first choke is received and every fourth packet is allowed into the network after the second choke arrives, and so on. Both algorithms use timers that are similar to the ISUP congestion control timers in SS7. The results indicate that the exponential mechanism overcontrols traffic during small overloads but performs well under very severe overload conditions, by permitting a higher throughput.

2.2.2.5 Other Congestion and Flow Control Methods

Rate based: In rate based methods the sender transmits at a fixed rate. The destination is able to send feedback information to dynamically adjust the transmission rate as resource availability changes. The choice between a window or rate based flow control depends partially upon the bottleneck resource. If processing is the bottleneck, then a rate based scheme is generally used; but if memory is the limited resource, then a window based scheme is used [Jain, 1990].

Stop and Go: Stop and Go flow control is mainly used by data link control procedures. In SS7 the receiving end sends a *status indication busy* and withholds acknowledgements (Section 1.9) to stop the transmitting end from sending MSUs. In this implementation, it is sometimes also necessary for the receiver to consider propagation delays, because several packets may still arrive before the transmitter receives the stop signal.

Timeout based congestion control: Jain [1990] proposed a timeout based scheme, where packet loss is used as a congestion indicator. Here a timer is started for each packet transmitted. If an acknowledgement is not received, before the timer expires, the sender assumes that the packet was lost. When a large number of timers expire within a small period of time, the sender assumes that packets were discarded in the network due to congestion and the outgoing load to the congested resource is then reduced. This type of flow control is a form of implicit feedback. However, in a network where timeout values are large, the load reduction may not be fast enough; in which case a large number of packets will be discarded before the source nodes respond. The source nodes could also be too slow to resume transmission at their normal transmission rate, once the congested resource has returned to normal.

Probe packets: Probe packets are packets sent into the network to extract route status information, e.g. the current state of a congested resource. The outgoing traffic load is then adjusted based on the delay experienced by the probe packets, or the congestion information returned in a feedback message. RCT messages, used by SS7 in the National Option with congestion priorities, are an example of probe packets (Section 1.9.1.2).

2.3 Performance Modelling of Multilayered Protocol Architectures

The OSI reference model was designed to provide a standard architecture for the development of communication protocols and networks. Concern for the processing overheads associated with the OSI suite of protocols and its effect on system performance, as perceived by the end users and processes, prompted research into this area. It is feared that the complexity of the OSI protocols could give rise to communication processing bottlenecks in the higher layers of the OSI reference model [Conway, 1989], and thus impede the achievable throughput of higher layer protocols in high-speed networks.

Previous analytical performance studies had concentrated on the performance of specific protocols in isolation, due to the difficulties posed by analytical modelling methods and the fact that most protocol functions cannot be easily modelled by product form queuing networks. Since the adoption of the OSI model as the standard reference architecture for protocol development, a number of researchers have proposed methods for the analysis of multilayered protocol architectures. One of the first methods was by Kritzinger [1986], where the performance of OSI compliant models is predicted directly from formal specification. This approach considers a single multi-layered system and is based on a multiple-chain queuing network. However the major shortcoming of his method is that it requires a detailed specification of transition probabilities and mean holding times of the states in the transition-relation graph [Conway, 1991]. If these quantities are not available as measurements from actual implementation or simulation, they have to be estimated and even if implementations are available, obtaining transition times for every possible transition scenario may not be practically realisable. Complex protocols with large state spaces could also create potential computational difficulties.

This following sub-sections examine two modelling methodologies that have been proposed for the analysis of SS7 networks. The method proposed by Willman & Kühn [1990] (Subsection 2.3.1) is applicable to protocols that require priority based scheduling strategies, segmenting and forking of messages - typical of the processing functions that are encountered in today's communication systems. Even though their work concentrates on SS7, a similar modelling approach can be used to analyse other layered protocol architectures.

In the technique proposed by Conway [1989] (Subsection 2.3.2) a generic queuing network is constructed from the structure and specifications of the OSI reference model. The resulting model is then solved by using an iterative decomposition technique to determine the network's performance parameters. In a later paper, Conway [1990] described how this method could be applied to analyse the SS7 protocol.

2.3.1 Hierarchical Decomposition and Aggregation

Willmann and Kühn [1989, 1990] analysed the performance of the SS7 protocol by constructing generic submodels for each layer of the protocol architecture. These generic submodels are derived from the functional specifications of each layer, as defined in the ITU-T Recommendations. The elements within each layer are represented by queuing network elements such as priority processors and multiple-chain multi-class routing chains. Figure 2-2 shows the generic submodel for MTP level 3, together with its functional block diagram. The generic submodel consists of a three-phase processor with a separate queue for each processing phase. The functions corresponding to control and management are not included in the model since their effects on system performance are considered to be negligible.

A layered model structure is constructed by combining the various queuing submodels; including the traffic sources, sinks and application level processes, into a single generic model of the network node (Figure 2-3). Each submodel has inputs and outputs for the protocol data units that are exchanged between the adjacent layers. The highest submodel is connected to the external sources and sinks of traffic. The infinite source corresponds to a Poisson traffic source and the infinite server queue models the effects of application level processes and customer response times.

For the analysis, the signalling links and the various submodels are analysed in isolation via decomposition and aggregation techniques. Decomposition allows complex systems to be broken into simpler subsystems that can be analysed in isolation; for example, a signalling link set can be decomposed into individual signalling links. The basic concept behind aggregation is that messages of the same class belonging to different chains have identical service

requirements in each layer, therefore aggregate traffic is simply a superposition of all chains and classes at a shared resource. This allows for the delay endured by individual messages to be calculated at resources that are influenced by the behaviour of aggregate traffic streams. To simplify the analysis, Willmann and Kühn assumed aggregate traffic to be Poisson in nature. Their analysis allows one to determine performance measures such as end-to-end transfer times, traffic flow rates and processor utilisation in each layer. This modelling approach was later extended in Bafutto et al. [1994] to account for implementation dependent characteristics in multi-vendor environments and to develop a signalling network planning tool. The conclusions from this study indicate that IN applications in future networks will have a significant impact on the processing load of signalling points and will also significantly increase the end-to-end message transfer delays.

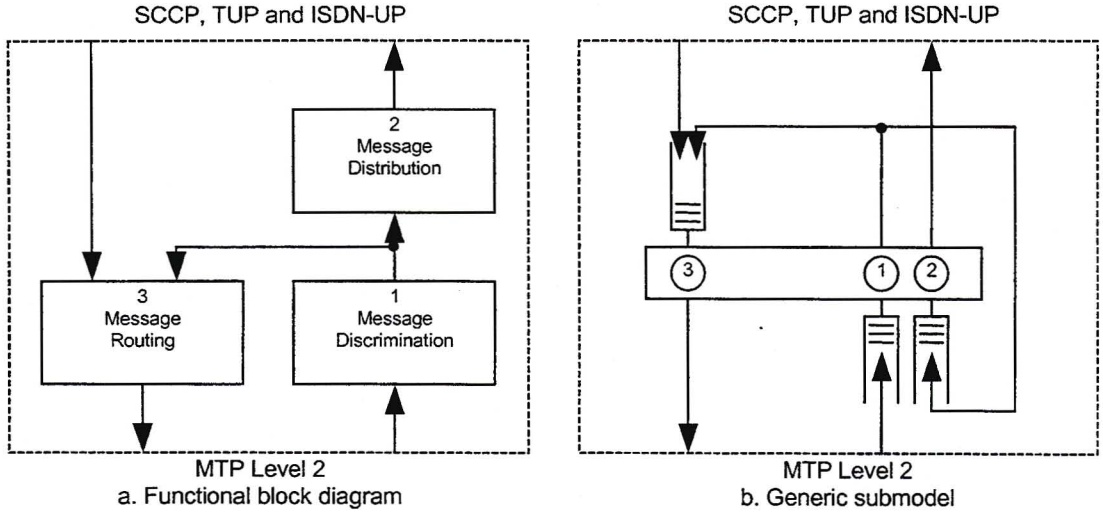


Figure 2-2. Generic submodel and functional block diagram of MTP level 3 [Willmann and Kühn, 1990].

A literature survey by Conway [1991] presents an analysis that is analogous to applying Norton's Theorem for solving queuing networks in a hierarchical manner. This method involves initially analysing the lowest layer, with the inputs to the second layer short-circuited with the outputs from the second layer. The submodel is analysed for each chain of customers. Thereafter the second layer is analysed, with the inputs and outputs interfacing with the third layer short-circuited, while the inputs and outputs connected to the first layer are connected to a flow-equivalent queue. A flow-equivalent queue accounts for the delay experienced by messages in the lower layers. The higher layers are analysed in a similar manner. Once all the layers have been analysed, the throughput and end-to-end response times between peer entities is obtainable.

One shortcoming of this method is that it does not consider multiple layers with shared memory and processing resources. The only interaction between layers is through the exchange of the messages via open connections. In addition, this modelling approach does not produce a product form equilibrium state distribution and as a result known efficient algorithms for queuing networks may not be applied to solve the analytical model [Conway, 1991].

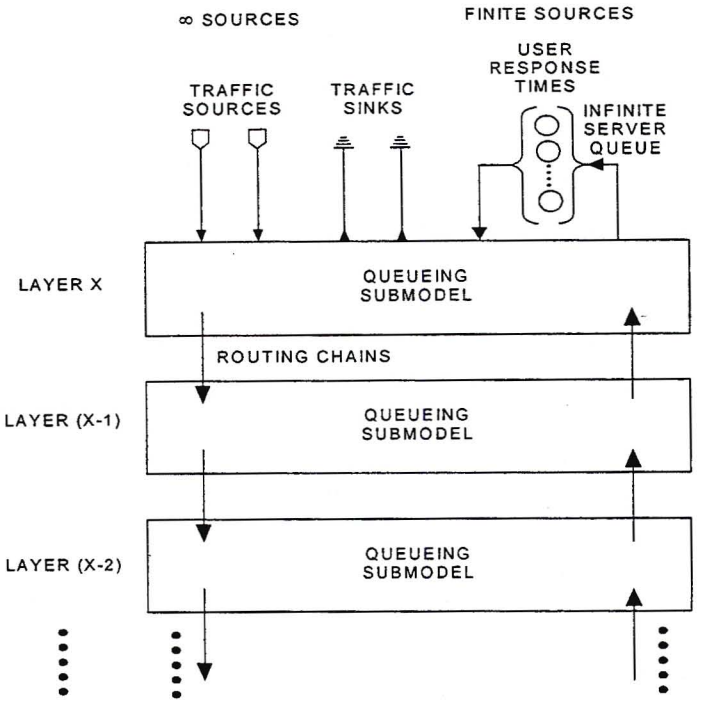


Figure 2-3. Interconnection of queuing submodels (adapted from [Conway, 1991]).

2.3.2 Iterative Decomposition of a Generic Model

Conway's [1989] modelling methodology is similar to that of the Hierarchical Decomposition and Aggregation method (Section 2.3.1), except that his model is based on the generic specifications of the OSI reference model. The model considers the processing overheads at each node, the underlying network topology and accommodates for multiple open systems.

This generic model accounts for the entire network structure, as shown in Figure 2-4. The functional units within each network layer are referred to as entities, and some layers may have more than one entity. Entities are able to carry out functions like flow control, blocking, relaying, error detection, etc. Application layer entities are connected to open and closed sources similar to those described in Section 2.3.1. Entities within each system in the network are interconnected through service access points, which are modelled as queues, while open systems or nodes at different locations within the network are interconnected via a queuing model of the physical layer. In addition, the processing overheads associated with each entity are modelled as queues with various service disciplines. Processing queues may also be shared by different entities in a node; to account for systems with shared processing resources. Each source is then associated with a routing chain that defines the path followed by customers through the network.

To analyse the system, the original network is decomposed and constructed into subsystems. This is done by replacing part of the network with flow-equivalent queues, which represent the delays incurred by customers outside the subsystem. In the determination of mean performance measures, the subsystems are solved in succession while continuously updating the service-time requirements of the flow-equivalent queues. This procedure is repeated iteratively, until convergence is obtained in the mean queue lengths of each subsystem or until some limit in the number of iterations is reached.

The main shortcoming of this methodology is that it is difficult to verify the accuracy of the technique for large complex networks [Conway, 1991].

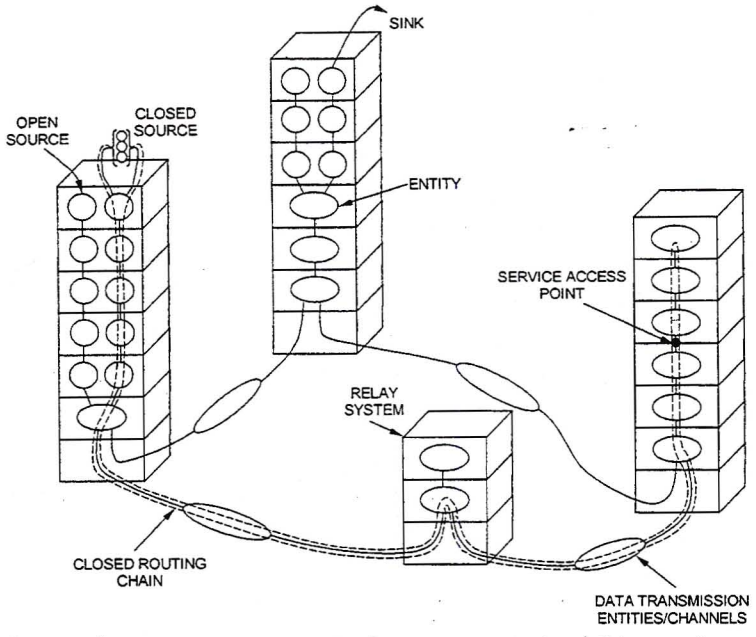


Figure 2-4. Network composed of open systems, entities and sources (adapted from [Conway, 1989]).

2.4 Performance Modelling and Analysis of SS7 Networks

Early performance studies of SS7 networks concentrated on the throughput and delay characteristics of the system. Since then, network failures and measurements of signalling traffic loads have led to a number of studies addressing the reliability of the SS7 protocols. Growing interest in the application of intelligent network services has also motivated researchers to examine the influence of new and emerging services on the existing CCS networks.

Previous studies assumed link bandwidth to be the primary bottleneck in CCS networks. But, as the integration of ATM and other high-speed technologies into existing telecommunications networks addresses the link bandwidth problem, the growth of mobile services, IN applications and other processing intensive services has shifted attention towards processing capacity as being the primary bottleneck. Subsections 2.4.1 and 2.4.2 discuss research investigations that analysed the impact of link level and processor congestion on network performance.

A number of studies have analysed the effectiveness of the current SS7 congestion and flow mechanisms in different overload scenarios. Subsections 2.4.3 through to 2.4.6 discuss the conclusions obtained from these studies together with their proposed improvements to address the shortcomings of the current implementation. Next, Subsections 2.4.7 and 2.4.8 analyse link level delays and consider the effect of clustered arrivals on the delay performance. Subsection 2.4.9 discusses the correlation that exists between call holding times and signalling traffic, and lastly Subsection 2.4.10 examines previous work on the impact of number portability and GPRS on signalling network load.

2.4.1 Link Level Congestion

Signalling links are prone to become congested under abnormal conditions or when traffic loads on the link tend to exceed the engineered maximum. Unexpected overloads can occur during link or node failures or when high link error rates induce link changeovers, or when processor congestion at a node indirectly causes link transmit congestion. Unbalanced loads due to IN services or ISDN supplementary services can also invoke link level congestion.

Skoog [1988] considered the criteria necessary for engineering signalling links carrying ISDN user-to-user information (UUI) message traffic. His results show that links carrying UUI traffic have to be engineered for a lower utilisation than the links that only carry call set-up traffic. In addition, during congestion the average message length of the controlled traffic can influence the time taken for the buffer occupancy to drop below the abatement threshold. For example, consider a link using the NOCP congestion control mechanism with two independent Poisson traffic streams as its input. One stream consists of UUI messages with a congestion discard priority of zero and the other stream consists of ISUP call set-up traffic with a congestion discard priority that is greater than zero. If only traffic that is controlled at the first onset threshold, O_1, is considered; then the expected transmit buffer occupancy, $\overline{N}_t$ at time t, is given by [Skoog, 1988]

$$\overline{N}_t = O_1 + M_0\left(\lambda_0\tau_0 + K - Ke^{-\lambda_0(t-\tau_0)/K}\right) + M_1\lambda_1 t - Ct, \qquad 0 \le t < \tau_r \qquad \textbf{(2-15)}$$

where the parameters for UUI messages, and ISUP call set-up messages are written with subscripts of 0 and 1 respectively. M_i is the average message length and λ_i the arrival rate of traffic stream i. C is the link speed and K is the number of source-destination pairs contributing to the UUI traffic stream. τ_0 is a reference time denoting when O_1 was reached and τ_r is the time taken to reach the abatement threshold, A_1, after time τ_0. The transmit buffer occupancy at time '$\tau_r + \delta$' is given by

$$\overline{N}_{\tau_r+\delta} = A_1 - \left[1 - \rho_1 - \rho_0\left(e^{-\lambda_0\tau_r/K} + \zeta\right)\right]C\delta - \zeta KM_0\left(1 - e^{-\lambda_0\delta/K}\right), \quad 0 \le \delta < t_c \qquad \textbf{(2-16)}$$

where

$$\zeta = \left(1 - e^{-\lambda_0\tau_r/K}\right)\left(1 - e^{-\lambda_0 t_c/K}\right), \qquad \rho_i = \frac{\lambda_i M_i}{C}, \qquad t_c = \mathrm{T15} + \mathrm{T16}.$$

The above analysis assumes that the traffic arrival rate is greater than the link speed, C, i.e., $(\rho_0 + \rho_1) > 1$. In addition, $\rho_1 < 1$, which implies that after the UUI traffic is removed the link will no longer be congested. Figure 2-5 shows the expected buffer occupancy for different UUI message sizes.

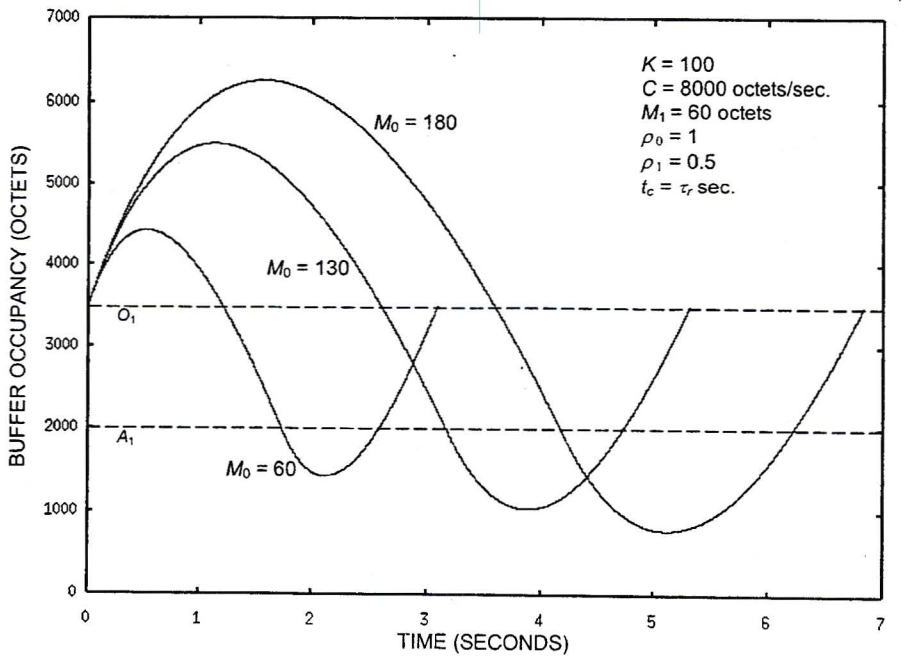


Figure 2-5. Expected buffer occupancy during congestion control.

Figure 2-5 clearly illustrates the oscillatory behaviour of the buffer occupancy during congestion. The net effect of an increased buffer occupancy, when larger UUI message lengths are used, is an increase in the average delay experienced by the high priority traffic. These graphs also highlight the importance of selecting appropriate timer values and threshold settings, for example, if the abatement threshold is too small or T15 is too large, the affected link would be under utilised once the traffic sources are throttled.

Zepf et al. [1991] analysed link level congestion with feedback control on a link set carrying an asymmetric load distribution. Their analysis is based on the International Option and utilises a hybrid iterative analysis technique to solve the equations. The results show that the route set flow controls throttle traffic on all the links of the link set and is therefore unfair to some users. They propose an alternate flow control mechanism, called the congested link method (CLM). In the CLM, the congestion indications that are sent to the user parts indicate which link is congested. The user parts then respond by only throttling traffic on the affected link, while the other links continue to carry their full loads. Alternatively, all new call traffic is routed via the uncongested links, until the congested link has returned to normal. As expected, the CLM significantly improves performance in an asymmetric load scenario. However, the CLM method is only effective if congestion occurs on the outgoing links of a SP, since TFC messages from STPs do not indicate which route on link in a route set is congested.

Nagarajan [1999] analysed the congestion controls of a signalling link from a BSC to a MSC in a GSM network. In his model a congestion onset message is sent to the SS7 user part when the link's buffer occupancy exceeds the onset threshold and a congestion abatement message is generated when the buffer occupancy drops below the abatement threshold. In his analysis three types of traffic streams were considered, namely:

- call processing traffic, including call set-up and mobility management,
- operations, administration and maintenance traffic for network administration, and
- all other traffic or miscellaneous traffic.

The following conclusions were obtained from his analysis:

- The throughput of call processing messages is insensitive to the choice of thresholds unless the abatement threshold is too close to zero, which is when link idling becomes apparent.
- Partial rather than complete throttling of call processing traffic has the advantage of increasing the call throughput at higher loads.
- The congestion control overhead is sensitive to the choice of thresholds. A lower abatement threshold results in a higher control overhead.

Nagarajan also suggests that during throttling, new calls and new location update traffic should first be dropped in favour of revenue generating calls already in progress. A token based method of calculating the number of call originations and location updates that should be admitted during congestion is also provided. This study provides a good understanding of the influence of generalised message streams on signalling link buffers, but it does not consider the effect of congestion on application level procedures.

2.4.2 Processor Congestion

Processor congestion occurs at a node when the message load arriving at MTP level 3 is faster than the processor's message handling capability, and as a result the processing queue is eventually exhausted. Under normal conditions, link utilisation is usually much lower than that of the routing processor. A small yet significant increase in the link loads could therefore overwhelm the routing processor. Congestion is also likely to occur when switches, that were originally designed to support the standard 56 or 64 kb/s links, are integrated with broadband technologies such as ATM [Kant, 1997]

Rumsewicz [1993] investigated the performance of message discard schemes during STP processor overload. His analysis considered two levels of message priorities and the effects of customer reattempts and application level recovery procedures. The results obtained show that after a brief overload, a network can continue to remain in the congested state. In a scenario, where all the messages have an equal probability of being discarded, the call completion rate drops close to zero. However, when messages are discarded on a priority basis the call completion rate drops to zero once the STP becomes congested. The results also show that network traffic is dominated by REL and RLC messages during the congestion period. The high network traffic load is maintained by reattempts from application level recovery procedures and customer reattempts, hence call throughput suffers. He states that the call completion rate is a more appropriate measure of signalling network performance than message throughput, as it more reliably indicates the quality-of-service from the customer's perspective. In addition, he states that the low call completion rates and sustained overloads illustrate the importance of implementing a feedback control mechanism in MTP level 3.

Rumsewicz's simulation results suggest that network performance is sensitive to the REL reattempt timer values. Simulations were conducted with a constant and variable REL timer. The results obtained with the variable timer correspond with those obtained analytically, while the call completion rate obtained with the constant REL timer was considerably higher during congestion. This discrepancy could be a consequence of the simulation not having modelled the transmission links and as a result the observed REL messages arrive at the STP level 3 queue in bursts.

Kant [1997] examined call throughput when ATM links are deployed in an existing SS7 network. This investigation is concerned with signalling links based on the high-speed ATM signalling link protocols in [ITU-T Recommendation Q.2210]. In the example given, a high-speed link is deployed due to a lack of communication ports and not due to an expected increase in traffic, e.g. one 1.5 Mb/s high-speed link can replace multiple 56 kb/s links and make physical ports available for use. In this scenario, the higher level processor does not have to be upgraded, unless the free ports are expected to carry significantly more new traffic. His simulation implements flow control at the link level and NOCP congestion control at level 3. The STP model used is based on a central processor architecture, but Kant states that the results also apply to distributed architectures. The results from the simulation indicate that link level flow control is inadequate, and MTP level 3 controls are necessary to obtain an acceptable call throughput. On its own, the level 3 congestion control performed reasonably well, but provided an undesirable global throttling, whereas the advantage of link level flow control is localised throttling. A combination of both the level 2 and 3 flow controls is therefore necessary to obtain the desired effect.

Kasera et al. [2001] analysed and compared the following processor overload detection algorithms:

- The Occupancy Algorithm measures load by examining the percentage of time that the processor is busy over an interval. This value is used to determine how many calls should be throttled in the next interval.
- The Signalling Random Early Detection Algorithm is a modified version of the Random Early Detection algorithm that is commonly used in IP routers.
- The Acceptance-Rate Occupancy Algorithm uses both the call acceptance rate and processor occupancy to determine the target call acceptance rate.

Their results show that the Signalling Random Early Detection Algorithm does not perform well for very high call arrival rates. Its throughput deteriorates significantly during steady state conditions. In the transient analysis the Signalling Random Early Detection Algorithm performed the best under a moderate load, while the Occupancy Algorithm performed poorly soon after the overload was introduced. The performance of the Acceptance-Rate Occupancy Algorithm was intermediate to that of the other algorithms, however it can be tuned to closely match the performance of the Signalling Random Early Detection Algorithm.

2.4.3 *Sustained Overloads*

Manfield et al. [1994] investigated the performance of various congestion control strategies in a sustained overload scenario. Their simulation employed a queuing model to examine the behaviour of a congested signalling link between two STPs. This simplified modelling approach does not consider message sequencing, application level recovery actions and the impact of caller reattempts. However, the simulation does account for message lengths, round-trip delays and the number of traffic sources in the network. A comparison between the IO and NOCP controls showed that the link level queuing delays are lower when the IO controls are used. Furthermore, varying T29 and T30 produced only a marginal improvement in the delay/throughput performance, and network performance is found to be insensitive to the selection of the T15 and T16 timer values. A priority based implementation of the IO congestion controls that throttles messages in ISUP, according to their NOCP congestion priorities, produced results that were similar to those obtained with the NOCP congestion controls. They concluded that user part congestion controls are more efficient than the MTP congestion controls since the flow controls are more closely coupled with the call processing functions, and messages are blocked before being processed in MTP.

Zepf & Rufa [1994] analysed a similar overload scenario, except that the objective of their investigation was to determine the impact of intelligent network services on the current congestion and flow control implementations. Their simulation model is more detailed than the one used by Manfield et al., and includes link level modelling of the propagation delays, transmission and retransmission buffers and protocol specific functions; such as error correction, congestion thresholds and flow control using SIB messages. An important observation from their results shows that since TFCs are sent on a message basis, the sources transmitting a large number of short MSUs are throttled more severely than the sources transmitting long MSUs. In addition, the congestion controls are also found to be ineffective if a very large number of users are transmitting over the same congested link. The results of their investigation indicate that there is no significant reduction in the signalling traffic load and more than half of the traffic over the congested link consists of RCT messages. Zepf & Rufa concluded that the IO and the NOCP do not work correctly in an environment with intelligent network services and a large number of users. They propose the use of a single TFC and RCT message to and from defined SP regions. For example, one TFC message could be sent to an entire signalling point cluster, where a local STP would duplicate and redistribute the message. A single RCT message from each region could also be used to probe the affected route's congestion status. However, this scheme would only be applicable in networks where the point codes can be used to distinguish between different networks and signalling point regions.

2.4.4 *Effectiveness of the TFC Procedure*

Smith [1994] studied the performance of the NOCP congestion control procedures in the presence of network latency (i.e. the delay between the onset of congestion and the reaction of traffic sources to the congestion). He modelled a single STP network, where the link between the STP and a target SP is congested during a focused overload. The call completion rate was used to measure the effectiveness of the congestion controls and the following results were obtained:

- The TFC procedure overcontrols traffic during congestion. Consequently, the update delay, caused by network latency, suppresses traffic from too many sources. Furthermore, the slow response of the MTP flow control timers prolongs the period of traffic suppression.
- Latency leads to traffic synchronisation between the source nodes and consequently leads to oscillation of the transmit buffer's queue length.

- Overcontrol and synchronisation between the traffic streams leads to a large number of message discards during the oscillatory peaks, and eventually induces an REL avalanche.
- The release message's reattempt behaviour aggravates congestion and impedes recovery once the overload traffic has subsided.

The modelling approach used attributes equal and constant delays to the STP processing queue, the link transmit queue and the link propagation delay. This constant update delay of all the sources, in Smith's simulation model, could have contributed largely to the highly oscillatory behaviour observed in the network traffic streams. Smith suggests that reducing the time intervals of the T15 and T16 timers has the effect of increasing the call completion rate, as the link idle time is reduced. To avoid REL avalanches, Smith also suggests that all the IAM messages should be suppressed at the sources until the REL backlog has been diminished.

Rumsewicz [1994] investigated the transient performance of a single STP network, in order to determine the efficacy of the TFC procedure in the NOCP. His simulation shows that TFC messages are generated in short oscillatory bursts and that the origination/destination pairs eventually become synchronised in their actions. However, the lack of a link level queuing model could have contributed to the highly periodic TFC message generation rate. His overall results show that a TFC message should be sent for one out of every eight low priority messages received, in order to obtain a reasonably good call completion rate. Whereas, in simulations where TFC messages are sent more frequently, the call completion rate is degraded by the large number of TFC and RCT messages present in the network. Furthermore, the use of a staggered congestion threshold scheme for each message priority is found to provide a superior performance. His results also confirm that network traffic via the congested entity is dominated by REL messages.

2.4.5 Multiple User Part Congestion Control Interactions

Mayer [1997] examined the interaction between ISUP and SCCP congestion controls, in a scenario where the services of both user parts are required to set-up a call. Here, the issue of fairness during congestion is important, as both user parts implement different flow control procedures. In his simulations, Mayer assumes that the SCCP query and response messages are assigned 8 priority levels and are throttled across 32 load reduction steps while ISUP traffic is throttled across 10 steps with the International Option. It therefore takes longer to reduce the SCCP load and his results show that during congestion ISUP traffic is blocked more severely than SCCP traffic. Unequal traffic loads from each user part can result in one user part's traffic dominating in the use of the congested resources and can therefore result in little or no useful throughput when the services of more than one user part are required. The following factors were found to contribute to the lack of load balancing that exists between the two flow controls [Mayer, 1997]:

- The ISUP and SCCP algorithms have a different number of reduction steps.
- Their traffic characteristics vary. ISUP has 70% of its octet traffic in the first message (IAM) while SCCP has 40% or less of its traffic in the initial query message.
- SCCP messages are on average five times longer than ISUP messages. ISUP messages therefore received five times more TFCs for the same amount of octet traffic. This can be avoided by sending TFCs for every n octets received, rather than for every n messages received.

The simulation results also showed that no stable point exists where the load can be balanced equally between both user parts. The load originating from each user part during congestion is influenced by several factors in the network's configuration, which inadvertently results in one user part dominating in the use of the congested facilities. Consequently, transactions requiring an interaction between both user parts are likely to fail.

2.4.6 Combined Control

Rumsewicz & Smith [1995], and Northcote & Rumsewicz [1995] compared the congestion handling capability of the IO, the NOCP and a combined control (CC), which incorporates both the NOCP and IO congestion controls. Their work examines link level congestion and STP congestion, respectively. In the CC the NOCP functions are implemented in MTP and the IO functions are implemented in the user parts. The STP uses the NOCP flow control mechanisms and sends TFC messages indicating the current level of congestion in the affected queue. A TFC arrival at a SP invokes both the NOCP and IO congestion controls. For example, when TFC messages with a congestion status of one arrive at a SP, the NOCP congestion control blocks all the IAM messages until T16 expires, thereafter the IO continues to control the IAM traffic load and gradually increases the load as the T30 timer expires. Their simulations considered various overload scenarios, including network wide increases and focused loads to a single node. The results obtained show that the network traffic is oscillatory when the International Option is used, because the T30 timer usually expires before the next TFC message arrives. However, they state that the International Option on its own can act as an effective control mechanism. The National Option with congestion priorities performs poorly during network wide overloads, but has the advantage of producing a relatively flat call completion rate profile. The combined control inherits the desirable characteristics of both the IO and the NOCP, such that it has a high throughput and a flat call completion rate profile. The combined control scheme performed as well as the better of IO and NOCP controls in most of the overload scenarios that were analysed. The authors conclude by stating that the CC deserves consideration, especially in networks currently employing the NOCP, as it may provide an added robustness under a variety of congestion situations.

2.4.7 Link Level Queuing Delays in Signalling Networks

Transmission delay is an important measure of signalling network performance. The most important factors that influence the end-to-end transmission delay within a signalling network are the propagation delays, the processing delays endured within a node, the number of network components encountered along a route and the traffic characteristics. In order to obtain a good approximation of link transmission delays, the links can be modelled as M/G/1[1] queues with non-preemptive priority queuing where the average waiting time, $E(W)$, for a link (in the absence of errors) is given by the following equations [Schwartz, 1987]:

$$E(W) = \tfrac{1}{2}T_f + \tfrac{1}{2}\lambda\frac{E(\tau^2)}{(1-\rho)} \quad \text{for basic transmission} \qquad \textbf{(2-17)}$$

and

$$E(W) = \frac{(1-e^{-\lambda T_L})E(\tau^2)}{2T_m(1-\rho)} + \frac{e^{-\lambda T_L}T_f}{2(1-\rho)} \quad \text{for preventive cyclic retransmission} \qquad \textbf{(2-18)}$$

where

T_f	is the emission time of fill-in signal units,
T_m	is the emission time of message signal units,
T_L	is the round trip propagation delay,
λ	is the arrival rate of message signal units,
ρ	is the link utilisation due to message signal units, and
$E(\tau^2)$	is the second moment of the message signal unit emission time.

Formulas for the queuing delay in the presence of disturbances, with error retransmission, are given in the ITU-T Recommendation Q.706. Research in this area still continues, as the M/G/1

[1] A *M/G/1* queue has a Markov (*M*) arrival process and a single server with a General (*G*) service process.

queuing model does not account for the effects of non-Poisson traffic streams, from TCAP and ISDN services, on the queuing delay. Likewise, no equivalent analysis exists for the propagation delays on high-speed ATM signalling links and SIGTRAN IP signalling links.

In a network with a mixture of long and short messages, the short messages tend to suffer from *agglutination*, whereby they experience longer delays and accumulate behind the long messages. Gianini & Pettitt [1997] examined this aspect, by modelling a system with a large number of sources and subsequent queues along each of the paths followed by the messages. The traffic arriving at the sources was composed of two traffic streams, each with different distributions of the long and short messages. At the final node, one traffic stream is removed while all the messages from the other traffic stream are allowed to enter the queue unimpeded. They discovered from their analysis of the final node that if there is a blending of traffic at the input to the final queue then the M/G/1 model is sufficient for the calculation of queuing delays.

Lazar et al. [1992] used the BCMP model by Baskett, Chandy, Muntz & Palacois [1975] to determine the mean and worst case packet delays in the Singapore SS7 network (Figure 1-2) during multiple STP node failures. By assigning a different class to traffic from each region of the network, they were able to calculate the end-to-end delay experienced by packets from each region during node failures. They later use their analytical model to complement the results obtained from a SS7 emulator [Lazar et al., 1994]. In their model there are M uni-directional links or queues and K sources of traffic, with a Poisson transmission rate of λ_i ($1 \le i \le K$). The steady state probability expression for the network is then given by

$$\pi(\overline{n}) = \pi_0 \prod_{i=1}^{M} \left(\frac{\overline{n}_i!}{\prod_{k=1}^{K} n_{ki}!} \prod_{k=1}^{K} \left(\frac{\lambda_{ki}}{\mu_i} \right)^{n_{ki}} \right), \tag{2-9}$$

where

$$\pi_0 = \prod_{i=1}^{M} \left(1 - \sum_{k=1}^{K} \frac{\lambda_{ki}}{\mu_i} \right). \tag{2-10}$$

The expected number of class l packets in the jth queue is given by

$$E[n_{lj}] = \left(\frac{\lambda_{lj}}{\mu_j - \sum_{k=1}^{K} \lambda_{kj}} \right) \tag{2-11}$$

and the mean number of class l packets in the network is given by

$$E[\overline{n}_l] = \sum_{j=1}^{M} E[n_{lj}] = \sum_{j=1}^{M} \left(\frac{\lambda_{lj}}{\mu_j - \sum_{k=1}^{K} \lambda_{kj}} \right). \tag{2-12}$$

By applying Little's formula, to the above equation, the mean time delay (T_l) for class l packets in the network is given by

$$E[T_l] = \sum_{j=1}^{M} \left(\frac{\lambda_{lj}}{\left(\mu_j - \sum_{k=1}^{K} \lambda_{kj} \right) \lambda_l} \right) \qquad \textbf{(2-13)}$$

and the mean time delay (T) for an arbitrary packet in the network is given by

$$E[T] = \sum_{j=1}^{M} \left(\frac{\sum_{k=1}^{K} \lambda_{kj}}{\left(\mu_j - \sum_{k=1}^{K} \lambda_{kj} \right) \sum_{k=1}^{K} \lambda_k} \right). \qquad \textbf{(2-14)}$$

2.4.8 Clustered Arrival Processes

In a signalling network the call arrival process is Poisson in nature, but the message arrival process at a node is not. For example, an IAM message is generated when a call is initiated (and can therefore be modelled as a Poisson process), but the generation of the subsequent messages of the call set-up procedure depends upon various system specific processes and factors (the REL reattempt process, for example, is a deterministic process). However, if the time separation between the correlated arrivals is large enough, the dependency between these messages can be ignored. Skoog [1991] modelled clustered arrival processes in signalling networks. His simulations confirm that the M/G/1 model is a good approximation for the determination of link queuing delays, when the interval between consecutive message arrivals is large and the link utilisation is below 80%. During heavy load conditions, the busy periods increase and the effect of interarrival time dependencies becomes apparent. Consequently, through the use of approximations, Skoog shows that the mean queuing delay for heavy traffic conditions can be approximated with the bulk arrival M/G/1 model.

In an ISDN environment users are able to send bursts of correlated messages over the SS7 network. Large user-to-user signalling messages are segmented into consecutive signalling messages that arrive within short intervals of each other. These messages have been found to behave like long messages and they significantly increase link queuing delays. Kosal & Skoog [1994] investigated a control mechanism to regulate the flow of correlated messages entering the signalling network. Their method attempts to increase the time separation between correlated messages in order to allow the dependency between them to be ignored. A credit manager (CM) at the source nodes was used to regulate the flow of correlated messages into the link transmission buffer. Small ISUP and TCAP messages are allowed to directly enter the link transmission queue, while the correlated messages are first queued at the CM. In the CM, credits are generated at a fixed rate and are allowed to accumulate until the maximum capacity of the credit bank is reached. The correlated message at the front of the CM queue is only allowed to enter the link transmission queue if the accumulated credit is greater than or equal to the message's octet length. If not, the message waits until the required credit has accumulated. Once the message has left the CM, the credit bank is reduced by an amount equal to the size of the message. Comparisons between the CM scheme and a M/G/1 queue showed that the queuing delay was significantly reduced when the CM was used. Depending on link utilisation, the time separation between correlated messages ranges from 0.2 to 1 second. As the time separation is increased the maximum throughput of the credit manager decreases and the available buffer capacity can be used by Poisson traffic, without enduring any significant queuing delays.

2.4.9 *Traffic Characteristics*

Most analytical models assume exponential call holding times. Bolotin [1994] used measurements of actual call holding times to develop analytical models. His measurements show that the actual holding time distributions differ from the exponential distribution. The exponential distribution underestimates the proportion of very short calls and the proportion calls lasting longer than the mean holding time. For calls with short holding times, the subsequent signalling messages arrive within close succession of each other and there exists a noticeable correlation between these messages. Consequently, call behaviour in the trunk network strongly dictates the behaviour of traffic in the signalling network. In addition, signalling traffic is also influenced by ISDN and IN services, which are responsible for highly correlated message streams (discussed in Section 2.4.8).

Bolotin states that calls within the telephone network can be isolated into several distinct classes, where each class represents a specific type of call; e.g. normal answered calls, calls with a busy condition, facsimile, voicemail, etc. Each class has its own holding time distribution; for example, voicemail calls usually have a short duration, while internet connections usually have a very long call holding time. Bolotin then demonstrated that the holding time distribution, $F(x)$, can be represented by a mixture of different call holding time distributions, as shown below:

$$F(x) = \sum_i p_i F_i(x), \quad \sum_i p_i = 1 \qquad \textbf{(2-19)}$$

where, probability p_i is the proportion of ith class calls in the mixture and $F_i(x)$ is the holding time distribution of the ith class. However, identifying the exact proportion of each class can be difficult and estimations have to be used to obtain distributions that match the measured holding time distributions. In later work by Chlebus [1997] on call holding time data in cellular networks, the lognormal distribution was found to provide the best fit but the author concludes that the exponential distribution is still suitable for modelling call holding times.

2.4.10 *Impact of New Services*

The introduction of new services leads to additional signalling messages (to support the new services) and possibly an increase in the size of existing signalling messages (e.g. due to larger subscriber profiles in mobile networks). Local Number Portability provides a person with the ability to retain a telephone number after switching to a new network operator, service provider or after moving to a new geographic location. In a Public Switched Telephone Network (PSTN) telephone number ranges are assigned to each exchange. A call is routed by examining the first few digits of the called number and circuits are seized to the destination exchange responsible for that number range. In order to support number portability, calls to a number range from which a number has been ported out are routed using a Location Routing Number, which is obtained by querying a Service Control Point (SCP). In [Chandramouli & Krishnan, 1999] the authors analysed the impact of number portability on a SS7 network where different proportions of number ranges contain ported numbers. Their results show that the signalling load increases by 48% when SCP queries are performed for all call attempts.

GSM operators are currently deploying (or have recently deployed) General Packet Radio Services (GPRS), yet very little literature exists on how it can impact the signalling network. Mayer [2000] investigated the impact of GPRS signalling on a GSM network via analytical and simulation models, but does not provide any details on the models used. He considered the following scenarios in a network with five MSCs (with integrated VLRs), two HLRs and one STP:

a) Two of the MSCs have combined Serving GPRS Support Nodes (SGSNs) and the other three are standalone MSCs.
b) All the MSCs have combined SGSNs.

His results show that for scenario A, the signalling load more than doubles on the links between the MSCs with combined SGSNs and the STPs, compared to a less than 60% increase for scenario B. Equivalently, significantly longer response times are also experienced in scenario A. These differences are due to the high concentration of GPRS subscribers on the two SGSNs in scenario A, while in B the subscribers (and their resulting signalling traffic) are distributed across more SGSNs. The graphs presented for the signalling traffic between the HLR and STP appear to be similar for both scenarios. However, one would naturally expect to see a higher signalling load on the HLR's links of scenario B, as more location updates would be the consequence of having more SGSNs.

2.5 The Effect of Message Reattempts on Network Performance

Repeated call attempts, from customers who were previously unsuccessful at establishing a call, can severely impact network performance. Calls that are initially rejected are often repeated within a short period of time. Re-dial and call-back facilities provided by modern telephones further exacerbate the traffic load due to caller reattempts during periods of overload. In addition they also produce highly correlated call attempts. The effect of repeated calls was addressed by Cohen [1957]. His model extended Erlang's loss model to account for the repetition time and staying time of lost calls in the system [Gross & Harris, 1985, p.101]. Since the analysis exhibits considerable complexity the interested reader is referred to Cohen [1957] or Syski [1986].

In packet switched networks, network layer recovery procedures are in place to ensure that packets are successfully transmitted to their respective destinations. In most implementations, if no acknowledgement is received within some timeout period, the source will reattempt to send all unacknowledged packets. This reattempt behaviour can have devastating effects on message throughput, especially during congestion situations. Rumsewicz [1993] observed that call completion suffers in SS7 networks when the REL recovery actions are activated during congestion. Release messages dominate the network traffic and as a result throughput of the other message types (including RLC messages) suffers. He also noticed that, due to caller reattempts, the network could remain in a congested state even after the overload traffic has been removed. A similar behaviour was observed by Kleinrock and Lam [1975] in slotted ALOHA systems with a finite population, where retransmissions resulted in the system having three equilibrium regions.

2.6 Mobility Management Research

Various researchers have analysed and proposed various mobility management techniques for mobile networks over the past decade. These efforts have generally used GSM and IS-41 as the basis for further research, by trying to address some of the shortcomings associated with these protocols. Their basic objectives were to attempt to reduce the signalling load and HLR transaction load. Proposals can generally be categorised as being either memoryless or memory-based methods [Tabbane, 1997]. The following subsections provide an overview of mobility management research. The interested reader is also referred to Tabbane [1997], Akyildiz et al. [1999] and Wong & Leung [2000] for comprehensive discussions on mobility management research and an extensive list of references.

2.6.1 Signalling Load in Mobile Networks

The additional signalling traffic load present in mobile networks due to call handovers, mobility management and authentication was analysed in [Pollini et al., 1995] and [Pollini et al., 1996].

The signalling load attributed to these procedures was determined from calculations of the number of handovers/call and location updates/call:

$$\frac{Handover}{Call} \leq 1.4 \frac{1}{\sqrt{N_{SW}}} \frac{V}{\lambda \sqrt{A_{cell}}} E_{term}, \text{ and}$$

$$\frac{Location\ Update}{Call} \leq 1.4 \frac{1}{\sqrt{N_{LA}}} \frac{V}{\lambda \sqrt{A_{cell}}} (1 - E_{term})$$

where N_{LA} is the number of calls per location area and N_{SW} is the number of cells per MSC. The average velocity of a mobile terminal (V), the cell area (A_{cell}) and the call arrival rate (λ) together determine the ratio of mobility to the call arrival rate. The constant (1.4) is used to account for the effect of cell geometry on the analysis. The main conclusions obtained from their calculations are as follows:

- GSM generates almost twice the signalling load of IS-41B, due to GSM's authentication procedures. However IS-41C, which includes authentication and is based on GSM is expected to perform in a similar manner to GSM.
- GSM generates 50% more signalling load with physically separated VLR and MSC nodes than the scenario where the MSC and VLR functionality resides in a single node.
- The signalling load is lower if multiple location areas are present on a few higher capacity MSCs than in the case where only one location area is present on each MSC (assuming each location area has a fixed size).

Additionally, signalling load and MSC/VLR processing load is also influenced by coverage area. Tabbane [1998] showed that an increase in the surface area by a factor of 25 results in an increase in processing load by a factor by approximately 5 because the number of handovers and location updates decreases as the cell area increases. He also finds that the coverage area of a rural MSC/VLR can be 10 times larger than that of an urban MSC/VLR to encounter the same processing load. This is due to the higher subscriber density in urban areas.

In GSM, a call between two subscribers who are registered on the same VLR generates the same MAP signalling traffic as the scenario where both subscribers are registered on different VLRs. Schopp [2000] compared standard GSM with a modified GSM, where the called party is first searched for at the local VLR before a query is sent to the HLR. As expected, the modified GSM scheme was found to perform better than standard GSM for different network topologies.

2.6.2 *Memoryless Mobility Management Methods*

2.6.2.1 Hierarchical Database Architecture

Database transaction load can be reduced through the design of an optimum database architecture. GSM is based on a centralised database architecture, where the HLR has to be queried for every call or service that needs to be delivered to a user, even if the called and calling parties are registered on the same VLR. Furthermore, the subscriber's profile needs to be transferred to a VLR every time a subscriber enters its coverage area and cancelled from the PVLR.

Wang [1993] proposed a distributed hierarchical database architecture as a method of reducing the distance travelled by signalling messages, and minimising the transaction load and delay on databases (Figure 2-6). When a call arrives (e.g. from MT2) the request is routed up the database tree until it arrives at a database that has a pointer to the called terminal (MT1). The request is routed down the tree with the help of pointers in the subsequent nodes. A call request from the *University* database (from MT3) would only progress up to the *Stanford* database, thereby reducing the load on high-level databases. Once the call request has been accepted a

direct path is used for call set-up. The disadvantages of Wang's method include; very large capacities are required in the higher level databases, multiple databases may need to be updated for subscriber's who move large distances and a call request delay may be greater for calls to subscribers who are not located locally. To address the higher layer database requirements, Eynard et al. [1995] analysed a hierarchical database architecture where only the lowest layer nodes have a HLR, while the higher nodes only assist the network in locating the correct database.

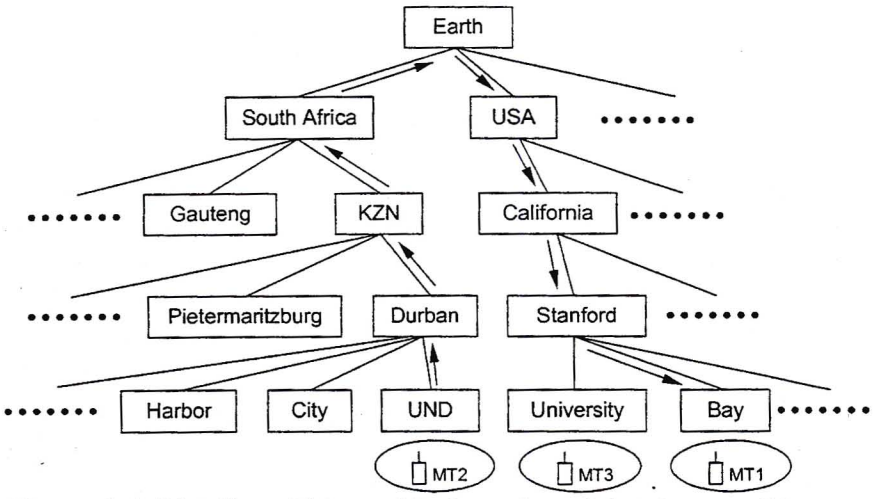


Figure 2-6. Distributed hierarchical tree based database architecture (adapted from [Wang, 1993]).

Schopp [1997] compared Wang's architecture with the GSM architecture by including a mobility model and call model in his analysis. In his analysis the various functional elements of a node are represented by queues, similar to the modelling approach proposed in [Willmann and Kühn, 1990]. The mean search delay for a called subscriber is found to be dependent on the call distribution matrix, but generally the distributed architecture has a mean search delay that is better or equivalent to the GSM architecture.

In [Schopp, 2000] the cost of location updates and calls in terms of the mean number of hops required to perform a location update or to retrieve call routing information is analysed in a multi-layered network. A standard GSM network is compared to architectures similar to Wang's distributed databases for various mobility coefficients. In his results the distributed architecture performs well for very low levels of mobility but the cost increases for higher levels of mobility. He also considers the mobility cost for roaming between multiple operators. His results indicate the following:

- Direct connections between the higher level databases or STPs result in a lower cost.
- Distributing the top level database across the nodes in the lower layer can reduce the cost of calls at the expense of a higher location update cost.
- The addition of Proxy-HLRs can help to further reduce the signalling cost.

Fasbender et al. [1995] extended Wang's model to support paging at a higher level database according to the subscriber's mobility profile. In the example shown in Figure 2-6, if MT1 regularly moves between the *University* and *Bay* databases, Fasbender at al. suggest that MT1 should be paged by the *Stanford* database, instead of allowing the call request to progress further down. Their analysis shows that higher layer paging reduces the average database access time and requires smaller location databases.

While the hierarchical database architecture presents one method of reducing the database transaction load, GSM operators have addressed the processing and capacity limitations of

HLRs by simply adding more HLRs to the network and distributing the subscriber profiles across more than one HLR.

2.6.2.2 Multiple Paging Areas per Location Area

In GSM a location area represents the area within which a newly arrived mobile terminal initiates a location update and it is also the area over which a terminal is paged for incoming calls. Location area size optimisation needs to consider the signalling load generated by frequent location updates between small location areas versus a high paging load in a large location area where a large number of mobile terminals are present. The method discussed in Tabbane [1997] aims to maximise the location area size by splitting it into several paging areas. A mobile terminal would initiate a location update on first entering the location area and thereafter only during periodic location updates. When a call arrives, a paging message is first broadcast in the paging area where radio contact with the terminal was last established and thereafter if the subscriber has not responded the terminal is paged in all the paging areas either globally or sequentially. While this scheme would help to reduce the signalling traffic due to frequent location updates, it could increase the call set-up delay.

2.6.2.3 Pointer Forwarding

Jain & Lin [1995] proposed a pointer forwarding strategy for location tracking when the call to mobility ratio is low. Their method reduces the number of HLR updates by creating a forwarding pointer from the PVLR to the new VLR for each move to a new location area. When a call arrives, a query from the HLR is sent to the first VLR and it is then forwarded along the pointer chain until it reaches the subscriber's current VLR, which responds to the query. This method also imposes an upper bound to the size of the pointer chain to prevent excessively long call set-up times. The pointer forwarding strategy is found to help reduce the signalling load towards the HLR by distributing the location update load between the VLRs, where resources are assumed to be more readily available.

While there has been further research on pointer forwarding techniques ([Lin & Tsai, 1998], [Krishna et al., 1996]) there are a number of shortcomings associated with this technique:

- None of the analytical models account for the transfer of subscriber profiles.
- Pointer forwarding increases the call set-up time.
- The failure of an intermediate VLR (or signalling link-set between VLRs) may lead to broken pointer chains and as a result the HLR would be unable to locate the called subscriber.
- The signalling traffic to the HLRs is reduced at the expense of higher signalling traffic between VLRs.
- In GSM the VLR has a higher transaction load than the HLRs, due to the frequent call processing and mobility management transactions triggered by mobile terminals. Pointer forwarding simply adds additional load to the already heavily loaded VLRs.

2.6.2.4 Lightweight Call Delivery Protocols

Cui et al. [1998] proposed two lightweight call delivery protocols. In the Lightweight Location Lookup Protocol (LiLLP) the originating MSC queries the HLR for the MSC address of the called party (Figure 2-7). The originating MSC then sends a connection set-up message directly to the called party's MSC. The LiLLP differs from the GSM call delivery procedure, as the HLR does not query the called party's VLR for a roaming number. The LiLLP is therefore found to significantly reduce signalling load and call set-up delay. However, this method lacks the ability to detect a detached subscriber until after circuits have been reserved to the destination MSC. Furthermore, this approach is not compatible with GSM's VLR fault recovery procedures (ETSI GSM 09.02), where the VLR attempts to restore a subscriber's profile after receiving a PRN message.

In the Reverse Connection Set-up protocol a connection request message is sent from the originating MSC to the called party's HLR. This message is forwarded to the called party's MSC, which initiates paging of the called subscriber. If the called party is able to accept the call, a circuit connection is set-up to the originating MSC (Figure 2-7). In addition to reducing the signalling load, this protocol helps to minimise unnecessary call set-up when the called party is unavailable. The performance of both protocols is found to be further improved if location caching and replication schemes are also implemented.

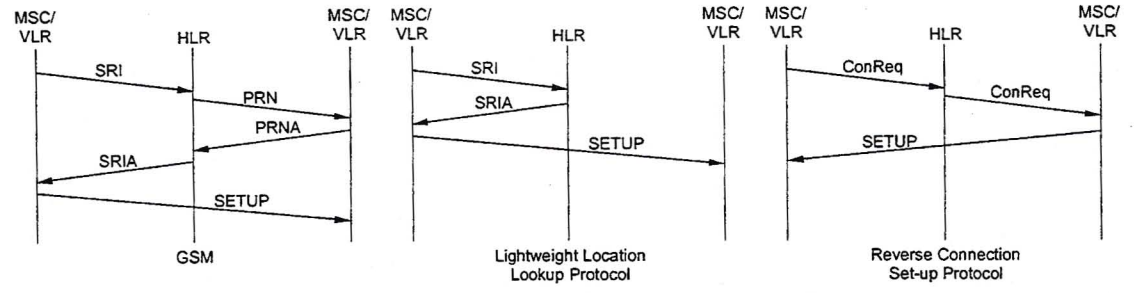


Figure 2-7. Location Lookup and Connection Set-up Message Flows for various protocols [Cui et al., 1998].

2.6.3 Memory-Based Mobility Management Methods

Memory-based mobility management techniques aim to reduce signalling load and database transaction load by trying to reduce the cost of repetitive actions and predictable behaviour. Typical examples include, trying to reduce the number of database lookups for a large number of calls to a particular subscriber or reducing the number of location updates from a subscriber who travels through the same VLRs from home to work on a daily basis.

2.6.3.1 Caching Strategies

Caching methods aim to reduce the signalling load and transaction load by storing subscriber profiles for call routing information in databases close to the source of the transactions. Jain et al. [1994] proposed a location caching strategy that maintains a record of previously called users in the MSC. This method was found to be effective if a large number of calls to a particular user originate from a MSC and the number of calls to this user is large compared to the user's mobility (referred to as a user's local call-to-mobility ratio). The cached information is used by the originating MSC to request a roaming number directly from the called party's VLR. If the called party has moved, a cache miss occurs and the HLR is queried for routing information.

The Gateway Location Register (GLR) [3GPP TR 23.909], proxy-HLR [Schopp, 2000] or mirror-HLR [Lingnau & Drobnik, 1998] is a node placed between HLRs and VLRs, that behaves like a HLR towards a VLR and like VLR towards a HLR. The GLR stores a subscriber's profile from the first location update and for subsequent location updates it behaves like the actual HLR by generating the necessary ISD and CL messages. While the subscriber roams on VLRs served by a GLR no signalling messages are sent to the subscriber's HLR for subsequent locations updates. When the subscriber leaves the area served by the GLR the subscriber's HLR cancels his/her profile from the GLR with the normal location cancellation procedure. Schopp [2000] showed via analysis that this method can be used to reduce inter-PLMN signalling traffic.

2.6.3.2 Long-term Profile Storage

Figure 2-8 shows the VLR utilisation for typical VLRs with city and residential coverage areas in a GSM network. During the morning when a subscriber commutes from home to work in the

city his/her MT performs a location update in each new VLR's coverage area and the subscriber's profile is deleted from the PVLR by the HLR. Hence, the utilisation of VLRs with residential coverage decreases during the morning and the utilisation of VLRs with coverage in the city increases. In the late afternoon/early evening as subscribers commute from work to home the reverse takes place. This type of behaviour is repetitive and is evident on every working day.

The basic objective of the Super-Charger concept [3GPP TR 23.912] is to reduce the signalling traffic load in mobile networks by using the memory resources in the VLR more efficiently. This is accomplished by not using the cancel location procedure to remove the subscriber's profile from the PVLR. If the subscriber returns to a PVLR where his/her profile is already present, a location update is performed but the insert subscriber data procedure is not performed. This has the effect of reducing the number of messages exchanged during the location update to only two, a UL and ULA message. If the VLR entered does not have the subscriber's profile the insert subscriber data procedure is used as in normal GSM. To ensure that Super-Charged networks are compatible with existing networks additional parameters are introduced into the existing messages to indicate whether a node supports Super-Charged functionality and to indicate the age of the subscriber profile stored in the VLR. If a UL message, for example, indicates the profile on the VLR is older than that currently stored on the HLR then the insert subscriber data procedure is initiated. In addition to reducing the signalling load, the Super-Charger mechanism also helps to reduce the transaction load on the VLR and HLR [3GPP TR 23.912].

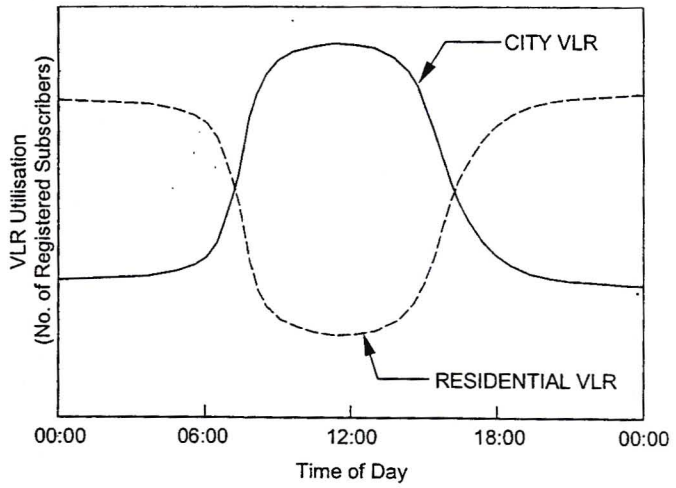


Figure 2-8. Graph showing the typical VLR Utilisation Characteristics (in terms of the number of subscribers registered on the VLR) for VLRs with city and residential coverage areas in a GSM network.

The Super-Charger mechanism can result in the VLR's database storage capacity being depleted if the number of stored profiles is allowed to grow uncontrollably and thus not allow new subscribers to register on the VLR. A database management mechanism is therefore required to constantly manage the database content. References [3GPP TR 23.912] and [3GPP TS 23.116] suggest the use of the following methods:

- A larger VLR database to ensure sufficient capacity is always available, but in combination with at least one of the methods below.
- A periodic audit that deletes old subscriber profiles, when no radio activity was detected from the subscriber over an extended period.
- A dynamic subscriber profile deletion mechanism that deletes a subscriber's from the VLR to make room for a newly arrived subscriber. Methods similar to those described in [Lin, 1998] and [Hung et al., 2001] can be used to delete old subscriber profiles.

2.6.3.3 Storage of User Mobility Patterns

The alternative strategy proposed by Tabbane [1995] attempts to reduce location updates triggered by repetitive mobility patterns by storing a subscriber's movement profile in the HLR and the terminal. The profile is defined such that various time intervals correspond to sets of location areas. A location area within a set is also associated with a value that indicates the probability with which a user is located in that area. When a call arrives for the subscriber, the network pages him sequentially by starting with the location area with the highest probability and if no response is received pages the location area with the next highest probability and so on. For delay sensitive calls, the subscriber can be paged in all the location areas defined simultaneously. If a subscriber enters a location area, other than those defined in the movement profile, the terminal initiates a voluntary location update to inform the network of its current position. Tabbane's results indicate this approach reduces the signalling load within the switching subsystem, base station subsystem and over the radio interface if the call-to-mobility ratio is low.

The two location algorithm [Lin, 1997] is a variation of the above method. Here the terminal and HLR maintain a record of current VLR's location area and the PVLR's location area (assuming each VLR is only responsible for one location area). If the subscriber returns to the previous location area no location update takes place. When a call arrives, the subscriber is first paged at the last recorded location area and if no response is obtained the subscriber is paged in the previous location area. The method is found to outperform IS-41 when the call-to-mobility ratio is low and is advantageous when there are frequent movements between adjacent location areas. In the worst case, when the subscriber does not return to the previous location area the location update traffic is equivalent to that of IS-41.

While the above approaches are effective at reducing the signalling load, they require additional memory capacity on the HLRs and lead to longer call set-up times. Furthermore, GSM's Short Message Service alerting procedures (ETSI GSM 09.02), which are used when a mobile terminal is temporarily unable, would be ineffective with the above methods.

2.7 Summary

In Section 2.2 the basic congestion and flow control techniques that were developed for packet switched networks are discussed. Congestion occurs when one or more network resources becomes scarce. The congestion and flow control mechanisms therefore attempt to alleviate congestion by discarding the excess packets and/or by throttling the input traffic until the congestion situation has subsided. Other techniques of flow control also involve the use of credit based schemes, to regulate the flow of traffic into the network.

Next, Section 2.3 discusses modelling methodologies for multilayered protocol architectures that have been applied to the analysis of SS7 protocols. These models were primarily designed to consider the effect of complex protocol interactions and processor sharing systems on the performance of communication protocols. However, some of these techniques are known to be very complex and detailed and are therefore difficult to solve analytically.

Section 2.4 examined and discussed previous performance studies of signalling networks. These studies show that the current congestion and flow control mechanisms perform poorly in some overload scenarios and are completely ineffective in an IN environment [Zepf & Rufa, 1994]. Various researchers have also observed that signalling networks are prone to experience REL message avalanches during congestion (Smith [1994] and Rumsewicz [1994]). As a result, various protocol enhancements have been proposed to address the shortcomings of the current implementations. A combined control scheme, for example, that incorporates both the IO and NOCP congestion controls, was found to perform better than the weaker of the two congestion

controls and at times it performed as well as the better of the two [Northcote & Rumsewicz, 1995].

Section 2.5 discussed previous research studies that had considered the impact of customer reattempts and application level recovery actions on network performance. The reattempt behaviour of customers and network layer protocols was found to significantly degrade network performance, especially during congestion situations.

Finally, Section 2.6 examined various studies that analysed the load attributed to mobility management and then discussed various proposals that aimed to alleviate the signalling load and database transaction load in GSM and IS-41 networks. The proposals include the use of a hierarchical database architecture, optimisation of the call delivery protocols, and various subscriber location and profile caching strategies. While some of these proposals are incompatible with existing mobility management procedures the Gateway Location Register and Super-Charger mechanism have been defined by 3GPP to help reduce signalling load.

3. Steady State Equilibrium Analysis of SS7 Network Performance in the PSTN

3.1 Introduction

Servicing packets in their order of arrival has the obvious advantages of fairness, simplicity and ease of implementation. The M/M/1/K[1] queue is one the simplest representations of a system with finite buffer storage resources and a first-come, first-serve (FCFS) queuing discipline. The FCFS queue can also be easily implemented in any system without an expensive processing overhead. Today's common channel signalling networks carry messages containing SCCP data, call set-up information and network management information. Since grade-of-service is of the utmost importance to a telecommunications company, the signalling network should maintain a high throughput of call related and management information, while discarding redundant data packets during congestion. To achieve this objective the *National Option with congestion priorities* allows for priority levels to be assigned to the individual signalling messages. During congestion the congested network elements should only discard low priority messages without disrupting the flow of high priority traffic, typically those messages from calls already in progress, and they should also not impede the distribution of network management messages containing the network status information.

The issue of signalling transfer point congestion was only addressed in the last decade. Message discard schemes were originally considered as being an acceptable flow control mechanism during processor overloads and are still widely used internationally. The ANSI Standards [ANSI T1.111] and research publications [Rumsewicz, 1993] have addressed this issue, and they propose the application of the flow control mechanisms currently defined for route set congestion at the network layer, in order to relay processor congestion information back to the traffic sources. However, it is possible that the additional load imposed by the feedback control messages will simply exacerbate the congestion situation, or the flow controls might not be able to throttle the source nodes fast enough. In order to develop congestion and flow control mechanisms that operate effectively in a large network environment and are also able to provide the optimum grade-of-service requirements, one has to initially evaluate the network's performance in the absence of these controls. This chapter focuses on developing a set of generalised equations to help determine the steady-state equilibrium performance of multiple STP networks in a PSTN during a worst-case scenario, when no flow control mechanisms are available or when they are completely ineffective.

In the analysis, below, each STP is assumed to be based on a central processor architecture (Section 1.9.5) and messages arriving to MTP level 3 are stored in a FCFS queue, before being serviced. Section 3.2 develops a prioritised queuing model with multiple discard thresholds, to control the admission of signalling messages into the level 3 processing buffer. The prioritised queuing model solves for the equilibrium state probabilities and the various performance measures that characterise the operation of the system. These analytical solutions are later used for determining the performance of the STPs and the probability with which signalling messages are successfully admitted into the level 3 queue.

The level 2 transmission queue is usually modelled as a M/G/1 queue. However, if the arrival rate of messages at the queue is small (relative to the service rate), the links are assumed to be free of transmission errors and the effect of fill-in signal units is negligible then the link can be

[1] A *M/M/1/K* queue has a Markov (*M*) arrival process, a single server with a Markov service process and a queue with a finite storage capacity, *K*.

approximated by a simple M/M/1 queue (as in Lazar et al. [1992]) from the equations provided in ITU-T Recommendation Q.706. A queuing network, of queues with negative exponential service times, can then be used to represent the level 2 and level 3 entities of a SS7 network. Section 3.3 develops a generalised equilibrium analysis for signalling networks with multiple STPs, and considers message sequencing, the effect of caller reattempts, application level recovery actions, message priorities and routing. The results obtained from this mathematical analysis are presented for various overload scenarios and validated with the aid of computer simulations in Section 3.4.

3.2 The Prioritised Queuing Discipline

Literature on queuing theory commonly focuses on prioritised service disciplines, where different classes of traffic are subjected to different service routines, e.g. preemptive queuing, non-preemptive queuing and BCMP class dependent service disciplines. A few studies have also considered class dependent queuing (or buffer management) disciplines. Lam & Luke Lien [1981], for example, allowed transit packets to consume more buffer resources than the new packets in their input buffer limit scheme; while Irland [1978] analysed different buffer management schemes in a switch with a shared pool of buffer resources (Section 2.2.2.2 provides a more detailed description of these buffer management schemes). Rumsewicz [1993] used a prioritised queuing discipline (a type of class dependent queuing discipline) to investigate the effects of message discarding, in SS7, on the call completion rate. However, his queuing model only considers two message priorities and is limited to a priority scheme where IAM messages are assigned the lowest priority and the other messages are assigned the higher priority. The analysis below develops a prioritised queuing model to accommodate for three levels of message priorities, as required in the NOCP, and also solves for the queue's performance measures.

Consider a queue with a negative exponential service distribution, where customers arrive according to a Poisson process. Assume that the arrivals can be categorised into three independent Poisson streams consisting of messages with low, medium and high priorities and with mean arrival rates of λ_0, λ_1 and λ_2 respectively. The queue is partitioned into three regions (Figure 3-1) such that:

- all the messages are accepted if the queue length is below K_1,
- the lowest priority messages (λ_0) will be discarded if the queue length is greater than or equal to K_1 but below K_2,
- only the highest priority messages (λ_2) will be accepted if the queue length is greater than or equal to K_2 but below K_3,
- none of the messages are accepted if the queue is filled to its maximum capacity, K_3.

This multiple discard-threshold queue is referred to as a M/M/1/K_1, K_2, K_3 queue in later text. The discard scheme expressed here is analogous to the message discard procedure defined for NOCP flow control. The message arrival rate, γ_n, into the queue can therefore be written as

$$\gamma_n = \begin{cases} \lambda & 0 \le n < K_1 \\ \lambda' & K_1 \le n < K_2 \\ \lambda'' & K_2 \le n < K_3 \\ 0 & n \ge K_3 \end{cases} \qquad \text{(3-1)}$$

where

$$\lambda = \lambda_0 + \lambda_1 + \lambda_2$$
$$\lambda' = \lambda_1 + \lambda_2$$
$$\lambda'' = \lambda_2$$

and n is the number of messages already in the queue or the state of the system. The combined message streams λ, λ' and λ'' are Poisson arrivals since their constituent streams are independent Poisson arrivals.

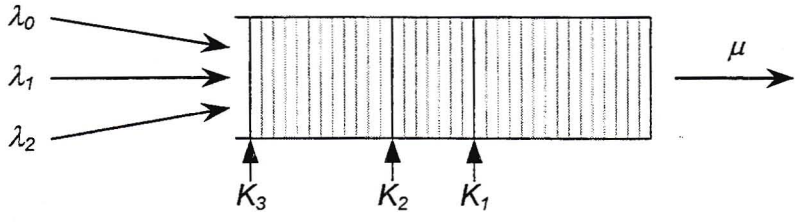


Figure 3-1. Priority thresholds of the M/M/1/K_1, K_2, K_3 queue.

The corresponding state transition diagram for the M/M/1/K_1, K_2, K_3 queue is shown in Figure 3-2. Calculation of the steady state equilibrium distribution yields the following:

$$\pi_n = \begin{cases} \pi_0 \rho^n & 0 \le n \le K_1 \\ \pi_0 \rho^{K_1} \rho'^{\,n-K_1} & K_1 < n \le K_2 \\ \pi_0 \rho^{K_1} \rho'^{\,K_2-K_1} \rho''^{\,n-K_2} & K_2 < n \le K_3 \end{cases} \qquad \textbf{(3-2)}$$

where

$$\rho = \frac{\lambda}{\mu}, \quad \rho' = \frac{\lambda'}{\mu} \quad \text{and} \quad \rho'' = \frac{\lambda''}{\mu}. \qquad \rho \ge \rho' \ge \rho''$$

One could also arrive at the above solution if a 3-dimensional state transition diagram is used where each axis represents the arrival rate of the separate message streams.

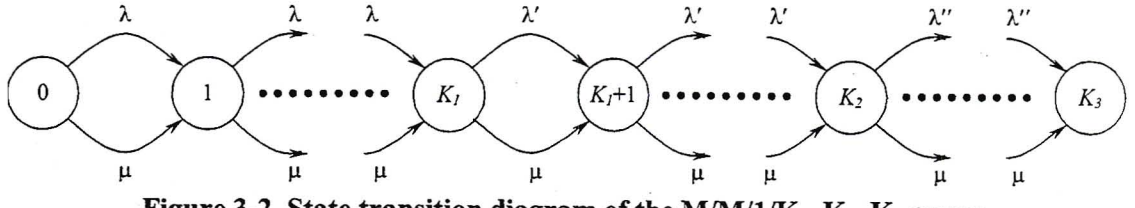


Figure 3-2. State transition diagram of the M/M/1/K_1, K_2, K_3 queue.

The normalising constant, π_0, can be calculated by equating the sum of the state probabilities to one,

$$\sum_{n=o}^{K_3} \pi_n = 1.$$

The following is obtained:

$$\pi_0 = \begin{cases} \left[\dfrac{1-\rho^{K_1+1}}{1-\rho} + \rho^{K_1}\rho'^{-K_1}\left(\dfrac{\rho'^{K_1+1}-\rho'^{K_2+1}}{1-\rho'}\right) + \rho^{K_1}\rho'^{K_2-K_1}\rho''^{-K_2}\left(\dfrac{\rho''^{K_2+1}-\rho''^{K_3+1}}{1-\rho''}\right)\right]^{-1} & \rho \neq 1, \rho' \neq 1, \rho'' \neq 1 \\ \left[\dfrac{1-\rho^{K_2+1}}{1-\rho} + \rho^{K_2}\rho''^{-K_2}\left(\dfrac{\rho''^{K_2+1}-\rho''^{K_3+1}}{1-\rho''}\right)\right]^{-1} & \rho = \rho' \neq 1, \rho'' \neq 1 \\ \left[\dfrac{1-\rho^{K_1+1}}{1-\rho} + \rho^{K_1}\rho'^{-K_1}\left(\dfrac{\rho'^{K_1+1}-\rho'^{K_3+1}}{1-\rho'}\right)\right]^{-1} & \rho \neq 1, \rho' = \rho'' \neq 1 \\ \left[\dfrac{1-\rho^{K_3+1}}{1-\rho}\right]^{-1} & \rho = \rho' = \rho'' \neq 1 \\ \left[K_1 + 1 + \rho^{K_1}\rho'^{-K_1}\left(\dfrac{\rho'^{K_1+1}-\rho'^{K_2+1}}{1-\rho'}\right) + \rho^{K_1}\rho'^{K_2-K_1}\rho''^{-K_2}\left(\dfrac{\rho''^{K_2+1}-\rho''^{K_3+1}}{1-\rho''}\right)\right]^{-1} & \rho = 1, \rho' \neq 1, \rho'' \neq 1 \\ \left[\dfrac{1-\rho^{K_1+1}}{1-\rho} + \rho^{K_1}\rho'^{-K_1}(K_2-K_1) + \rho^{K_1}\rho'^{K_2-K_1}\rho''^{-K_2}\left(\dfrac{\rho''^{K_2+1}-\rho''^{K_3+1}}{1-\rho''}\right)\right]^{-1} & \rho \neq 1, \rho' = 1, \rho'' \neq 1 \\ \left[\dfrac{1-\rho^{K_1+1}}{1-\rho} + \rho^{K_1}\rho'^{-K_1}\left(\dfrac{\rho'^{K_1+1}-\rho'^{K_2+1}}{1-\rho'}\right) + \rho^{K_1}\rho'^{K_2-K_1}\rho''^{-K_2}(K_3-K_2)\right]^{-1} & \rho \neq 1, \rho' \neq 1, \rho'' = 1 \\ \left[K_2 + 1 + \rho^{K_2}\rho''^{-K_2}\left(\dfrac{\rho''^{K_2+1}-\rho''^{K_3+1}}{1-\rho''}\right)\right]^{-1} & \rho = \rho' = 1, \rho'' \neq 1 \\ \left[\dfrac{1-\rho^{K_1+1}}{1-\rho} + \rho^{K_1}\rho'^{-K_1}(K_3-K_1)\right]^{-1} & \rho \neq 1, \rho' = \rho'' = 1 \\ [K_3+1]^{-1} & \rho = \rho' = \rho'' = 1 \end{cases}$$

The probability that messages in traffic streams λ_0, λ_1 or λ_2 will be discarded is given by

$$p_0 = \sum_{n=K_1}^{K_3}\pi_n, \quad p_1 = \sum_{n=K_2}^{K_3}\pi_n \quad \text{and} \quad p_2 = \pi_{K_3} \tag{3-3}$$

respectively.

The manner in which the different messages accumulate within a queue can be used to determine the effectiveness of a three-priority scheme, and to inspect buffer utilisation during congestion. In addition, these measurements can also be used to select the optimum buffer threshold settings, necessary to obtain maximum utilisation of the available resources. The expected number of messages, L, in the system at steady state is

$$L = \sum_{n=0}^{K_3} n\pi_n = \pi_0 \sum_{n=0}^{K_1} n\rho^n + \pi_0\rho^{K_1}\rho'^{-K_1}\sum_{n=K_1+1}^{K_2} n\rho'^n + \pi_0\rho^{K_1}\rho'^{K_2-K_1}\rho''^{-K_2}\sum_{n=K_2+1}^{K_3} n\rho''^n \tag{3-4}$$

where

$$\sum_{n=0}^{K_1} n\rho^n = \rho \cdot \frac{1-(K_1+1)\rho^{K_1} + K_1\rho^{K_1+1}}{(1-\rho)^2}$$

$$\sum_{n=K_1+1}^{K_2} n\rho'^n = \rho' \cdot \frac{(K_1+1-K_1\rho')\rho'^{K_2} - (K_2+1-K_2\rho')\rho'^{K_2}}{(1-\rho')^2}$$

$$\sum_{n=K_2+1}^{K_3} n\rho''^n = \rho'' \cdot \frac{(K_2+1-K_2\rho'')\rho''^{K_2} - (K_3+1-K_3\rho'')\rho''^{K_3}}{(1-\rho'')^2}$$

and the mean queue length is given by

$$L_Q = L - (1 - \pi_0). \tag{3-5}$$

The expected number of type j messages, L_j, is of particular interest in a multiple threshold queue. L_j indicates which messages consume the most resources during an overload. This performance parameter is given by the following expression:

$$L_j = L \frac{\lambda_j (1 - p_j)}{\sum_{n=1}^{N} \lambda_n (1 - p_n)} \quad \text{where N = no. of message streams.} \tag{3-6}$$

Calculation of the waiting time distribution helps to estimate the maximum queuing delay that can be expected during an overload. The queuing delay also helps in the selection of buffer sizes, as a large buffer can lead to excessive delays during an overload and consequently lead to timeouts of application level processes. The expected waiting time of a message in a $M/M/1/K_1$, K_2, K_3 queue can be obtained from Little's formula, and is therefore given by

$$W = \frac{L}{\lambda_s} \tag{3-7}$$

where

$$\lambda_s = (\lambda - \lambda')(1 - p_0) + (\lambda' - \lambda'')(1 - p_1) + \lambda''(1 - p_2)$$

or

$$\lambda_s = \lambda_0 (1 - p_0) + \lambda_1 (1 - p_1) + \lambda_2 (1 - p_2) \tag{3-8}$$

if there are only three distinct messages classes with distinct priorities. In equation (3-7) λ_s is the throughput of a queue during equilibrium.

The mean waiting time is given by

$$W_Q = \frac{L_Q}{\lambda_s} \tag{3-9}$$

and the delay experienced by type j messages is

$$W_j = \frac{L_j}{\lambda_j}. \tag{3-10}$$

Lastly, the processor utilisation, U, is given by

$$U = (1 - \pi_0). \tag{3-11}$$

3.3 Equilibrium Analysis

Most analytical and simulation models tend to analyse SS7 network performance by examining a network with a single STP node (e.g. the SS7 performance investigations by Rumsewicz [1993] and Smith [1994]). In models where two STPs are used, the analysis usually concentrates on the performance of the link set between the STPs (e.g. the analysis on the impact of sustained overloads by Manfield et al. [1994]). However, while these analytical models do highlight the shortcomings of the current SS7 congestion and flow controls, they cannot be used to evaluate

the performance of large signalling networks or different network architectures. The analysis, below, (an extension of Rumsewicz's [1993] analysis on the effects message discard schemes on call completion rates) develops a generalised analytical model to evaluate the performance and call handling capability of SS7 networks during STP congestion. The analytical model presented here has the following advantages over previous models:

- one or more STPs may be included in the network,
- the model is not restricted to a single network architecture,
- different message priority schemes can be analysed,
- different routing algorithms can be evaluated, and
- different traffic loads can be used for each source-destination pair.

The simple ISUP call set-up message sequence shown in Figure 1.9 is used to determine the traffic load of each message type. Furthermore, the analytical model also accounts for the effects of application level recovery actions and customer reattempts on network performance.

To model the traffic processes in a simple packet switched network at least two principal parameters are required, namely: the *routing matrix* and the *traffic matrix*. Other parameters, such as the number of routes available between two nodes and the number of nodes traversed over a particular route, can be derived directly from these parameters. These additional parameters are used in the analytical model to simplify the realisation of the system equations. The following is a list of notation and definitions used in the subsequent analysis:

i) N STPs and M SP regions are defined within the network. Each STP is assigned a value n ($1 \le n \le N$) and each SP region is assigned a value m ($1 \le m \le M$) as shown in Figure 3-4. M_{Rx} is the number of SPs in region x.

ii) α is a $M \times M$ source/destination traffic matrix of new call traffic (i.e. the first offered load) and α_{ij} is the number of new call arrivals per second at region i that are designated to region j. The first offered load from region i is given by

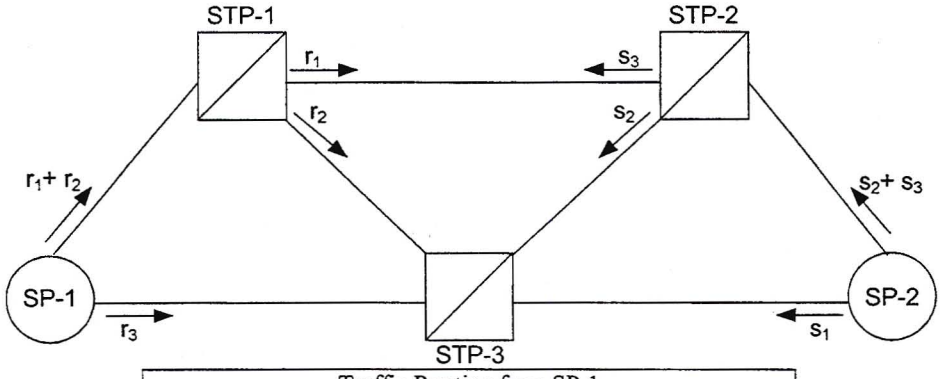
$$\alpha_i = \sum_{j=1}^{M} \alpha_{ij} .$$

iii) R is a $M \times M$ matrix where R_{ij} is the number of routes available between region i and region j. Since all the links are bi-directional $R_{ij} = R_{ji}$. Let A = maximum(R_{ij}).

iv) Q_F is a $M \times M \times A$ matrix where $Q_{F\,ijk}$ is the number of STPs traversed by a message in the forward direction from region i to region j on route k. Let B_F = maximum($Q_{F\,ijk}$). Note that if the number of routes between two nodes is less than A, then $Q_{F\,ijk}$ for the undefined routes is zero.

v) Q_R is a $M \times M \times A$ matrix where $Q_{R\,ijk}$ is the number of STPs traversed by a message in the reverse direction on route k from region i to region j. If the same nodes are traversed in both directions then $Q_{F\,ijk} = Q_{R\,ijk}$. Let B_R = maximum($Q_{R\,ijk}$).

vi) $\Gamma(i,j,k)$ is a one-to-one mapping function that maps the reverse route (for route k from i to j) to the equivalent forward route at the destination node. For example, consider the network configuration, shown in Figure 3-3, and its associated routing tables. The outgoing routes from each signalling point are labelled from left to right (r_1 to r_3 are the forward routes from SP-1). Here $Q_{F\,123} = 1$ and $Q_{R\,123} = 2$, for calls originating from SP-1. But, in order to calculate the outgoing message load from SP-2, a mapping function is required to translate the route followed by traffic in the reverse direction from r_3 to the equivalent route in the forward direction from SP-2 (s_2 in this case). Therefore $\Gamma(1,2,3) = 2$, indicates that s_2 is used to return responses for messages arriving on r_3. If a simple mapping function is not attainable a lookup table (or $M \times M \times A$ matrix) can be used to accomplish this task.



Traffic Routing from SP-1	
Forward	Reverse
SP-1→STP-1→STP-2→SP-2	SP-2→STP-2→STP-1→SP-1
SP-1→STP-1→STP-3→SP-2	SP-2→STP-3→SP-1
SP-1→STP-3→SP-2	SP-2→STP-2→STP-3→SP-1

Traffic Routing from SP-2	
Forward	Reverse
SP-2→STP-3→SP-1	SP-1→STP-1→STP-3→SP-2
SP-2→STP-2→STP-3→SP-1	SP-1→STP-3→SP-2
SP-2→STP-2→STP-1→SP-1	SP-1→STP-1→STP-2→SP-2

Figure 3-3. Example of routing in a simple network.

vii) $\mathcal{R}_F$ is a $M \times M \times A \times B_F$ forward routing matrix where $\mathcal{R}_{F\,ijkl}$ indicates which STP is encountered by a message from region i to region j on route k at node l. In the above example, $\mathcal{R}_{F\,1212} = 2$ indicates that STP-2 is the second STP traversed in the forward direction on route r_1 from SP-1 to SP-2.

viii) $\mathcal{R}_R$ is a $M \times M \times A \times B_R$ reverse routing matrix where $\mathcal{R}_{R\,ijkl}$ indicates which STP is encountered by a message in the reverse direction from region i to region j on route k at node l. In the above example, $\mathcal{R}_{R\,1212} = 1$ indicates that STP-1 is the second STP traversed in the reverse direction on the route r_1 from SP-1 to SP-2.

ix) $\psi(i,j,k,l,n)$ is a Boolean function that returns either 1 (true) or 0 (false) if the lth node on route k from region i to region j is STP-n ($1 \le n \le N$).

x) The transmission rate of a traffic stream with type T messages is denoted by λ_T, where T denotes the message type, e.g. IAM, ACM, etc. $\lambda_{T\,ijk}$ refers to the transmission rate of type T messages from region i to region j on route k (in the forward direction).

$$\lambda_{T\,ij} = \sum_{k=1}^{R_{ij}} \lambda_{T\,ijk}; \quad \lambda_{T\,i} = \sum_{j=1}^{M} \sum_{k=1}^{R_{ij}} \lambda_{T\,ijk} \quad \text{and} \quad \lambda_T = \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=1}^{R_{ij}} \lambda_{T\,ijk} \qquad \textbf{(3-12)}$$

xi) ϕ_T is the priority of message T, where T can be IAM, ACM, etc. ($0 \le \phi_T \le 2$).

xii) σ_T is the length of message T in octets ($\sigma_T > 0$).

xiii) $P_D(\phi_T,d,i,j,k,l)$ is the probability that a message T from region i to region j on route k with priority ϕ_T is *discarded* at the lth node. $P_S(\phi_T,d,i,j,k,l)$ is the probability that a message T from region i to region j on route k with priority ϕ_T is *successfully* admitted into the level 3 queue at the lth node. $d = 0$ or 1 indicates whether the other variables refer to a route in the forward or reverse direction, respectively.

xiv) c and r are the reattempt probabilities for failed calls and failed release procedures, respectively.

xv) The signalling transfer points have a single MTP level 3 processor with a service rate of μ messages/sec.

xvi) λ_n is the arrival rate of messages at STP-*n*, λ_n' is the arrival rate of messages with priorities of one and two at STP-*n* and λ_n'' is the arrival rate of messages with a priority of two at STP-*n*.

xvii) The signalling points have two MTP level 3 processors, one for incoming messages and the other for outgoing messages, each with a service rate of μ_{SP} messages/sec.

xviii) All the links have a transmission speed of μ_L bits/sec.

xix) ν_{Tij} is the arrival rate of type *T* messages at the links from region *i* to STP-*j* in bits/sec, ν'_{Tij} is the arrival rate of type *T* messages at the links from STP-*i* to STP-*j* in bits/sec and ν''_{Tij} is the arrival rate of type *T* messages at the links from STP-*i* to region *j* in bits/sec.

The total number of IAM messages generated is equal to the total number of new call arrivals plus reattempts due to failures. When an IAM message is created a timer, T_{IAM}, is started for the associated call. If an ANM is not received before T_{IAM} expires, then the call is assumed to have failed and the circuit release procedure is initiated. A call may fail if either the IAM message or its resulting ANM message is discarded at a congested STP. The unsuccessful calls are later reattempted with some probability *c*. The probability of an IAM message being discarded is equal to the sum of the discard probabilities at each node along its route. While the probability of an ANM message being discarded is equal to the product of the probability that the IAM message was successful and the sum of discard probabilities at each node on the return route. The number of IAM messages sent from one region, *x*, to another, *y*, can therefore be written as follows:

$$\lambda_{IAM\,xy} = \text{new call arrivals} + \text{first IAM reattempts} + \text{second IAM reattempts} + \ldots \quad \textbf{(3-13)}$$

where

$$\text{new call arrivals} = \alpha_{xy},$$

$$\text{first IAM reattempts} = \alpha_{xy}\left[\sum_{i=1}^{R_{xy}}\left(\sum_{j=1}^{Q_{F\,xyi}}\frac{1}{R_{xy}}P_D(\phi_{IAM},0,x,y,i;j)\right)\right]\cdot c + \alpha_{xy}\left[\sum_{i=1}^{R_{xy}}\left(\sum_{j=1}^{Q_{R\,xyi}}\frac{1}{R_{xy}}P_S(\phi_{IAM},0,x,y,i,Q_{F\,xyi})P_D(\phi_{ANM},1,x,y,i,j)\right)\right]\cdot c$$

and

$$\text{second IAM reattempts} = (\text{first IAM reattempts})\cdot\left[\sum_{i=1}^{R_{xy}}\left(\sum_{j=1}^{Q_{F\,xyi}}\frac{1}{R_{xy}}P_D(\phi_{IAM},0,x,y,i,j)\right)\right]\cdot c + (\text{first IAM reattempts})\cdot\left[\sum_{i=1}^{R_{xy}}\left(\sum_{j=1}^{Q_{R\,xyi}}\frac{1}{R_{xy}}P_S(\phi_{IAM},0,x,y,i,Q_{F\,xyi})P_D(\phi_{ANM},1,x,y,i,j)\right)\right]\cdot c$$

The geometric series in (3-13) can thus be rewritten as

$$\lambda_{IAM\,xy} = \frac{\alpha_{xy}}{\left[1-\left(\sum_{i=1}^{R_{xy}}\left(\sum_{j=1}^{Q_{F\,xyi}}\frac{1}{R_{xy}}P_D(\phi_{IAM},0,x,y,i,j)+\sum_{j=1}^{Q_{R\,xyi}}\frac{1}{R_{yx}}P_S(\phi_{IAM},0,x,y,i,Q_{F\,xyi})P_D(\phi_{ANM},1,x,y,i,j)\right)\right)\cdot c\right]} \quad \textbf{(3-14)}$$

The above equation assumes that load sharing is used in the network, to allow the IAM traffic to be distributed equally across all possible routes to the destination. The IAM traffic along each route, z, from region x to region y is therefore given by

$$\lambda_{IAM\,xyz} = (1/R_{xy})\lambda_{IAM\,xy}, \qquad \text{where } 1 \le z \le R_{xy}.$$

ACM and ANM messages are generated for each IAM message that is successfully received at the destination. Let $z = \Gamma(y,x,i)$, then

$$\lambda_{ANM\,xyz} = \lambda_{ACM\,xyz} = \frac{1}{R_{yx}} \lambda_{IAM\,yx} P_S\left(\phi_{IAM}, 0, y, x, i, Q_{F\,yxi}\right). \tag{3-15}$$

For each IAM message created at least one subsequent REL message is created. When the first REL is sent two timers, T1 and T5 (T1 < T5), are started. If a RLC is not received before T1 expires, a new REL is sent and T1 is restarted. This process continues until T5 expires. The REL messages may therefore be repeated for up to 'T1/T5' times. A REL reattempt occurs if either a REL message or its associated RLC message is discarded within the network. For modelling purposes r, the effective reattempt probability for repeating a REL message when the release procedure fails, is used with the equations. The effective reattempt probability is given by the following equation: 'r = (T5/T1)/(T5/T1+1)' [Rumsewicz, 1993], where the numerator gives the maximum number of REL messages that may be generated due to reattempts and the denominator gives the total number of REL messages that may be generated. By using a method that is similar to the one used to calculate the outgoing IAM message load, the number of REL messages transmitted over each route is given by

$$\lambda_{RELxyz} = \frac{\lambda_{IAM\,xyz}}{\left[1 - \left(\sum_{i=1}^{Q_{F\,xyz}} P_D\left(\phi_{REL}, 0, x, y, z, i\right) + \sum_{i=1}^{Q_{R\,xyz}} P_S\left(\phi_{REL}, 0, x, y, z, Q_{F\,xyz}\right) P_D\left(\phi_{RLC}, 1, x, y, z, i\right)\right) \cdot r\right]}. \tag{3-16}$$

Note, that while the IAM message load (including reattempts) is equally distributed over all possible routes the subsequent REL messages (including reattempts) will follow the same route as their associated IAM messages. Hence, the REL load across each route is calculated separately. It is therefore possible for the REL and RLC loads to be unequal across each route when congestion, for example, only occurs on one route.

The rate at which RLC messages are transmitted over each route is given by

$$\lambda_{RLC\,xyz} = \lambda_{REL\,yxi} P_S\left(\phi_{REL}, 0, y, x, i, Q_{F\,yxi}\right) \tag{3-17}$$

where $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

The call completion rate, $S_{C\,x}$, for a region x is obtained by determining the total number of ANM messages successfully received:

$$S_{Cx} = \sum_{i=1}^{M} \sum_{j=1}^{R_{ix}} \lambda_{ANM\,ixj} P_S\left(\phi_{ANM}, 0, i, x, j, Q_{F\,ixj}\right) \tag{3-18}$$

The message throughput, S_M, for each route is given by

$$S_{M\,xyz} = \sum_T \lambda_{T\,xyz} P_S\left(\phi_T, 0, x, y, z, Q_{F\,xyz}\right) \tag{3-19}$$

where T identifies to the type of message (either IAM, ACM, ANM, etc.).

The message throughput of traffic to destination y is given by

$$S_{M\,xy} = \sum_{i=1}^{R_{xy}} S_{M\,xyi}$$

and the total message throughput of traffic from region x is

$$S_{M\,x} = \sum_{i=1}^{M} \sum_{j=1}^{R_{xi}} S_{M\,xij} \cdot$$

The total number of messages successfully arriving at region y is given by

$$S'_{M\,y} = \sum_{i=1}^{M} \sum_{j=1}^{R_{iy}} S_{M\,iyj} \cdot \tag{3-20}$$

The total number of messages, λ_n, arriving at STP-n is equal to the total number of new messages arriving into the network plus the total number of transit messages from the other STPs:

$$\lambda_n = \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=1}^{R_{ij}} \Bigg(\underbrace{\sum_T \lambda_{Tijk} \cdot \psi(i,j,k,1,n)}_{\text{New messages}} + \underbrace{\sum_{l=2}^{Q_{Fijk}} \left(\sum_T \lambda_{Tijk} P_S\left(\phi_T, 0, i, j, k, l-1\right) \right) \cdot \psi(i,j,k,l,n)}_{\text{Transit messages}} \Bigg) \tag{3-21}$$

A similar summation process can be used to obtain λ_n' (the arrival rate of messages with a priority of one or two) and λ_n'' (the arrival rate of messages with a priority of two) for each STP. Once λ_n, λ_n', and λ_n'' are available, Equation (3-3) can then be used to calculate the blocking probabilities at STP-n. Even though the message streams are correlated with each other, they are assumed be independent Poisson traffic streams in order to simplify the analysis. In addition to the blocking probabilities, the performance measures for each STP's processing queue can be evaluated by using equations (3-4) through to (3-11).

Once the probabilities of discard at each STP are available, P_S and P_D can be calculated from

$$P_S\left(\phi_m, d, x, y, z, i\right) = \begin{cases} \prod_{n=1}^{i} \left(1 - p_{ae}\right) & \text{for} \quad d = 0 \\ \prod_{n=1}^{i} \left(1 - p_{af}\right) & \text{for} \quad d = 1 \end{cases} \tag{3-22}$$

and

$$P_D(\phi_m,d,x,y,z,i)=\begin{cases} p_{ag} & \text{for } i=1 \text{ and } d=0 \\ p_{ag}\left(\prod_{n=1}^{i-1}(1-p_{ae})\right) & \text{for } i>1 \text{ and } d=0 \\ p_{ah} & \text{for } i=1 \text{ and } d=1 \\ p_{ah}\left(\prod_{n=1}^{i-1}(1-p_{af})\right) & \text{for } i>1 \text{ and } d=1 \end{cases} \quad \textbf{(3-23)}$$

where

$$a=\phi_m,\quad e=\Re_{F\,xyzn},\quad f=\Re_{R\,xyzn},\quad g=\Re_{F\,xyzi} \text{ and } h=\Re_{R\,xyzi}.$$

In the above equations, p_{xy} is the blocking probability at threshold 'x + 1' of STP-y. The value for p_{xy} can be calculated from equation (3-3) where p_x is the blocking probability at threshold 'x + 1' of a M/M/1/K_1, K_2, K_3 queue. The M/M/1/K_1, K_2, K_3 queuing model is used to approximate the behaviour of the STP's level 3 processing queue. It is important to note that the message arrival process at the STP is not Poisson due to call reattempts and REL reattempts. In Section 3.4 simulation models are used to examine and compare the accuracy of these approximations. Equation (3-22) calculates the probability that a message is successfully admitted into a STP and was also successfully admitted at each of the preceding STPs, while equation (3-23) calculates the probability that a message is discarded at a STP but was successfully admitted at each of the preceding STPs (when i > 1).

A message traversing a network will encounter queuing and processing delays in the level 3 processing buffers and the level 2 transmission buffers. Hence, a delay analysis is also necessary to gauge the grade-of-service of a signalling network and to ensure that application level recovery timers do not expire due to excessive network latencies.

The mean delay endured by messages, from region x, in the level 3 processing buffers is given by:

$$W_{MTPx}=\left(\sum_{i=1}^{M}\sum_{j=1}^{R_{xi}}\sum_{k=1}^{Q_{Fxij}}\frac{L_n}{A_n}\cdot\frac{S_{M\,xij}}{S_{M\,x}}\right)+\frac{1}{\mu_{SP}-\sum_{T}\frac{\lambda_{T\,x}}{M_{Rx}}}+\left(\sum_{i=1}^{M}\frac{1}{\mu_{SP}-\frac{S'_{M\,i}}{M_{Ri}}}\cdot\frac{S_{M\,xi}}{S_{M\,x}}\right) \quad \textbf{(3-24)}$$

where

$$A_n=\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{R_{ij}}\left(\sum_{l=1}^{Q_{Fijk}}\left(\sum_{T}\lambda_{T\,ijk}P_S(\phi_T,0,i,j,k,l)\right)\cdot\psi(i,j,k,l,n)\right) \quad \textbf{(3-25)}$$

and

$$n=\Re_{F\,xijk}.$$

In the above equations, A_n is the rate at which messages are admitted into the processing buffer of STP-n (i.e. the successful messages), and is calculated by determining the total number of messages from each region that are routed via STP-n. The expected number of messages, L_n, in STP-n can be obtained from equation (3-4). Each term in equation (3-24) is derived with Little's theorem. The first term in equation (3-24) determines the mean delay experienced by messages in the STP processing buffers, the second term determines the mean delay experienced in the outgoing processing buffer of the originating SP and the third term determines the mean delay experienced in the incoming processing buffer of the destination SPs.

To calculate the link level transmission delays, the traffic load carried by each signalling link is required, and is given by

$$v_{T\,xm} = \sum_{i=1}^{M}\sum_{j=1}^{R_{xi}}\left(\lambda_{T\,xij}\cdot\psi(x,i,j,1,m)\right) \tag{3-26}$$

for SP to STP incoming links,

$$v'_{T\,mn} = \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{R_{ij}}\left(\sum_{l=1}^{Q_{Fijk}-1}\lambda_{T\,ijk}P_S(\phi_T,0,i,j,k,l)\cdot\psi(i,j,k,l,m)\psi(i,j,k,l+1,n)\right) \tag{3-27}$$

for STP to STP transit links, and

$$v''_{T\,nx} = \sum_{i=1}^{M}\sum_{j=1}^{R_{ix}}\left(\lambda_{T\,ixj}P_S(\phi_T,0,i,x,j,Q_{F\,ixj})\cdot\psi(i,x,j,Q_{F\,ixj},n)\right) \tag{3-28}$$

for STP to SP outgoing links.

The MTP level 2 signalling links are modelled as M/M/1 queues. Within the analysis they are assumed to have a sufficiently large capacity in order to remain uncongested under high loads as the analysis focuses on the impact of congestion and discarding in the MTP level 3 buffers of STPs. By applying Little's theorem, the mean delay endured by messages, from region x, across the signalling links is given by

$$W_{LINK\,x} = \sum_{i=1}^{M}\sum_{j=1}^{R_{xi}}\sum_{T}\left(\left(\frac{8\sigma_T S_{M\,xij}}{\left(\mu_L - \frac{\sum_T 8\sigma_T v_{T\,xa}}{M_{Rx}}\right)S_{M\,x}}\cdot\frac{v_{T\,xa}}{\sum_T v_{T\,xa}}\right)+\omega+\left(\frac{8\sigma_T S_{M\,xij}}{\left(\mu_L - \frac{\sum_T 8\sigma_T v''_{T\,di}}{M_{Ri}}\right)\lambda_{S\,x}}\cdot\frac{v''_{T\,di}}{\sum_T v''_{T\,di}}\right)\right) \tag{3-29}$$

where

$$\omega = \begin{cases}\sum_{k=1}^{Q_{F\,xij}-1}\left(\frac{8\sigma_T S_{M\,xij}}{\left(\mu_L - \sum_T 8\sigma_T v'_{T\,bc}\right)S_{M\,x}}\cdot\frac{v'_{T\,bc}}{\sum_T v'_{T\,bc}}\right) & Q_{F\,xij} > 1\\ 0 & Q_{F\,xij} \le 1\end{cases}$$

and $a = \Re_{F\,xij1}, b = \Re_{F\,xijk}, c = \Re_{F\,xije}, d = \Re_{F\,xijf}, e = k+1, f = Q_{F\,xij}$,

$\sum_T v_{T\,xa} \neq 0, \sum_T v'_{T\,bc} \neq 0, \sum_T v''_{T\,di} \neq 0.$

Finally, the mean end-to-end message transfer delay, W_x, for an arbitrary message from region x is given by the following expression

$$W_x = W_{MTP\,x} + W_{LINK\,x} \tag{3-30}$$

In the above analysis, the transmission rate of each message stream depends on its probability of discard in the STP processing buffers. However, the discard probabilities are calculated from the message arrival rates at each STP. An iterative algorithm, supplied in Appendix A.1, was therefore used to solve the system of equations. The algorithm initially guesses values for the probabilities of discard in each STP, and then uses these values to calculate the transmission rate

of the various traffic streams. The transmission rates are then used to calculate the probabilities of discard at each STP, which are again used to recalculate the message transmission rates in an iterative manner until the discard probabilities converge to a particular value.

3.4 Numerical Results and Discussion

This section presents the numerical results obtained for the steady state equilibrium analysis of a fully connected multiple STP network. The network examined here (illustrated in Figure 3-4) is structurally similar to the Singapore National network, discussed in Section 1.4. Each region consists of 6 SPs and each SP is connected to its adjacent STPs. All links have a transmission rate of 128 kb/s. Messages within the network are routed using the shortest-path routing algorithm and where multiple routes exist, the load is balanced equally across each route. Messages sent in the reverse direction travel along the same path as the messages that initiated them. The subsequent messages of a call follow the same route as the IAM message, in order to maintain message sequencing. The Singapore network has 16 or 17 SPs within each SP region, but the following analysis only uses 6 SPs, as the computer memory and execution time required for simulations were found to be greatly reduced.

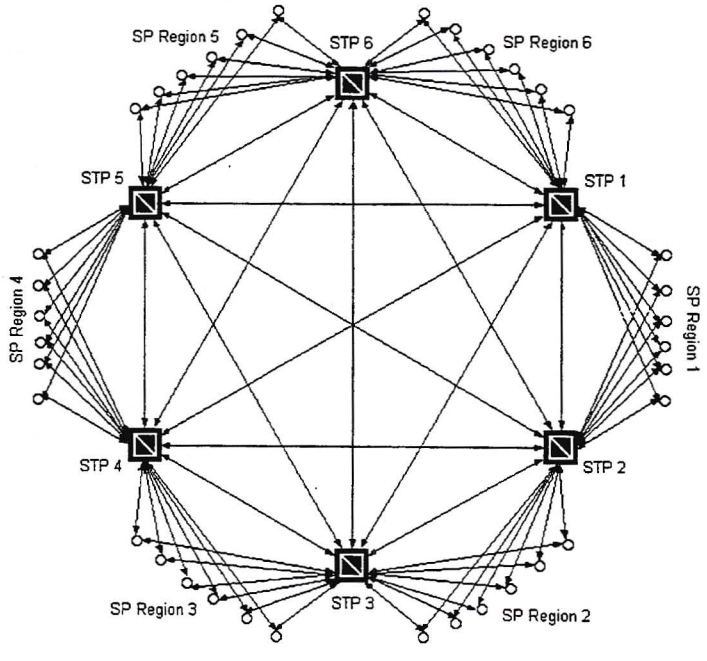


Figure 3-4. SS7 network of fully connected STPs.

To demonstrate the effectiveness and accuracy of the analysis, the numerical results presented here are compared to those obtained by simulations. Unlike the analysis, the simulation explicitly models the call set-up message sequence, the application level recovery timers and the effect of link transmission delays. Routing tables and a description of the simulation model are given in Appendix A.2 and A.3, respectively. Throughout the subsequent discussion the following parameters are used (unless otherwise stated):

- Each STP has a message handling capacity of $\mu = 1000$ messages/second.
- From the network structure: $N = 6$, $M = 6$ and $M_{Rm} = 6$ $(1 \leq m \leq M)$
- The input and output level 3 processing buffers in the SPs have a message handling capacity of $\mu_{SP} = 5000$ messages/second.
- All the signalling links in the network have a transfer rate of $\mu_L = 128$ kb/s.
- The customer reattempt probability, c, is 0.7.

- REL messages are repeated a maximum of 12 times, therefore $r = 12/13$ is used in the analysis. In the simulation, timers T1 = 5 seconds and T5 = 60 seconds are used for the call release procedure.
- The individual message lengths are $\sigma_{IAM} = 48$ octets, $\sigma_{ACM} = 20$ octets, $\sigma_{ANM} = 18$ octets, $\sigma_{REL} = 22$ octets and $\sigma_{RLC} = 17$ octets.

3.4.1 Network Wide Overloads

In this scenario the analysis assumes that the first offered load (or new call arrival rate) from each region is equal. A call originating at a SP has an equal probability of being destined to any of the other 35 SPs in the network. For example, if the first offered load is 70 calls/sec from each region, then 10 calls/sec (70×5/35) would be destined to the originating region and 12 calls/sec (70×6/35) to each of the other regions.

The simulation results presented in this section and the subsequent sections were obtained by commencing the simulations with a first offered load that is well above the maximum call handling capability of the network. Once the system reaches equilibrium, the results are recorded and the first offered load is reduced slightly. The system is again allowed to reach equilibrium before the results are recorded and then the first offered load is again reduced. This process is repeated until the network enters the uncongested state.

3.4.1.1 Single Discard Threshold

The single discard threshold scheme is a special case of the multiple discard threshold scheme, where $K_1 = K_2 = K_3$. Discarding here is analogous to the message discard flow control procedure defined for the International Option, since the message priorities are ignored and excess messages are simply discarded once the level 3 buffer reaches its maximum capacity. This scheme is also representative of current implementations that simply discard messages when the level 3 buffer resources are exhausted.

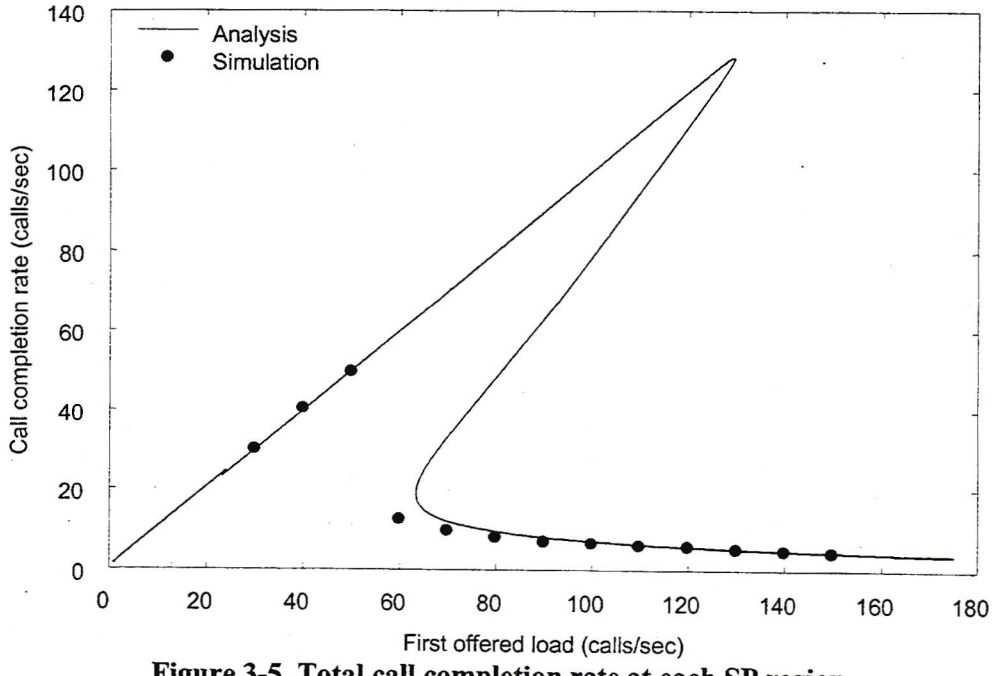


Figure 3-5. Total call completion rate at each SP region (for a single discard threshold scheme with $K_1 = 100$).

Figure 3-5 shows the call completion rate for various first offered loads from each SP region, when $K_1 = 100$ messages. The shape of the curve obtained is similar to that observed by

Rumsewicz [1993] for a single STP network. The call completion rate increases linearly until the first offered load reaches approximately 130 calls/sec, which is the maximum number of calls that can be supported by the network. Then the curve drops and bends back on itself twice. This yields three possible solutions for some of the first offered loads.

The linear (or top) section of the curve corresponds to the call completion rates obtained when the network is in the uncongested state. When the network is operating in this region, all the offered calls are successfully completed and none of the call set-up messages are discarded in the STPs. In this region of the curve, the message arrival rate at each STP is less than the processing speed (μ) of the level 3 processors. Figure 3-6 shows the message arrival rate at each STP together with the traffic load generated by each SP region. Note; the message load arriving at each STP is slightly higher than the input traffic load from each region since transit traffic from the other STPs also contributes to a STP's input traffic load.

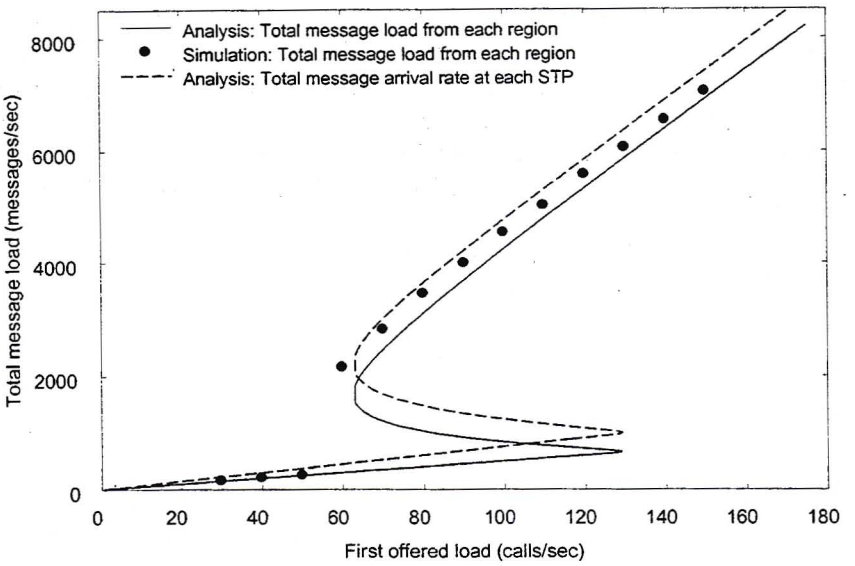


Figure 3-6. Total message loads leaving each SP region and arriving at each STP (for a single discard threshold scheme with $K_1 = 100$).

The bottom section of the curve, in Figure 3-5, corresponds to call completion rates obtained when the entire network is in the congested state. Only a small fraction of the offered calls (including reattempts) are successfully completed during congestion. Part of the congested section also lies in the region with three solutions. This suggests that the network can become congested even if the first offered load from each SP region is less than the network's maximum call handling capacity. The transient analysis, in Chapter 4, shows that if the network is subjected to a brief period of overload (above the network's call handing capability) for a long enough period to induce congestion, then the network can remain in a congested state even after the overload traffic has been removed. For example, consider a network where the first offered load from each region is 70 calls/sec (i.e. each STP is operating at ~53% of its maximum message handling capability under normal conditions). Now, if the first offered load doubles and persists long enough for the network to become congested and then returns to 70 calls/sec, the network's operating point would move along the bottom section of the curve while it is in the congested state. In order to return to the uncongested state, the first offered load would have to drop below 60 calls/sec for a long enough period to allow the system to recover. The simulation results illustrated in Figure 3-5 (and in the subsequent figures) depict a scenario where the system commences in the congested state and the first offered load is gradually reduced until the system returns to the uncongested state. Both the top and bottom sections of the curve represent stable equilibrium operating regions. However, since the arrival process is Poisson in nature, it is possible for the first offered load to increase or decrease significantly for

some period of time to allow for the operating point to move between the two equilibrium regions.

The middle section of the curve represents a meta-stable region. A network operating in this region would drift towards one the two stable regions once a slight perturbation is experienced in the traffic arrival rate. This behaviour can be observed in the transient analysis of Section 4.3.4 and is also confirmed by Rumsewicz [1993]. The middle region is also not practically realisable in simulations.

While the network is in the congested state, signalling traffic between the SPs is dominated by REL messages (Figure 3-7). For example, when the first offered load is 65 calls/sec, and the network is in the congested state, at least 56% of the messages arriving into the network are REL messages, and the ratio of REL messages to the total message load increases for higher first offered loads. The large volume of REL messages is therefore sufficient to keep the network in a congested state at the lower first offered loads and it also impedes the throughput of the other signalling messages. The blocked IAM, ANM and RLC messages help to maintain the high REL volume by generating additional REL messages. The reattempt behaviour of IAM and REL messages further aggravates the congestion situation by generating additional traffic. The IAM message load can thus be greater than 130 calls/sec even though the first offered load contributes to only half (or less) of the total IAM traffic volume.

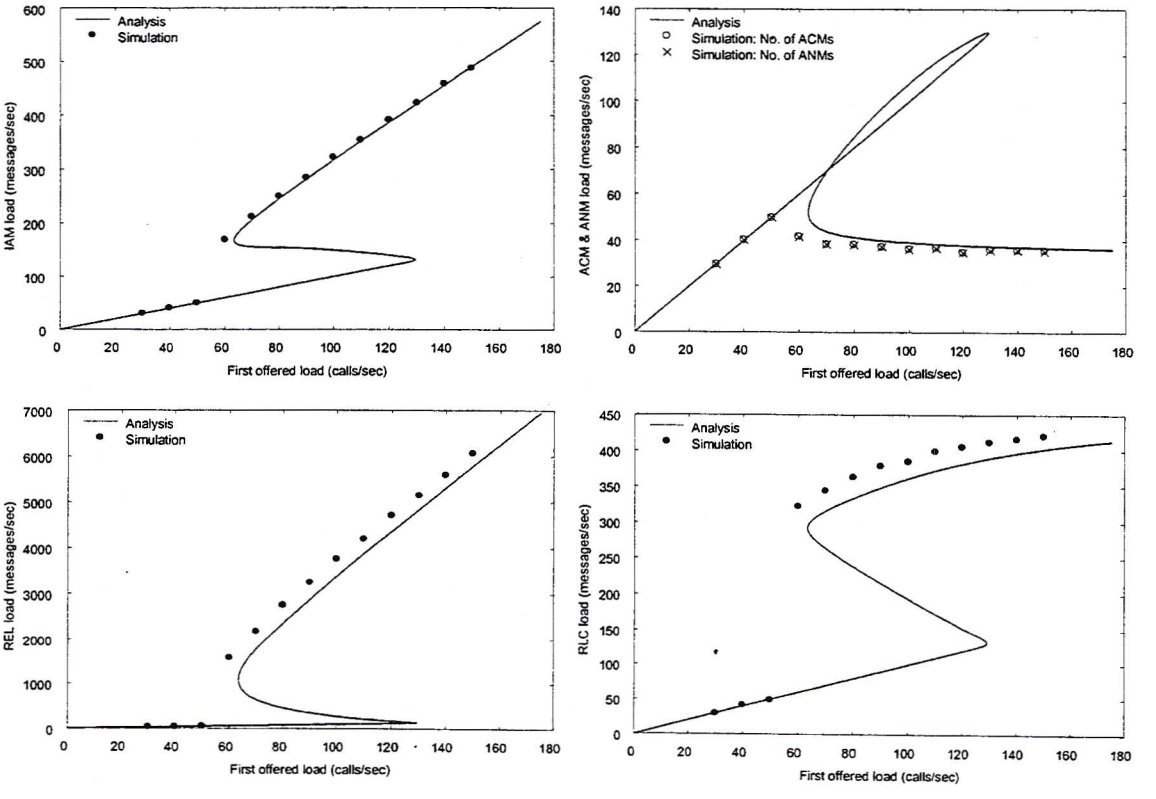


Figure 3-7. Individual message loads from each SP region.

The analytical results obtained, in Figure 3-7, for the IAM, ACM and ANM message loads are very good approximations when compared to the simulation results, but the REL and RLC loads observed in simulations are slightly higher than those obtained by analysis. This discrepancy is due to the correlation that exists in the REL message streams. The source nodes become synchronised with each other soon after the onset of congestion and a burst of REL messages is repeated periodically every T1 seconds (discussed further in Chapter 4). The burst of REL messages exhausts the STP queues within a small period of time, and any additional REL messages that are received thereafter are discarded. The successful REL messages then provoke

a burst of RLC messages. The RLC messages also exhaust the STP queues within a small period of time and the subsequent RLC messages are then discarded. The REL and RLC messages are therefore effectively subjected to a slightly higher average discard probability than the other traffic streams, and hence more REL messages are generated by the reattempt procedures.

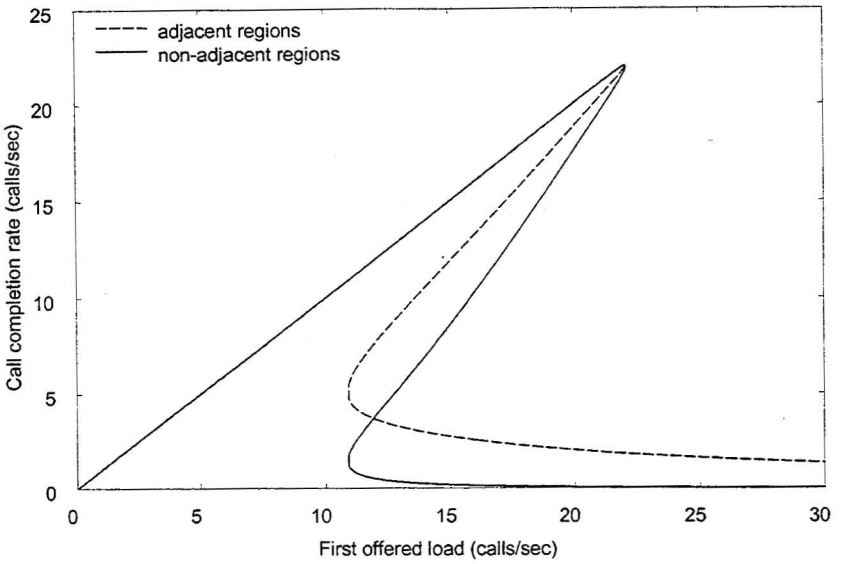


Figure 3-8. Call completion rate for calls to the adjacent and non-adjacent regions.

Figure 3-8 examines the call completion rates between the adjacent and non-adjacent regions. Traffic between the adjacent regions has to traverse a single STP, while traffic between the non-adjacent regions is switched through two STPs. The above figure shows that during a network wide congestion situation the call completion rate between the non-adjacent regions is zero (or negligible), while the few calls that are successful reside between SPs in the same or adjacent regions. The traffic between the non-adjacent regions is composed almost completely of REL messages. The other messages that are successfully admitted into the first STP are usually discarded at the second STP. In addition, throughput to the non-adjacent regions decreases as the first offered load is increased (Figure 3-9). Buffer resources are therefore wasted on messages that are accepted at the first STP but are then discarded at the second STP. The following proposals could help to improve the overall call completion rate by ensuring optimum utilisation of the STP buffer capacity:

a) Discard all message traffic to the non-adjacent regions during congestion: In this scheme the flow control mechanisms favour calls between SPs in the same region and adjacent regions, and therefore messages that are admitted into the first STP will successfully reach their destination. Consequently, messages throughput to the non-adjacent regions suffers, and will drop to zero.
b) Implementation of the buffer allocation algorithms discussed in Section 2.2.2.2 in the STP buffers: For example, Lam & Luke Lien's [1981] *input buffer limit* scheme distinguishes between transit traffic and new incoming traffic. Here, new messages may only consume a finite capacity, while transit messages are able to fill the entire buffer. Transit messages therefore have a lower probability of discard and consequently message throughput to the non-adjacent regions is improved as buffer resources are utilised more efficiently.

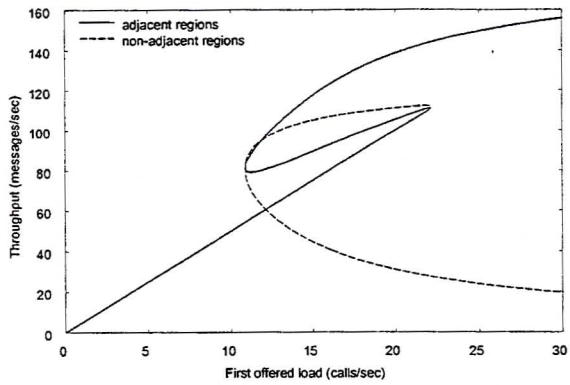


Figure 3-9. Message throughput for traffic to the adjacent and non-adjacent regions.

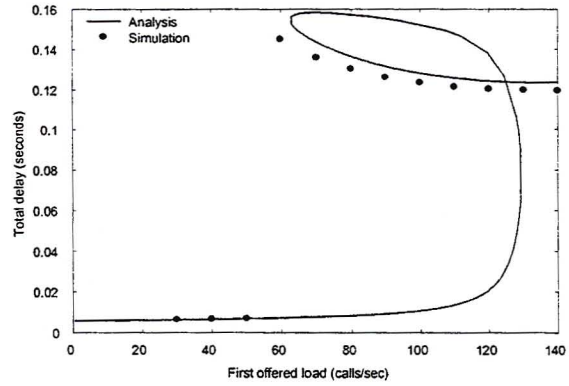


Figure 3-10. Total end-to-end message transfer delay.

Figure 3-10 shows the average end-to-end transfer delay for an arbitrary but successful message in the network. Even though the traffic loads are higher for very large first offered loads, the average message transfer delay decreases and is less than the delay experienced when the network is operating in the meta-stable region. This is due to the higher message throughput attained to the non-adjacent regions when the network is in the meta-stable region, and as network throughput to the non-adjacent regions deteriorates the message transfer delay is dominated by the waiting time endured in a single STP's processing buffer. At very low traffic loads the link level delay accounts for more than 50% of the end-to-end message transfer delay, while at very high loads the effect of STP delays becomes more pronounced.

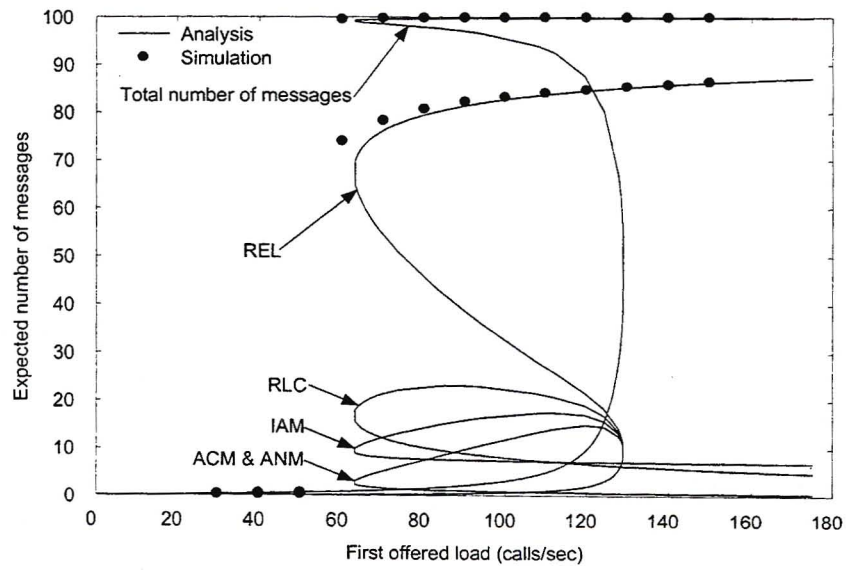


Figure 3-11. Expected number of messages in a STP queue.

Figure 3-11 shows the expected number of messages in a STP buffer for various first offered loads (to avoid cluttering the figure with too many points, only some of the simulation results are shown). During congestion the STP routing buffer is filled to its maximum capacity and REL messages consume over 75% of the available resources. Increasing the first offered load, while the network is congested, increases the number of REL messages within the routing buffer at the expense of the other signalling messages. The throughput of the other messages and the call completion rate consequently deteriorates.

3.4.1.2 Multiple Discard Thresholds

The objective of schemes with multiple discard thresholds is to assign a higher priority to the important messages, thereby reducing their probability of discard and ultimately improving the

overall network performance. The previous section showed that network traffic was dominated by REL messages during congestion. One should therefore be able to reduce the network traffic volume by assigning a higher priority to the REL messages, and thus improve their throughput. Furthermore, by assigning a much higher priority to ANM and RLC messages one would be able to improve the call completion rate and increase the number of free voice circuits, during congestion scenarios. The successful ANM messages could also postpone the arrival of the REL messages, while successful RLC messages would stop the generation of further REL messages for the associated voice circuit. Lastly, by assigning the lowest priority to IAM messages, the message load is reduced to three messages per failed call attempt, compared to five messages if the IAM message is successful. In most studies on SS7 congestion and flow control, employing multiple message priority schemes, the lowest priority (0) is assigned to the IAM messages, ACM and REL messages are assigned a higher priority (1) and ANM and RLC messages are assigned the highest priority (2). This section examines the performance of the above message priority scheme and various other message priority schemes, and assesses their relative advantages and disadvantages.

To easily make reference to the priority assignments used in multiple discard threshold schemes, a five digit notation of the type $abcde_{ISUP}$ is used to distinguish between the different schemes; where a is the priority of the IAM messages, b is the priority of the ACM messages, etc. (e.g. the priority scheme described above would be denoted by the notation: 01212_{ISUP})

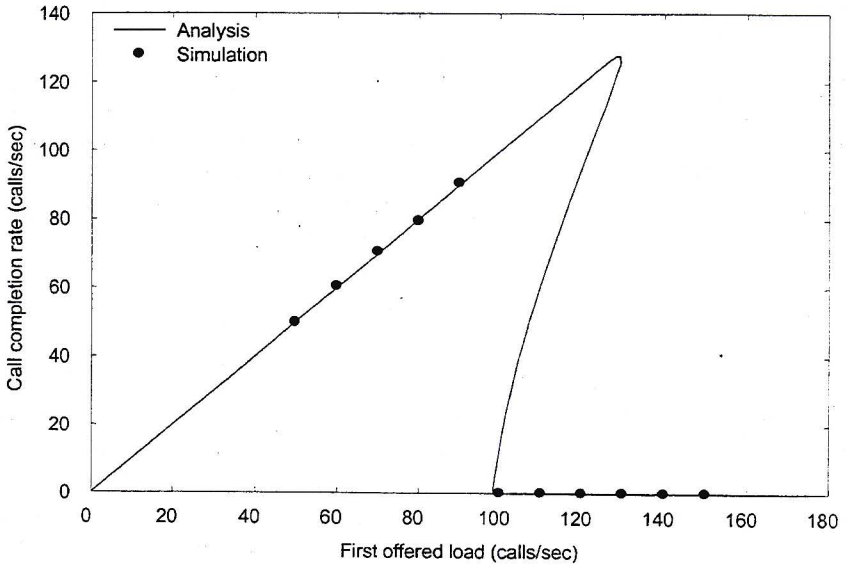
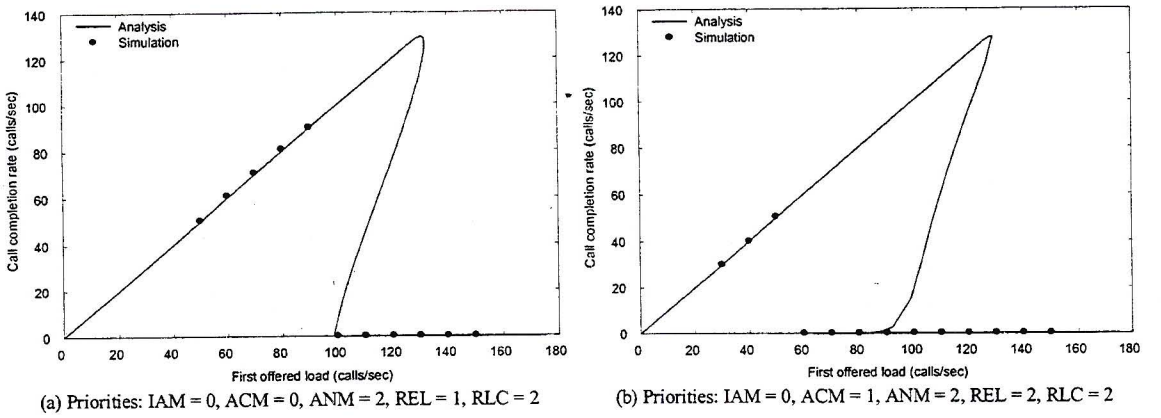


Figure 3-12. Total call completion rate at each SP region for the 01212_{ISUP} priority scheme ($K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

Figure 3-12 shows the call completion rate obtained for the 01212_{ISUP} priority scheme when $K_1 = 100$, $K_2 = 120$ and $K_3 = 130$ messages. The curve resembles the call completion rate curve obtained for the single discard threshold scheme. Again, there are three possible solutions for some of the first offered loads, but in this scheme the call completion rate drops to zero, rather than becoming very small, during congestion. As in the previous section the REL messages still predominate in the network traffic load, during congestion. However, here the volume of REL messages into the network is significantly reduced and as a result a congested network will return to an uncongested state when the first offered load drops below a 100 calls/sec for a long enough period. Section 4.3.1.2 examines the dynamic behaviour of a network as it returns to the uncongested state after a brief period of congestion. Furthermore, unlike the single discard threshold scheme, if all the STPs are operated at below 66% of their message handling capacity (i.e. less than 85.8 calls/sec originate from each region) under normal conditions, the network would be able to recover from congestion once the first offered load has returned to its normal level.



(a) Priorities: IAM = 0, ACM = 0, ANM = 2, REL = 1, RLC = 2

(b) Priorities: IAM = 0, ACM = 1, ANM = 2, REL = 2, RLC = 2

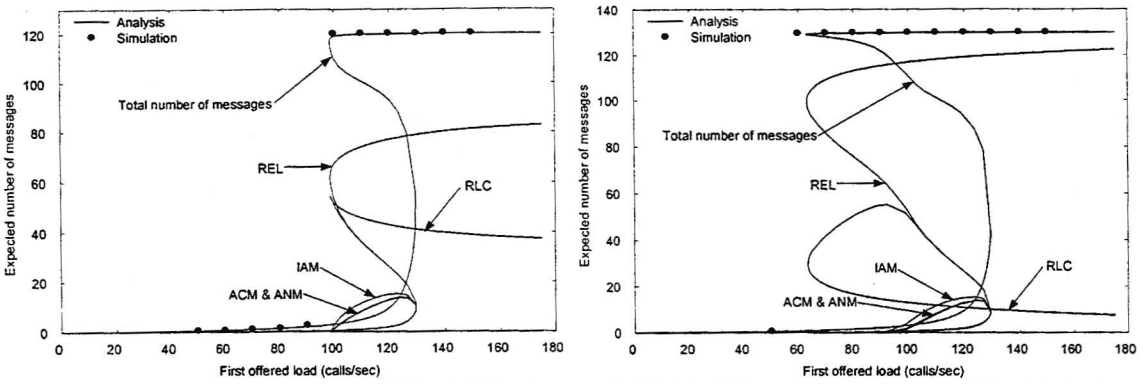
(c) Priorities: IAM = 0, ACM = 0, ANM = 2, REL = 2, RLC = 2

(d) Priorities: IAM = 1, ACM = 0, ANM = 2, REL = 1, RLC = 2

Figure 3-13. Total call completion rate at each SP region for different multiple discard threshold schemes ($K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

The call completion rate obtained for various other priority assignment schemes is shown in Figure 3-13. The general observations and conclusions derived from the above results are as follows:

- In priority schemes where both the IAM and ACM messages are assigned the lowest priority – zero, only a small improvement in the maximum call handling capability of the network is evident when compared to similar schemes where the ACM messages are assigned a priority of one.
- In priority schemes where REL messages are assigned the highest priority, the congested region is extended to a lower first offered load. As in the single discard threshold scheme, the first offered load would have to drop below 60 calls/sec for a long enough duration before the network can return to the uncongested state. Assigning the highest priority to REL messages simply results in more RLC messages being discarded. RLC discarding invokes more REL message reattempts, which effectively intensifies the traffic volume. The 01212_{ISUP} and 00212_{ISUP} priority schemes are therefore the most favourable priority assignments as they have lower network traffic loads and they allow the network to recover at a much higher first offered load.
- In the 10212_{ISUP} priority scheme the network has a maximum call handling capacity of ~155 calls/sec. The disadvantage of this scheme is that only a few ACM messages are admitted into the network during overloads. If no ACM is received by the originating exchange, the calling party does not hear a ringing tone and could hang-up if the call is not answered within a short period of time. Nevertheless, this scheme does have advantages in applications where calls are answered by automated services (or a large pool of operators) that are capable of supporting a large number of short duration calls, e.g. tele-voting and call-in competitions. Section 3.4.3 examines the advantages of this scheme in a focused overload scenario (possibly a more realistic overload scenario where this scheme would be applicable).



(a) Priorities: IAM = 0, ACM = 1, ANM = 2, REL = 1, RLC = 2

(b) Priorities: IAM = 0, ACM = 1, ANM = 2, REL = 2, RLC = 2

Figure 3-14. Expected number of messages in a STP queue.

Figures 3-14 (a) and (b) show the message occupancy of STP buffers using the 01212_{ISUP} and 01222_{ISUP} message priority schemes, respectively. In all of the above investigations, when REL messages are assigned a priority of one, the average number of messages in the STP buffer is approximately equal to K_2. But when the REL messages are assigned a priority of two, the average number of messages in the STP buffer is approximately equal to K_3. In the latter scheme the REL messages consume most of the available buffer resources, at the expense of the RLC messages. Hence, most of the RLC messages are discarded and the subsequent REL reattempts maintain the very high traffic volumes that are observed at the lower first offered loads. In both cases, the number of REL and RLC messages is sufficient to keep the total number messages in the STP processing buffer above the level one discard threshold, during congestion, and therefore no IAM messages are admitted into any of the STPs. As a result, no ACM and ANM messages are generated and therefore no calls are successful.

In Figure 3-14 (a), the number of RLC messages in the STP processing buffer decreases for an increasing first offered load. Here, the reduction in the number of RLC messages present in the STP for higher first offered loads is due to the reduction in the number of REL messages that are successfully delivered to the non-adjacent regions. Hence, the majority of the RLC messages are attributed to messages between the adjacent regions. Consequently, message throughput to the non-adjacent regions is also very small. The performance of the multiple discard threshold schemes could therefore be improved with the application of buffer allocation algorithms that give priority to the transit traffic. As discussed in Section 3.4.1.1, these schemes could improve REL message throughput to non-adjacent regions and the STP buffers would also be better utilised.

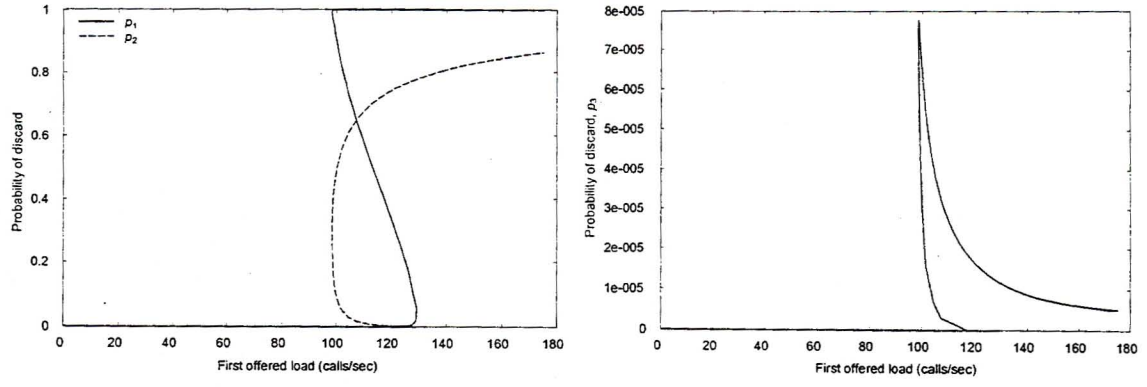


Figure 3-15. Message discard probabilities at a STP (for 01212_{ISUP} priority scheme with $K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

Figure 3-15 shows the probability of a message being discarded at each of the three discard thresholds in the 01212_{ISUP} priority scheme. As expected, through inspection of the buffer utilisation graphs, the probability of discard (when the network is congested) at the first discard threshold is one, i.e. all the IAM messages are discarded. The probability of discard at the second discard threshold is a function of the REL and RLC input loads. But unlike p_1 and p_2, the probability of discard at the level 3 threshold decreases for increasing first offered loads, when the network is in the congested state. This decrease is due to the reduction in the number of RLC messages transmitted between the non-adjacent regions.

3.4.1.3 Selection of the STP discard thresholds

This section studies the impact of threshold settings on network performance. But, rather than carry out an exhaustive search for the optimal choice of thresholds, the threshold settings in the single priority (00000_{ISUP}) and two priority (01111_{ISUP}) schemes are examined to determine appropriate settings for the lower threshold limits in the three priority scheme.

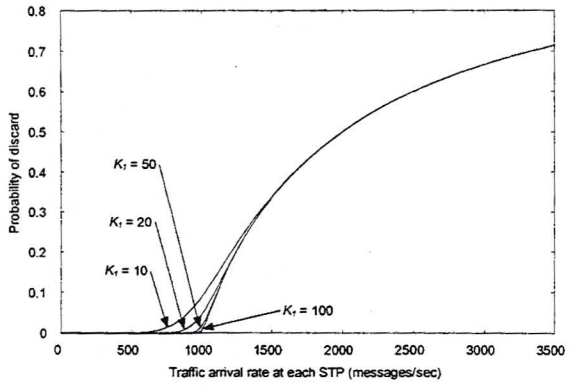


Figure 3-16. Probability of discard, p_1, in a single priority scheme ($K_1 = K_2 = K_3$).

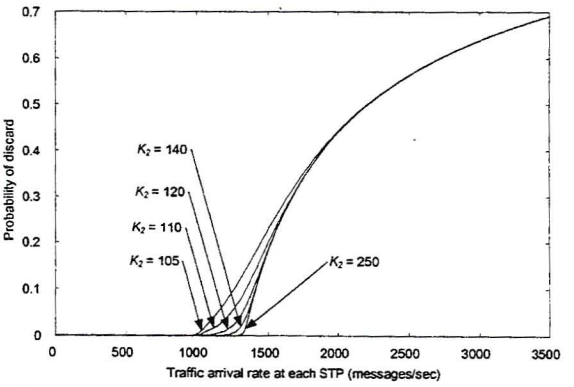


Figure 3-17. Probability of discard, p_2, in a two priority scheme ($K_1 = 100$ messages and $K_2 = K_3$).

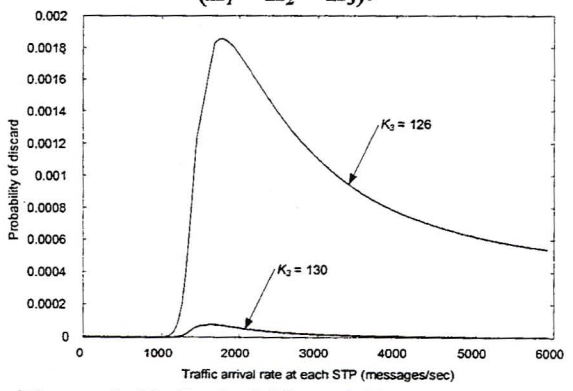


Figure 3-18. Probability of discard, p_3, in a three priority scheme ($K_1 = 100$ messages and $K_2 = 120$ messages).

K_3 (messages)	Maximum probability of discard
123	24.61×10^{-3}
126	1.86×10^{-3}
130	77.74×10^{-6}
140	39.21×10^{-9}

Table 3-1. Maximum value of p_3 for various sizes of K_3 ($K_1 = 100$ messages and $K_2 = 120$ messages).

In the single priority scheme (Figure 3-16), increasing the buffer threshold above 50 messages does not significantly decrease the probability of discard for traffic arrival rates close to 1000 messages/sec. The probability of discard is the same at very high traffic loads, for all possible threshold settings. No substantial performance improvement can thus be achieved by increasing the buffer size above 100 messages. A very large level 3 processing buffer will only allow the network to support the overload traffic for a slightly longer period, before message discarding commences.

In the two priority scheme (Figure 3-17), $K_1 = 100$ messages and K_2 was varied over a range of values. During congestion, when very high volumes of REL messages arrive at each STP, $p_1 \approx$

1.0 (as in Figure 3-15) and none of the IAM messages are admitted into the network. Increasing the buffer threshold above 250 messages in the two priority scheme does significantly improve performance. At a traffic load of ~1325 messages/sec the arrival rate of the high priority messages is ~1000 messages/sec, and the discarding of high priority messages only commences at this point. However, traffic arrival rates from 1000 messages/sec to ~2200 messages/sec lie in the meta-stable region of the call completion rate curve, therefore a buffer size of 120 messages would be a more reasonable and practical choice.

In the three priority scheme (Figure 3-18 and Table 3-1) the level three probability of discard reaches its maximum value when the maximum number of RLC messages are present in the network. For $K_3 \geq 130$ messages, almost all the RLC messages are successful and therefore the effect of level 3 discarding on the system's throughput performance is negligible. Here, the buffer resources that are made available by the processed REL messages are occupied by their subsequent RLC messages within a small period of time. As a result, the buffer resources available exclusively for priority 2 messages are under utilised. Setting the third discard threshold size to at least 130 messages would therefore be appropriate. The low probability of discard obtained at the third threshold is also highly favourable since it indicates that network management messages (which are usually assigned the highest priority), including TFC messages, will experience minimal discarding within the network. In the single priority and two priority schemes the probability of discarding high priority messages is very large during congestion and in both these cases it is possible that the network management messages would never reach their destinations.

These results suggest that selecting very large buffers does not necessarily improve the network performance under steady state conditions. Contrary to previous studies by Northcote & Rumsewicz [1995] and Kant [1997] where very large buffer thresholds are selected on the basis of the maximum tolerable transfer delay, the above results show that smaller buffers not only provide the same probability of discard but ultimately they also provide the same call completion rates and throughput with the advantage of a smaller queuing delay in the processing buffer.

Analyses through simulations (not shown) for various arbitrary discard threshold setting confirm the accuracy of the above results. In all cases the average number of messages in the STP processing buffer, for the 01212_{ISUP} priority scheme, is approximately equal to K_2. These results are representative of a worst case scenario where feedback control mechanisms are not available or are either ineffective or too slow to react to buffer congestion. Experiments with level 3 feedback controls show that the average number of messages in the buffer is usually less than K_2 (refer to Section 7 and [Northcote & Rumsewicz, 1995]).

3.4.1.4 Different Network sizes

The previous sections examined the performance of a network with a backbone of six fully interconnected STPs and six SP regions. This section investigates the influence of network size on performance. Here the network is assumed to consist of backbone of n STPs and n SP regions, where the SPs in each region are connected to their two adjacent STPs (as in the six STP network).

Figure 3-19 shows the call completion rate versus the first offered load from each SP region, when various sizes of fully connected STP networks are examined. The call completion rate obtained at each region in a three STP network are identical to that obtained in a single STP network [Rumsewicz, 1993], since all the messages are routed through one STP. For very large networks the maximum number of calls supported from each region is $\sim\mu/10$ calls/sec.

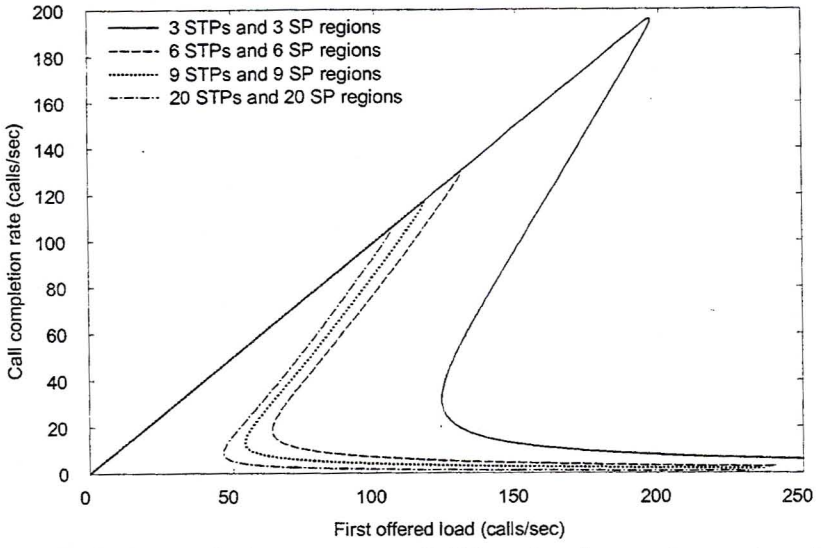


Figure 3-19. Call completion rate at each SP region for various network sizes ($K_1 = K_2 = K_3 = 100$).

Consider the following where the message handling capability of the STPs in both the three and six STP networks is assumed to be $\mu = 1000$ messages/sec:

(a) In a three STP network, a maximum of 200 calls/sec is supported from each region, i.e. 600 calls/sec in the entire network. For a first offered load of 160 calls/sec from each region (i.e. there are 480 calls/sec in the network), each STP operates at 80% of its maximum capacity (when the network is not congested).

(b) In a six STP network, a maximum of 130 calls/sec is supported from each region, i.e. 780 calls/sec in the entire network. For a first offered load of 80 calls/sec from each region (i.e. there are 480 calls/sec in the network), each STP operates at 62% of its maximum capacity (when the network is not congested).

As expected the six STP network is able to support a larger network load and the traffic load imposed on each STP for the same number of calls is smaller. But in (a) the first offered load from each region has to drop below 120 calls/sec (60% of the maximum capacity) before the network can recover from congestion, while in (b) the first offered load has to drop below 60 calls/sec (46% of the maximum capacity). In both cases the network wide load has to be reduced to below 360 calls/sec. This suggests that even though a six STP network can support more network traffic than a three STP network, increasing the network size does not effectively allow the network to recover from congestion at a higher first offered load but instead requires a much lower traffic load to each STP to achieve the same effect.

3.4.1.5 The impact of caller reattempts and REL reattempts

Most studies, that consider the effect of customer behaviour on network performance, assume a reattempt probability of 0.7 ([Rumsewicz, 1994] & [Smith, 1994]). This value is generally assumed to serve as a realistic approximation and represents an average of between 3 and 4 attempts from a customer. But this value could be higher, possibly over 0.9, if automatic redial facilities are considered.

Figures 3-20 and 3-21 illustrate the sensitivity of call completion rates to the customer's reattempt probability and the release procedure's reattempt probability. In both instances the region with multiple solutions only exists for the higher reattempt probabilities. Furthermore, the call completion rate is higher for the lower reattempt probabilities, as fewer messages are generated by the failed call attempts and failed release procedures [Rumsewicz, 1993]. These results therefore highlight the necessity to account for the effects of customer behaviour and application level recovery actions when quantifying the performance of a communications network during an overload.

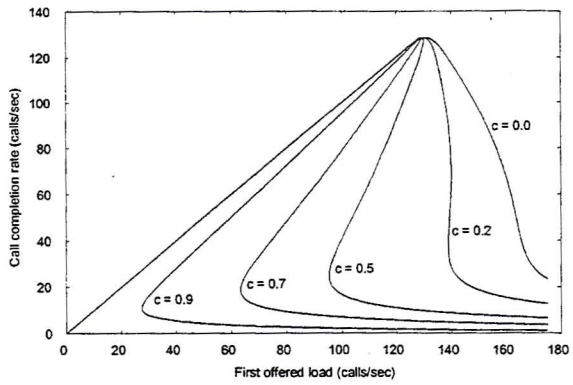


Figure 3-20. Call completion rates for various customer reattempt probabilities ($K_1 = K_2 = K_3 = 100$, r = 12/13).

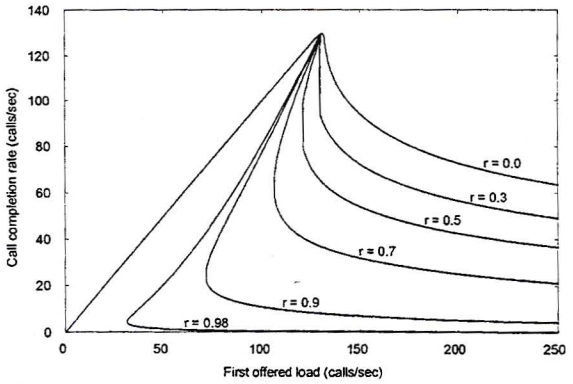


Figure 3-21. Call completion rates for various REL reattempt probabilities ($K_1 = K_2 = K_3 = 100$, c = 0.7).

These results also suggest that suppressing caller reattempts, when the rate of call failures is high, could be a viable solution in networks where feedback control mechanisms are not implemented in the STP. Congestion detection, in this case, can be accomplished with the implementation of a timeout based congestion control mechanism that is described in Section 2.2.2.5 and Jain [1990].

3.4.2 Network failures

In a telecommunications network, network survivability in the event of node or link failures is important in order to maintain the required grade of service. This section examines the call handling capability of the six STP network in the event of single or multiple STP failures. However, even though the results presented here concentrate on STP failures, similar results could be obtained for non-symmetrical signalling network configurations. The following solutions are obtained by simply modifying the variables associated with message routing to now consider the new routes followed after the node failures.

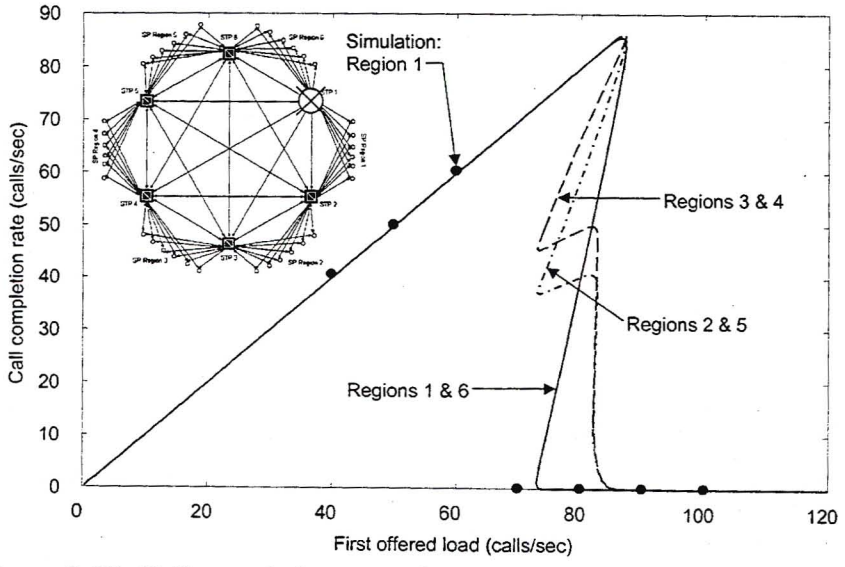


Figure 3-22. Call completion rate of each region when STP-1 has failed (01212_{ISUP} priority scheme with $K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

Figure 3-22 shows the call completion rate for the 01212_{ISUP} message priority scheme when STP-1 has failed. In failure scenarios all six regions perform similarly while the entire network

is uncongested, but different solutions are obtained for each region when part of the network or the entire network is congested or operating in the meta-stable region. For the scenario shown in Figure 3-22 the performance of each of the six regions can be represented by three curves, where the performance of regions 1 and 6 is equivalent, the performance of regions 2 and 5 is equivalent, and the call completion rate of regions 3 and 4 is also equivalent. Unlike the quad STP structure (which requires a 100% redundancy), the six STP network only requires a 50% redundancy. In the event of a single STP failure, the two adjacent STPs are subjected to a 50% load increase. The network should therefore be designed to operate at less than 2/3 of its maximum call handling capacity under normal conditions (i.e. the call arrival rate should be less and 86 calls/sec from each region). For example, if the call arrival rate from each region is 68 calls/sec under normal conditions (80% of the engineered maximum load of 86 calls/sec) then the network would be able to recover from congestion once the first offered load returns to normal.

In the above figure, the region with three solutions still exists for first offered loads below the maximum call handling capability of the network, but only part of the network remains congested once the overload traffic is removed. STPs 2 and 6 absorb the entire message load to and from regions 1 and 6, respectively, and are therefore more susceptible to congestion. These STPs are also the first to become congested when the first offered load exceeds 86 calls/sec, before the congestion propagates to the other STPs. In the region with three solutions, the message volume arriving at the other STPs is less than 1000 messages/sec and therefore the other SP regions are able to maintain a non-zero call completion rate. For example, when the first offered load is 80 calls/sec and STPs 2 and 6 are congested then all the calls originating from region-3 that are destined to SPs in regions 2, 3 and 4 are successful, while at least half of the calls to region-5 are successful and none of the calls to regions 1 and 6 are successful. Hence, the call completion rate for region-3 is approximately 46 calls/sec.

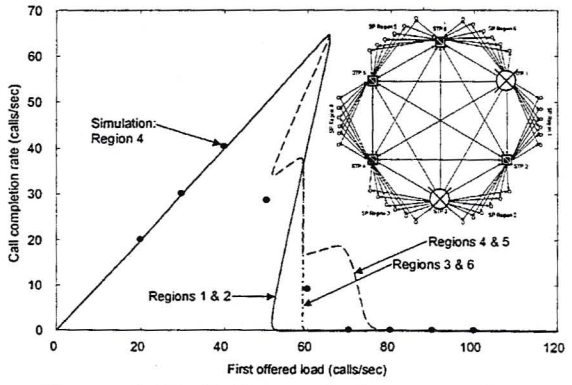


Figure 3-23. Call completion rate when STP-1 and STP-3 have failed (01212_{ISUP} priority scheme with $K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

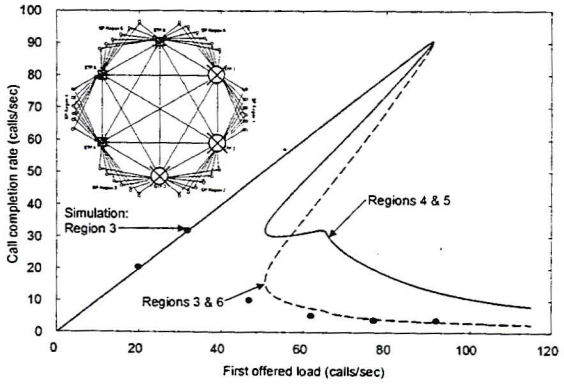


Figure 3-24. Call completion rate when STP-1, STP-2 and STP-3 have failed ($K_1 = K_2 = K_3 = 100$).

Figures 3-23 and 3-24 show the call completions rates obtained for networks with two and three STP failures, respectively. For the failure scenario analysed in Figure 3-23, STP-2 supports the entire traffic load to and from regions 1 and 2 and is therefore the first node to become congested. Congestion then propagates to STPs 4 and 6 and eventually to STP-5 if the first offered load is very large. Since STP-5 supports the smallest traffic load in network, regions 4 and 5 are able to maintain a non-zero call completion rate between each other for first offered loads exceeding the maximum call handling capacity of the network. In Figure 3-24 the network has a maximum call handling capacity of 90 calls/sec since there is no traffic to or from regions 1 and 2. The call completion rate for this scenario is non-zero as a single discard threshold scheme is implemented at each STP and each message type therefore has an equal probability of being discarded. The lower call completion rates observed in the simulations are

due to the increased discarding of IAM messages, as the REL load is larger than that obtained by analysis.

The above results indicate that unless failure scenarios are considered in the network planning process it is possible for either part of the network or the entire network to become congested in the event of link or node failures. Likewise, non-symmetrical network configurations could also induce localised congestion. The high traffic load to a congested part of the network could consequently trigger congestion in other parts of the network.

3.4.3 *Focused Overloads*

In this scenario, SP region-1 is the target of all the calls from the overload traffic streams. The network has a background traffic load of 50 calls/sec, from each region. Calls from an SP that correspond to the background traffic stream have a equal probability of being destined to any of the other 35 SPs, while calls corresponding to the focused overload have an equal probability of being destined to any one of the six SPs in region-1. To reduce the complexity of the system, the following analysis assumes that region-1 does not contribute to the focused overload. This type of scenario could represent a media stimulated event where a large number of calls are targeted towards a specific region, exchange or number.

Figures 3-25 and 3-26 show the call completion rates obtained for the 01212_{ISUP} and 10212_{ISUP} message priority schemes. The network utilising the 01212_{ISUP} message priority scheme is able to support a focused overload of 48 calls/sec (98 calls/sec – 50 calls/sec). The network can therefore support a maximum of 540 calls/sec during the focused overload scenario, significantly less than the 780 calls/sec supported during a network wide overload. The network utilising the 10212_{ISUP} message priority scheme is able to support a focused overload of 65 calls/sec (a 35% improvement over the 01212_{ISUP} scheme). In both scenarios, STPs 1 and 2 become congested soon after the maximum call handling capability of the networks is exceeded. Soon thereafter, the congestion propagates to STPs 4 and 5 and the call completion rate of region-4 deteriorates. STPs 3 and 6 are last to become congested, as they carry the smallest volume of overload traffic.

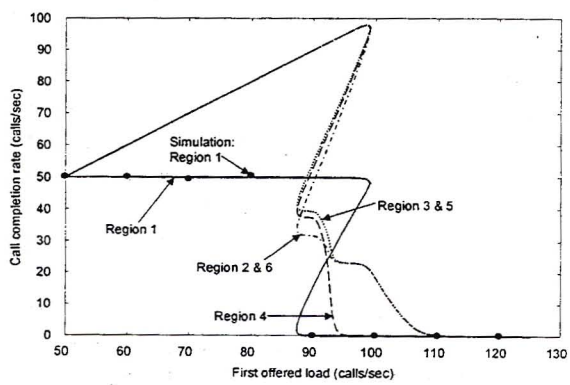


Figure 3-25. Call completion rate for a focused overload to SP region-1 (01212_{ISUP} priority scheme with $K_1 = 100, K_2 = 120, K_3 = 130$).

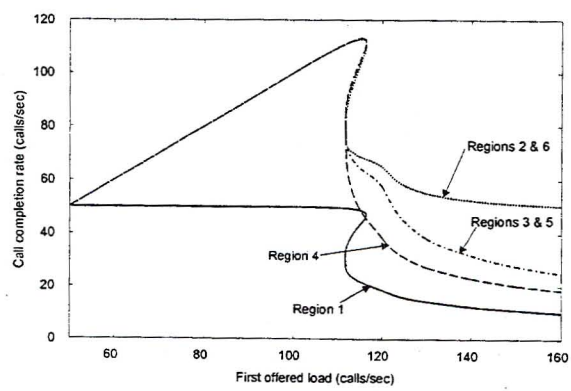


Figure 3-26. Call completion rate for a focused overload to SP region-1 (10212_{ISUP} priority scheme with $K_1 = 100, K_2 = 120, K_3 = 130$).

The results obtained indicate that congestion induced by a focused overload can eventually propagate throughout the network and consequently degrade the overall network performance. However, the 10212_{ISUP} message priority scheme helps to increase the maximum call handling capability of the network, at the expense of the ACM messages. In addition, when the 10212_{ISUP} scheme is used a substantial number of calls from the background traffic load are still successful while the STPs are congested.

3.5 Summary

Previous studies of SS7 network performance analysed simple network structures, where the overload is directed towards a single STP or signalling link set. The analysis in these studies is also limited to the performance evaluation of the congested entity and the source nodes that utilise the affected entity. As these studies were primarily concerned with emphasising the shortcomings of the current SS7 protocol implementation they have not attempted to develop models that can assist in the performance evaluation of large signalling networks.

This chapter developed mathematical models to analyse the performance of a priority based queuing discipline and of signalling network configurations with multiple STPs. In the prioritised queuing discipline, multiple discard thresholds are used to control the admission of the various message types into the queue. The discard process modelled here is analogous to message discarding in the National Option with congestion priorities, when the feedback controls are ineffective. The generalised analytical model developed for the performance evaluation of multiple STP networks has the following improvements over previous models:

- any number of STPs and SPs may be included in the network structure,
- different message priority schemes may be analysed,
- different routing algorithms may be evaluated,
- a source-destination traffic matrix allows for different traffic scenarios to be examined,
- and the analytical model is not confined for a single network configuration.

In addition, the model also accounts for message sequencing and the effects of caller reattempts and application level recovery actions. Comparisons with simulation results show that the analysis provides a good representation of network performance.

The results indicate that large networks are susceptible to sustained overloads once the original source of the overload has subsided. The sustained overloads observed here are similar to the sustained overloads observed in the single STP network by Rumsewicz [1993]. However, in networks with unbalanced traffic loads, node failures or focused overloads, it is possible for only part of the network to be become congested or remain congested once the overload traffic has decreased.

Furthermore, the results show that in large signalling networks, transit messages (i.e. messages that traverse two STPs) have a much lower throughput than signalling messages that only traverse a single STP, especially during severe congestion situations. The call completion rate between SPs in the non-adjacent regions therefore drops to zero when the network is congested. Assigning a higher priority to transit messages could therefore help to better utilise the available resources and improve message throughput.

In addition to the overload traffic, customer reattempts and application level recovery actions are also responsible for the substantial increase in signalling traffic, during congestion situations. Discarded IAM and ANM messages result in failed call attempts that consequently initiate the release procedure and possibly a further call attempt. If a REL or RLC message is discarded then there exists a very high probability that the release procedure will be reattempted. Eventually network traffic is dominated by REL messages. However, fewer REL messages are generated with the 01212_{ISUP} message priority scheme than with the single discard threshold scheme. In the 01212_{ISUP} priority scheme, REL and RLC messages have a higher probability of successful throughput and only three messages are generated per failed call attempt (compared to possibly four or five messages in the single discard threshold scheme, excluding reattempts). The lower message load therefore allows the network to recover from congestion at a higher first offered load.

An investigation on the influence of buffer threshold settings on network performance indicates that very large buffers do not necessarily improve message throughput or the call completion rate once the network is congested; but they only increase the queuing delay endured by signalling messages in the MTP level 3 processing buffer. The probability of discarding high priority messages, such as ANM, RLC and network management messages, in the 01212_{ISUP} message priority scheme is found to be negligible. These messages therefore have a high throughput as the processing buffer is rarely filled to its maximum capacity.

Lastly, an examination of different message priority schemes indicates that the maximum message handling capability of the network is increased when the 10212_{ISUP} message priority scheme is used. In addition, a non-zero call completion rate is also achieved when the network is congested. The improved performance is achieved at the expense of ACM message throughput. This scheme would therefore be beneficial in focused overload scenarios where a large number of calls, with short call ringing times, are directed towards a single region, exchange or number. Generally, the 01212_{ISUP} and 00212_{ISUP} priority schemes were found to be the most favourable priority assignments since the networks' have lower traffic loads, which allows for recovery from congestion at a much higher first offered load.

4. Transient Analysis of SS7 Network Performance in the PSTN

4.1 Introduction

Most analytical models that address the performance of packet switched networks concentrate on system behaviour during steady state conditions. In the past, most researchers have favoured the steady state analytical approach over a transient analysis method because:

a) The system equations are greatly simplified once the time varying quantities are replaced by their means and variances.
b) The steady state equations can be easily solved using standard numerical solution techniques.

The transient analysis is also complicated by the fact that it requires explicit consideration of the underlying stochastic processes, e.g. the distribution of call holding times, and the current state of the system usually depends on all the past events.

Only a few researchers have attempted to develop analytical models for the transient analysis of SS7 networks. Skoog [1988] analysed the buffer occupancy of a link level transmission buffer that is carrying short ISUP signalling messages together with low priority UUI messages. However, his method is limited to the analysis of a single transmission queue. Zepf & Willmann [1991] used a hybrid iterative analysis technique to calculate the link utilisation of a congested link set. Their method is limited to the analysis of a single link set. Rumsewicz [1993] constructed a mathematical model that explicitly considered different message types, message sequencing and the effect of reattempt traffic. His model is limited to the analysis of a single STP's processing queue, with either a single discard threshold or two discard thresholds.

The analysis presented in this chapter extends Rumsewicz's model to allow for the transient analysis of a multiple STP network. As in the previous chapter the transient model considers the network structure, routing and allows for the analysis of different message priority schemes, failure scenarios and different traffic loads from each region. The steady state equilibrium analysis in the previous chapter allows one to dimension a signalling network in order to allow it to operate effectively during periods of overload and failure, while the transient analysis provides a more intuitive perspective of the traffic processes that lead to network congestion and sustain the overload situation. The analysis may also be used to determine how the selection of different system parameters may influence recovery from congestion during very brief periods of overload.

4.2 Transient Mathematical Analysis

Unlike the equilibrium analysis the transient analysis requires explicit consideration of the individual stochastic processes, and there also exists a dependence between the current state of the system and all the previous events. The notation used in this section is identical to that used for the equilibrium analysis, except for the following additions:

i) $\alpha(t)$ is a $M \times M$ source/destination traffic matrix of new call traffic (i.e. the first offered load) at time t and $\alpha_{ij}(t)$ is the number of new call arrivals per second at region i that are designated to region j. The first offered load from region i is given by

$$\alpha_i(t) = \sum_{j=1}^{M} \alpha_{ij}(t).$$

i) The transmission rate of a traffic stream with type T messages at time t is denoted by $\lambda_T(t)$, where T can be IAM, ACM, etc. $\lambda_{T\,ijk}(t)$ is the instantaneous outgoing transmission rate of type T messages from region i to region j on route k (in the forward direction) at time t.

$$\lambda_{T\,ij}(t)=\sum_{k=1}^{R_{ij}}\lambda_{T\,ijk}(t);\quad \lambda_{T\,i}(t)=\sum_{j=1}^{M}\sum_{k=1}^{R_{ij}}\lambda_{T\,ijk}(t)\quad\text{and}\quad \lambda_{T}(t)=\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{R_{ij}}\lambda_{T\,ijk}(t) \qquad \textbf{(4-1)}$$

ii) $P_D(\phi_T,d,i,j,k,l,t)$ is the probability that a message T from region i to region j on route k with priority ϕ_T is *discarded* at time t in lth node. $P_S(\phi_T,d,i,j,k,l,t)$ is the probability that a message T from region i to region j on route k with priority ϕ_T is *successfully* admitted into the queue of the lth node at time t. $d = 0$ or 1 indicates whether the other variables refer to a route in the forward or reverse direction, respectively.

iii) $\lambda_n(t)$ is the arrival rate of messages to STP-n at time t, $\lambda_n'(t)$ is the arrival rate of messages with priorities of one and two to STP-n at time t and $\lambda_n''(t)$ is the arrival rate of messages with a priority of two to STP-n at time t.

iv) T_{IAM} is the timer started when an IAM message is sent. If an answer message is not received before T_{IAM} expires, the call is assumed to have failed, and the release procedure is initiated.

v) T_{REL} is the timer started when a release message is sent (i.e. T_{REL} = T1). If a release complete message is not received before T_{REL} expires the release procedure is reattempted. r = (T5/T1)/(T5/T1+1) determines the probability with which a failed release procedure is reattempted. T1 and T5 are the REL reattempt timers described in Section 1.6.

vi) D is the average queuing delay endured by a signalling message in the level 3 processing buffer of a STP. A more accurate value can be obtained by calculating the instantaneous delay at each STP with equation (3.7). However, experimentation with various values of D indicates that if a delay in the order of one second or less (but greater than the maximum possible delay) is used in the following calculations, the resulting error in the call completion rate is negligible.

vii) The ringing time of a telephone has a uniform distribution $Q(x)$, with a corresponding density function $q(x)$.

viii) The call holding time has an exponential distribution $V(x)$, with a corresponding density function $v(x)$.

ix) The time during which the customer realises that the call has failed and reattempts is a uniform distribution $W(x)$, with a corresponding density function $w(x)$.

The following expression gives the transmission rate of IAM messages that are leaving region m and are destined to region n at time t:

$$\begin{aligned}\lambda_{IAM\,mn}(t)=\alpha_{mn}(t)&+c\int_0^{\infty}\sum_{i=1}^{R_{mn}}\lambda_{IAM\,mni}(t-x-T_{IAM})\sum_{j=1}^{Q_{F\,mni}}P_D(\phi_{IAM},0,m,n,i,j,t-x-T_{IAM})w(x)dx\\&+c\int_0^{\infty}\sum_{i=1}^{R_{mn}}\sum_{j=1}^{Q_{R\,mni}}\int_0^{t-y}\lambda_{ANM\,nmi}\left(t-y-\left(T_{IAM}-x-Q_{R\,mni}D\right)\right)\\&\times P_D\left(\phi_{ANM},1,m,n,i,j,t-y-\left(T_{IAM}-x-Q_{R\,mni}D\right)\right)q(x)dx\,w(y)dy\end{aligned} \qquad \textbf{(4-2)}$$

The first term denotes the total number of new call arrivals or the first offered load to destination n, the second term represents the number of reattempted calls due to IAM discarding in the network and the third term denotes the number of reattempted calls due to ANM discarding in the network. The IAM traffic load across each route, i, from region m to region n is therefore given by

$$\lambda_{IAM\ mni}(t) = \frac{1}{R_{mn}} \lambda_{IAM\ mn}(t) \tag{4-3}$$

To calculate the ANM transmission rates over each route let $z = \Gamma(n,m,i)$ where $1 \le i \le R_{mn}$, then

$$\lambda_{ACM\ mnz}(t) = \frac{1}{R_{nm}} \lambda_{IAM\ nm}(t - Q_{F\ nmi}D) P_S(\phi_{IAM}, 0, n, m, i, Q_{F\ nmi}, t - Q_{F\ nmi}D), \tag{4-4}$$

which is equivalent to the number of successful IAM messages received from each source n.

The number of ANM messages generated is equal to the number of ACM messages generated earlier (assuming that all the calls are answered).

$$\lambda_{ANM\ mni}(t) = \int_0^\infty \lambda_{ACM\ mni}(t-x) q(x) dx \tag{4-5}$$

The REL transmission rate over each route is given by

$$\begin{aligned}
\lambda_{REL\ mni}(t) &= \lambda_{IAM\ mni}(t - T_{IAM}) \sum_{k=1}^{Q_{F\ mni}} P_D(\phi_{IAM}, 0, m, n, i, k, t - T_{IAM}) \\
&+ \int_0^\infty \lambda_{ANM\ nmz}(t - x - Q_{R\ mni}D) P_S(\phi_{ANM}, 1, m, n, i, Q_{R\ mni}, t - x - Q_{R\ mni}D) v(x) dx \\
&+ \int_0^t \lambda_{ANM\ nmz}(t - T_{IAM} + x + Q_{R\ mni}D) \sum_{k=1}^{Q_{R\ mni}} P_D(\phi_{ANM}, 1, m, n, i, k, t - T_{IAM} + x + Q_{R\ mni}D) q(x) dx \\
&+ \lambda_{REL\ mni}(t - T_{REL}) \sum_{k=1}^{Q_{F\ mni}} P_D(\phi_{REL}, 0, m, n, i, k, t - T_{REL}) r \\
&+ \lambda_{RLC\ nmz}(t - T_{REL} + Q_{R\ mni}D) \sum_{k=1}^{Q_{R\ mni}} P_D(\phi_{RLC}, 1, m, n, i, k, t - T_{REL} + Q_{R\ mni}D) r
\end{aligned} \tag{4-6}$$

where the first term denotes the number of REL message generated due to IAM discarding, the second term represents the number of REL messages generated at the end of the successful calls, while the third term denotes the number of REL messages arising from calls lost due to ANM discarding. The fourth term determines the number of REL reattempts initiated due to REL messages that were discarded earlier and the last term calculates the number of REL reattempts due to discarded RLC messages.

The RLC transmission rate is given by

$$\lambda_{RLC\ mnz}(t) = \lambda_{REL\ nmi}(t - Q_{F\ nmi}D) P_S(\phi_{REL}, 0, n, m, i, Q_{F\ nmi}, t - Q_{F\ nmi}D) \tag{4-7}$$

where $z = \Gamma(n,m,i)$ and $1 \le i \le R_{mn}$.

The call completion rate for a particular region m, is given by

$$S_{Cm}(t)=\sum_{i=1}^{M}\sum_{j=1}^{R_{im}}\lambda_{ANM\,imj}\left(t-Q_{F\,imj}D\right)P_S\left(\phi_{ANM},0,i,m,j,Q_{F\,imj},t-Q_{F\,imj}D\right), \quad \textbf{(4-8)}$$

which determines the number of successful ANM messages received by the originating point of the call.

The total number of messages arriving at STP-n at time t is given by

$$\begin{aligned}\lambda_n(t)=\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{R_{ij}}\Bigg(&\sum_{T}\lambda_{T\,ijk}(t)\cdot\psi(i,j,k,1,n)\\&+\sum_{l=2}^{Q_{F\,ijk}}\left(\sum_{T}\lambda_{T\,ijk}(t)P_S\left(\phi_T,0,i,j,k,l-1,t-(l-1)D\right)\right)\cdot\psi(i,j,k,l,n)\Bigg)\end{aligned} \quad \textbf{(4-9)}$$

A similar summation process can be used to obtain $\lambda_n'(t)$ (the arrival rate of messages with a priority of one or two) and $\lambda_n''(t)$ (the arrival rate of messages with a priority of two) for each STP. Even though $\lambda_{T\,ijk}(t)$ is dependent of previous events in the system, all the message streams are assumed to be independent Poisson processes, when determining the equilibrium queue length distribution.

The probability that messages in traffic streams $\lambda_n(t)$, $\lambda_n'(t)$, and $\lambda_n''(t)$ arriving at STP-n will be discarded is given by

$$p_{0n}(t)=\sum_{i=K_1}^{K_3}\pi_{i\,n}(t), \quad p_{1n}(t)=\sum_{i=K_2}^{K_3}\pi_{i\,n}(t) \quad \text{and} \quad p_{2n}(t)=\pi_{K_3\,n}(t) \qquad \textbf{(4-10)}$$

respectively, where π_{in} is the queue length distribution of the routing processor's input buffer, i is the number of messages in the queue and p_{xy} is the blocking probability at threshold 'x + 1' of STP-y

Once the probabilities of discard at each STP are available, P_S and P_D can be calculated from

$$P_S(\phi_m,d,x,y,z,i,t)=\begin{cases}\prod_{n=1}^{i}\left[1-p_{ae}(t+(n-1)D)\right] & \text{for} \quad d=0\\ \prod_{n=1}^{i}\left[1-p_{af}(t+(n-1)D)\right] & \text{for} \quad d=1\end{cases} \quad \textbf{(4-11)}$$

$$P_D(\phi_m,d,x,y,z,i,t)=\begin{cases}p_{ag}(t) & \text{for} \quad i=1 \quad \text{and} \quad d=0\\ p_{ag}(t+(i-1)D)\left[\prod_{n=1}^{i-1}\left[1-p_{ae}(t+(n-1)D)\right]\right] & \text{for} \quad i>1 \quad \text{and} \quad d=0\\ p_{ah}(t) & \text{for} \quad i=1 \quad \text{and} \quad d=1\\ p_{ah}(t+(i-1)D)\left[\prod_{n=1}^{i-1}\left[1-p_{af}(t+(n-1)D)\right]\right] & \text{for} \quad i>1 \quad \text{and} \quad d=1\end{cases} \quad \textbf{(4-12)}$$

where

$$a=\phi_m, \quad e=\Re_{F\,xyzn}, \quad f=\Re_{R\,xyzn}, \quad g=\Re_{F\,xyzi} \quad \text{and} \quad h=\Re_{R\,xyzi}.$$

To solve the above system of equations, one has to assume that the history of the system for time $t < 0$ is known. The simplest analytical solution would assume that all the transmission rates and discard probabilities are zero for $t < 0$, as in the simulation. Alternatively, the transmission rates and discard probabilities for a specific operating point can be obtained via the steady state equilibrium analysis, and then used in the transient analysis for $t < 0$.

Instead of solving the above equations numerically (which could be a very tedious task) for an exact solution, the numerical approximation technique suggested by Rumsewicz [1993] is used to calculate the transmission rates of the various message types. To use this approach, first determine the discrete time functions of equations (4-2) through to (4-8). Next, determine the corresponding probability mass functions q_i, w_i and v_i of $q(x)$, $w(x)$ and $v(x)$, respectively. Finally, select the size of time increments, T_{INC}, for the analysis and then solve the system of equations for each time interval from $t = 0$ seconds. In the following results, a time increment of $T_{INC} = 1$ second and an average STP queuing delay of $D = 1$ second is used.

4.3 Numerical Results and Discussion

This section presents the numerical results obtained from the transient analysis of a fully connected multiple STP network. The network examined here and its system parameters are identical to those used in Section 3.4. The following timer settings and distributions are also used in the transient analysis and simulation model:

- The answer time, $Q(x)$, has a uniform distribution from 2 to 22 seconds.
- The reattempt time, $W(x)$, has a uniform distribution from 2 to 32 seconds.
- The call holding time, $V(x)$, has an exponential distribution, with a mean of 150 seconds.
- The call set-up timeout period $T_{IAM} = 24$ seconds.
- The release procedure timeout period $T_{REL} = 5$ seconds.

The above parameters are consistent with those used in other publications on signalling network performance [Rumsewicz, 1993 and Northcote & Rumsewicz, 1995].

4.3.1 Network Wide Overloads

Both the simulation and analysis are started from an initial state of rest, when no calls or messages are present in the network. A first offered load of 107 calls/sec is introduced into the network at time $t = 0$ seconds. The first offered load is increased to 150 calls/sec at $t = 1300$ seconds, when the average number of messages in the network is constant. The overload is then maintained for a period of 500 seconds before the first offered load is returned to 107 calls/sec. The call arrival traffic patterns, for both the simulation and the mathematical analysis, are illustrated in Figure 4-1. The analytical results, presented below, represent mean values of the system's state (hence the smooth lines), while the simulations explicitly model the random processes that exist in the system and therefore the simulation results vary about the system mean.

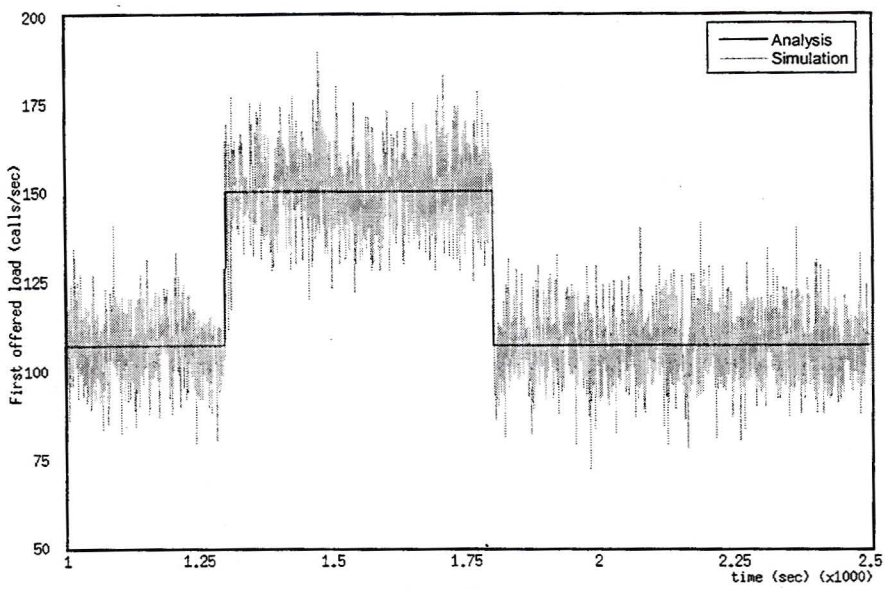


Figure 4-1. Call arrival rate at each SP region (excluding reattempts).

4.3.1.1 Single Discard Threshold

Figure 4-2 shows the predicted and simulated call completion rates for each SP region when $K_1 = K_2 = K_3 = 100$ messages. When the overload is introduced there is an increase in the call completion rate, which grows until the traffic arrival rates exceed the processing speeds of the STPs. The STP processing buffers are then quickly filled until no buffer storage capacity is available. Any further messages that are received are then discarded, until resources are freed by the processed messages. Most of the IAM, ACM and ANM messages from the overload are discarded and the call completion rate subsequently drops. Some REL and RLC messages are also discarded after the initial onset of congestion. A few seconds later, when the T_{IAM} and T_{REL} timers expire, the REL message load increases and the traffic load towards the STPs is further intensified. This increase in the traffic volume results in more IAM and ANM messages being discarded and consequently more REL message are generated. Customer reattempts due to call failures increase the number IAM messages generated. The increasing traffic volume reduces throughput of ANM message traffic, and hence the call completion rate decreases. The network traffic volume continues to increase until the system stabilises in the congested region.

In a scenario where none of the IAM or REL messages are successful, the maximum number of IAM messages expected in the outgoing message stream of SP region x is

$$\lambda_{IAM\,x} = \frac{\alpha_x}{(1-c)} \tag{4-13}$$

and the maximum number of REL messages is

$$\lambda_{REL\,x} = \frac{\lambda_{IAM\,x}}{(1-r)}, \tag{4-14}$$

once the system has stabilised. The system will eventually stabilise because the maximum number of calls and release attempts for each call is finite. Although the above calculation of the maximum number of messages expected overestimates the network traffic load, these equations provide a simple (but rough) method of determining the network load in a worst-case scenario.

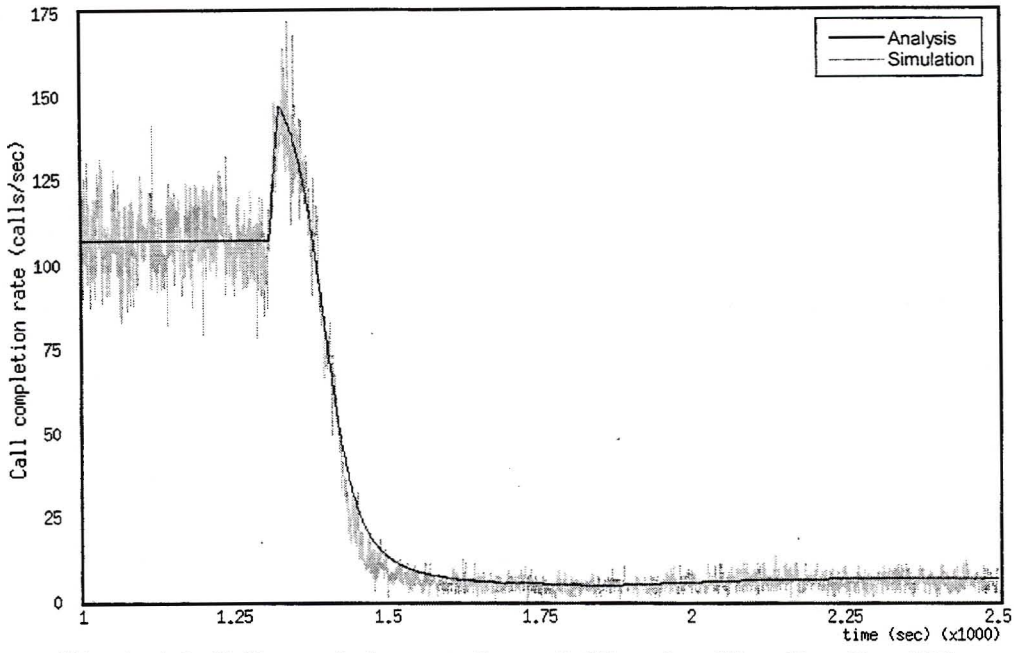


Figure 4-2. Call completion rate for each SP region ($K_1 = K_2 = K_3 = 100$).

When the overload is removed the network remains congested, even though there is a decrease in the network traffic volume. The reattempts are sufficient to keep the network in a congested state. However, a very small improvement (of less than 5 calls/sec) in the call completion rate is evident when the overload traffic is removed. The above results provide further evidence that a congestion situation is possible at first offered loads that are below the network's maximum call handling capacity.

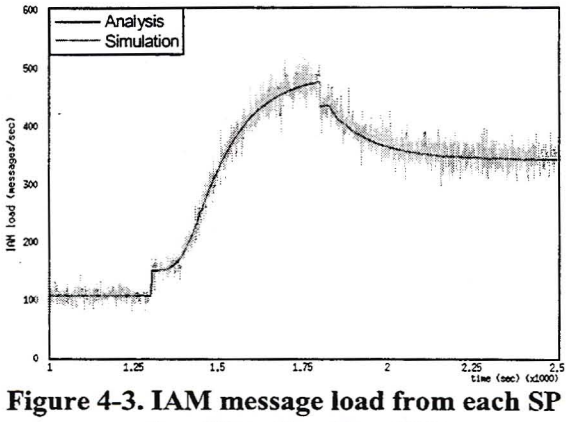


Figure 4-3. IAM message load from each SP region ($K_1 = K_2 = K_3 = 100$).

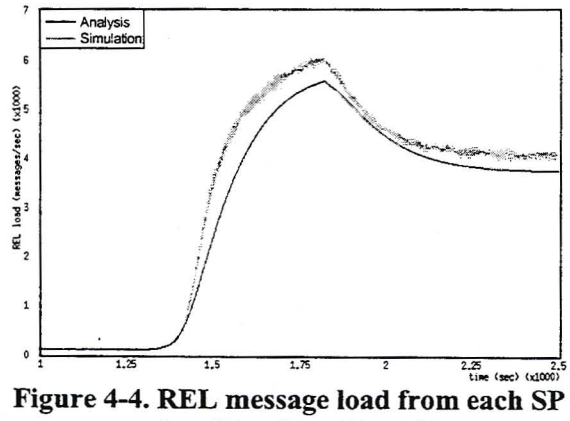


Figure 4-4. REL message load from each SP region ($K_1 = K_2 = K_3 = 100$).

Figures 4-3 and 4-4 show the IAM and REL message arrival rates into the network, from each SP region. The IAM load calculated by analysis serves as a very good approximation when compared to the simulation results. When the overload traffic is removed the number of IAM messages generated decreases and the system stabilises at a lower IAM message load. The predicted REL message load is lower than the load obtained by simulation. As mentioned in Section 3.4.1.1, this discrepancy is due to the correlated nature of the REL message stream, which is repeated every T_{REL} seconds. The REL message stream is therefore subjected to a higher effective discard probability and consequently more REL messages are discarded when a burst of messages arrive at the STPs.

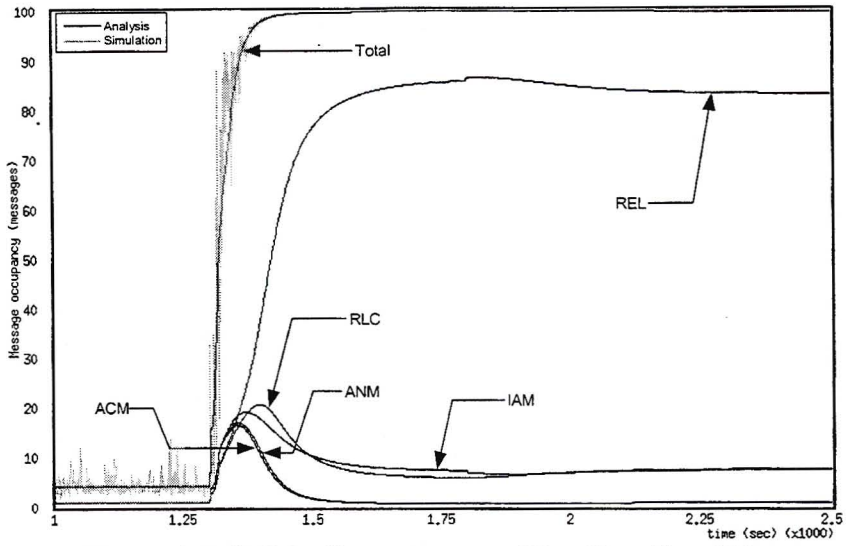


Figure 4-5. STP buffer occupancy ($K_1 = K_2 = K_3 = 100$).

Figure 4-5 shows the queue length and message occupancy of a STP buffer. When the overload is introduced the queue length begins to increase and the number of REL messages in the buffer becomes more prevalent once message discarding commences. Note, the queue length data shown in Figure 4-5 for the simulation is averaged over a one second period and even though the average queue length shown only approaches 100 at ~1400 seconds, the queue actually reaches its maximum size soon after the onset of congestion. The queue length then drops, as no subsequent ACM messages are received from the discarded IAM messages. The figure also shows that the queue length does not decrease significantly when the overload traffic is removed, but there is a small decrease in the number of REL messages present in the queue. This reduction in the number of REL messages is accompanied by a small increase in the number of other messages admitted into the queue.

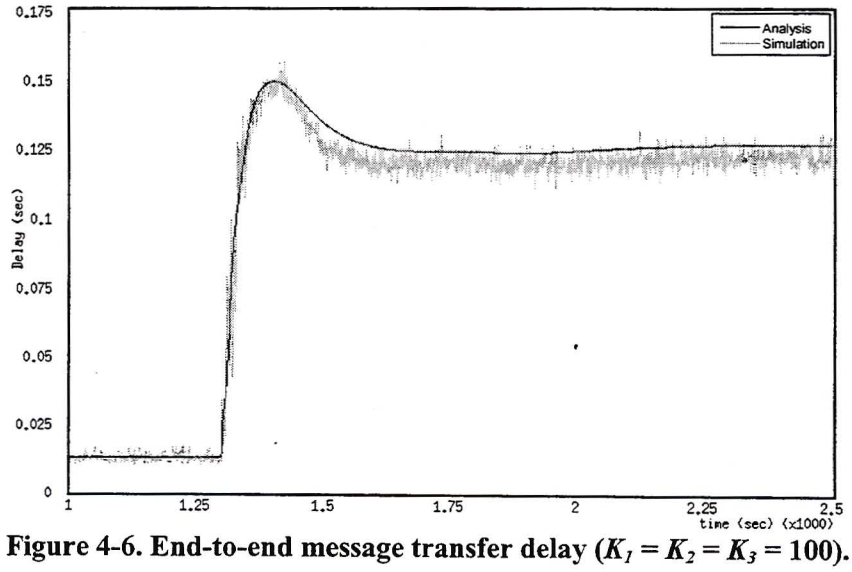


Figure 4-6. End-to-end message transfer delay ($K_1 = K_2 = K_3 = 100$).

Figure 4-6 shows the end-to-end message transfer delay of an arbitrary message. The large delay observed at the initial onset of congestion is due to the high message throughput attained to the non-adjacent regions. Eventually, as the network traffic load increases, the number of successful messages to the non-adjacent regions decreases and throughput to SPs in the same region or in the adjacent regions increases. The number of successful messages to the non-adjacent regions improves slightly when the overload traffic is removed, and hence a very small increase in the average end-to-end message transfer delay is also observed.

4.3.1.2 Multiple Discard Thresholds

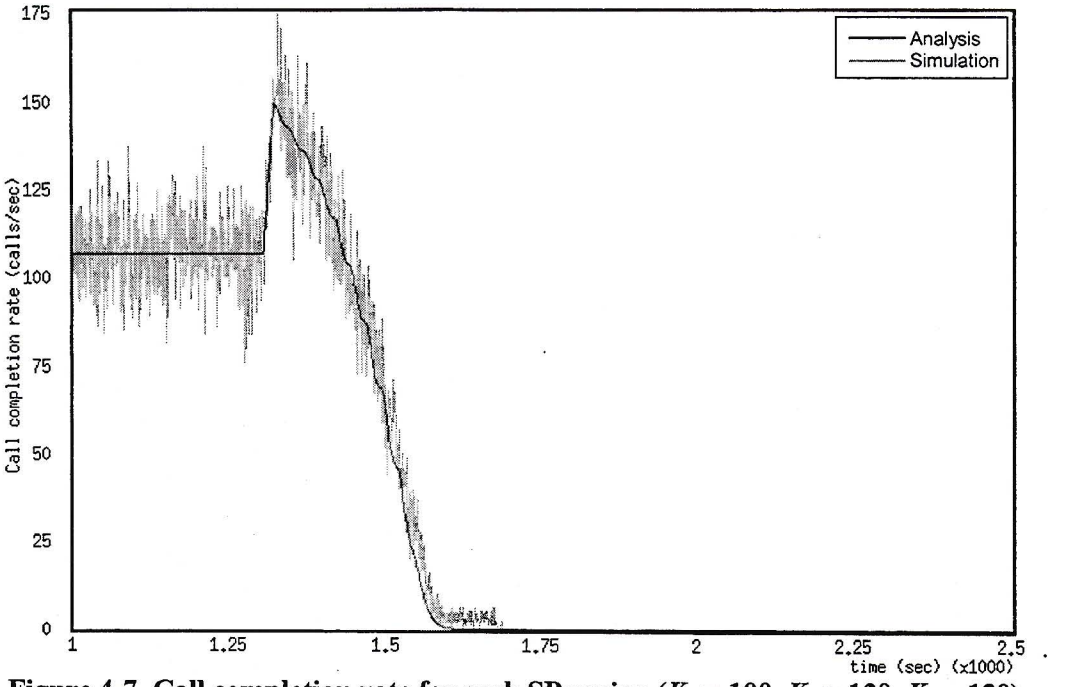


Figure 4-7. Call completion rate for each SP region ($K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

Figure 4-7 shows the call completion rate obtained for the 01212_{ISUP} message priority scheme. Unlike the single discard threshold scheme here the call completion rate takes much longer to deteriorate and therefore more calls are successfully completed during the first 300 seconds of the overload. However, the call completion rate eventually drops to zero when no further unanswered calls are still pending in the network.

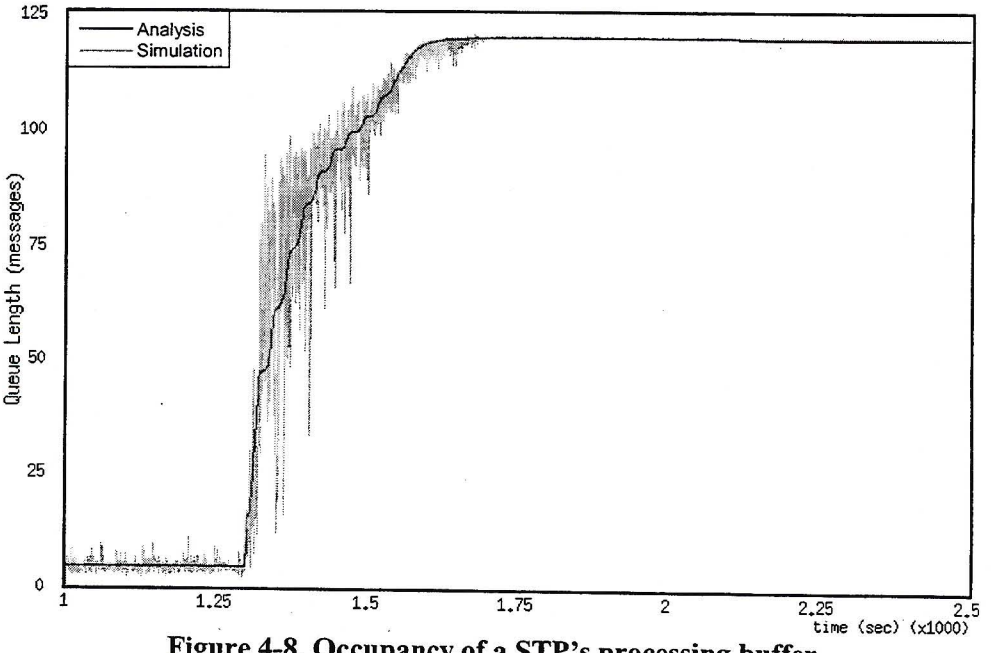


Figure 4-8. Occupancy of a STP's processing buffer ($K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

When the overload is triggered a large number of IAM messages are admitted into the network. The MTP level 3 processing buffers in the STPs rapidly grow until K_1 is attained. Soon thereafter, the successful IAM messages trigger a surge in the number of ACM messages that

are admitted into the network. These ACM messages are eventually followed by a large number of ANM messages. These messages cause the queue length of the level 3 processing buffer to increase until K_2 is reached, and as a result no further IAM messages are admitted into the queue. Consequently, the blocking of IAM messages results in a decrease in the number of ACM and ANM messages generated, and the queue length eventually drops to below K_1. IAM messages are then admitted into the queue as resources are made available by the processed messages. The IAM messages again trigger another flood of ACM and ANM messages and the queue length increases until K_2 is attained. This process is repeated a number of times, with fewer IAM messages being admitted on subsequent intervals as more buffer resources are consumed by the growing volume of REL messages. T_{IAM} seconds after the first IAM messages were discarded, a large number of REL messages were generated. Together with the REL and RLC messages from the normal call release procedure, the level 3 processing buffer is maintained above K_1 for a longer duration on subsequent T_{IAM} intervals. This process leads to more IAM messages being discarded at the STPs as time progresses and consequently the number of REL messages generated every T_{IAM} seconds also increases. The number of IAM and REL messages generated by the SPs gradually increases until the network is overwhelmed with only REL and RLC messages. The high volume of REL and RLC messages results in the STP's processing queue length remaining continuously above K_1 (but fluctuating about K_2) once the system reaches equilibrium. Figure 4-8 shows the growth in the number of messages present in a STP's processing buffer during the overload. The periodic surge in the number of REL messages every T_{IAM} seconds (followed by a subsequent increase in the IAM message load) results in a step-wise decrease in the mean call completion rate and a step-wise increase in the mean STP queue length.

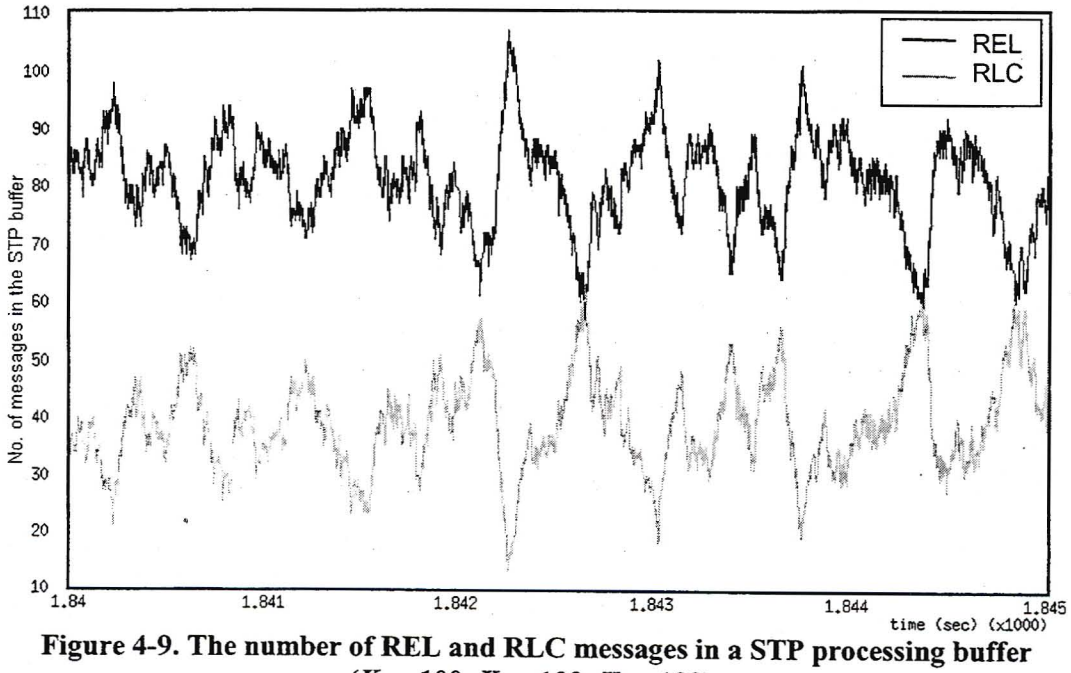


Figure 4-9. The number of REL and RLC messages in a STP processing buffer ($K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

In a large network, there are usually more REL messages than RLC messages present in the STP's queue as most of the REL messages are destined to SPs in the non-adjacent regions and of these only a few are successfully admitted into the second STP's processing buffer. The REL messages are also discarded in a periodic manner while the network is in a congested state. Initially, when a large number of REL messages, possibly due to IAM discarding T_{IAM} seconds earlier, arrive at a STP – the queue length increases to K_2. The resulting RLC messages either occupy the positions made available by the preceding REL messages or the next available position in the queue. Figure 4-9 shows the number of REL and RLC messages present in a STP processing buffer over a small period of time, from simulations. The number of REL messages

in the queue decreases as the number of RLC messages in the queue increases, while the queue length is maintained at or above K_2. The blocked REL messages are reattempted T_{REL} seconds later, which results in a periodic burst in the number of REL messages every T_{REL} seconds thereafter.

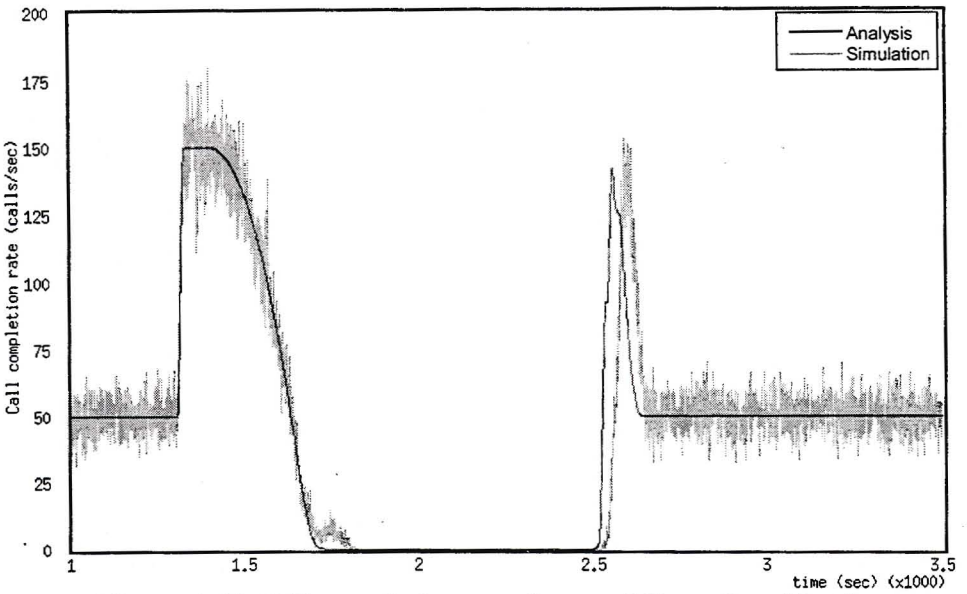


Figure 4-10. Call completion rate for small first offered loads ($K_1 = 100, K_2 = 120, K_3 = 130$).

Figure 4-10 examines a scenario where the background traffic load is small enough to allow the network to recover from the congested state. The background traffic consists of a first offered load with an arrival rate of 50 calls/sec. At time $t = 1300$ seconds the first offered load is increased to 150 calls/sec and maintained at this level for a period of 1000 seconds before the first offered load is returned to 50 calls/sec. The overload period selected is sufficiently large to allow the network to attain equilibrium at a call arrival rate of 150 calls/sec. When the first offered load is reduced, the network traffic load decreases rapidly and eventually a large peak is observed in the call completion rate before it returns to 50 calls/sec. This peak is due to the large number of backlogged calls that are finally successful. The analytical model advances sooner from the congested state than the simulation model, since the REL message load predicted by analysis is smaller than the load observed in the simulation.

4.3.1.3 Different message priority schemes

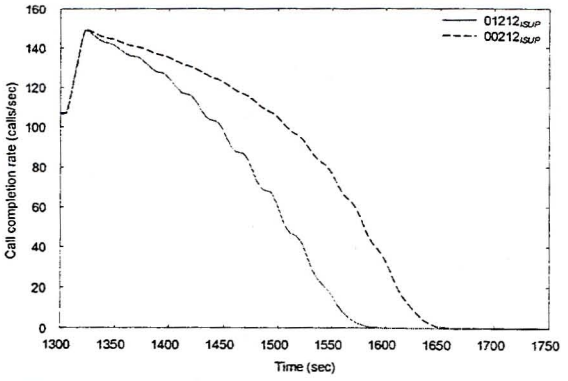


Figure 4-11. Call completion rates for the 01212_{ISUP} and 00212_{ISUP} message priority schemes.

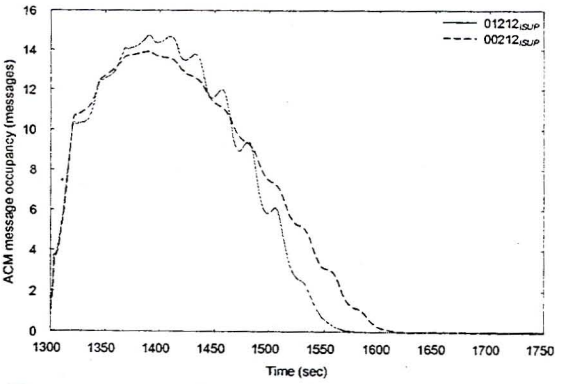


Figure 4-12. ACM message occupancy in a STP for the 01212_{ISUP} and 00212_{ISUP} message priority schemes.

The steady state equilibrium analysis in the previous chapter concluded that the 01212_{ISUP} and 00212_{ISUP} message priority schemes are the most suitable priority assignment schemes, since they have lower network traffic volumes and are able to recover from congestion at a higher first offered load. Figures 4-11 and 4-12 compare differences between the call completion rates and the ACM message occupancy of a STP buffer, for both priority assignment schemes. When the 00212_{ISUP} scheme is used, the queue length remains above K_1 for a smaller duration, as fewer high priority messages are present in the network. Conveniently, more IAM messages are admitted into the STPs and therefore more calls are successful before the call completion rate settles at zero. Fewer REL messages are also discarded when the 00212_{ISUP} priority scheme is used, as REL messages do not have to compete with ACM messages for the buffer resources between K_1 and K_2. Figure 4-12 shows that reducing the priority of the ACM messages does not severely impact the number of ACM messages present in a STP's processing buffer. However, the low priority ACM messages are present in the buffer for a longer duration than when the 01212_{ISUP} message priority scheme is used.

Figures 4-13 and 4-14 show the call completion rate and STP buffer occupancy obtained for the 10212_{ISUP} message priority scheme. The system has a background traffic load of 107 calls/sec. At $t = 1300$ seconds an overload of 43 calls/sec is introduced into the network, and overload persists for a period of 1200 seconds. The overloaded system stabilises at $t = 1800$ seconds, with a call completion rate of 148.7 calls/sec (i.e. a 99% success rate is obtained). During the overload most of the IAM and REL messages are successful and all the ANM and RLC messages are successful, at the expense of ACM message throughput. The average queue length of the STP processing buffer lies below K_1, and therefore some of the ACM messages are also successful. The system returns to its normal operating point soon after the overload traffic is removed.

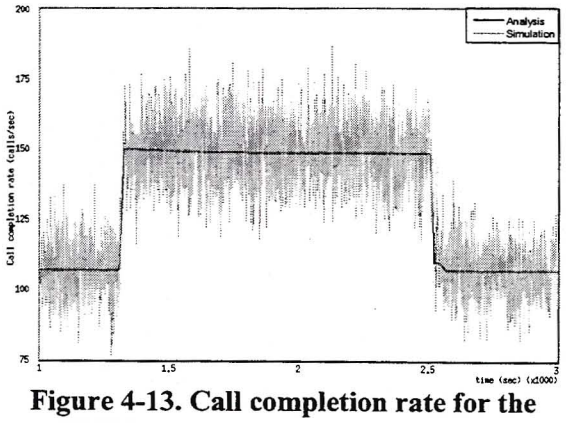


Figure 4-13. Call completion rate for the 10212_{ISUP} message priority scheme.

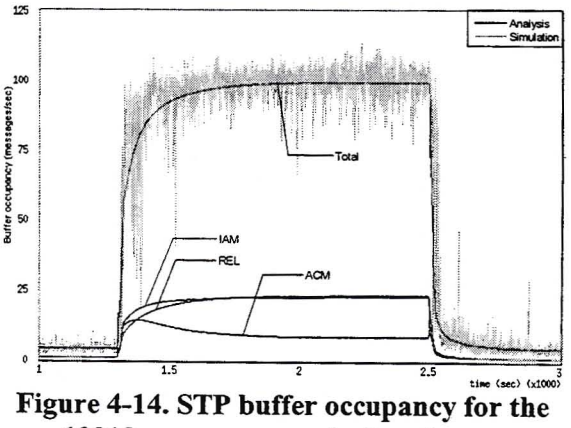


Figure 4-14. STP buffer occupancy for the 10212_{ISUP} message priority scheme.

4.3.1.4 Influence of threshold settings on network performance

Section 3.4.1.3 studied the impact of buffer threshold settings on message discarding. This section examines the impact of buffer threshold settings further, in order to investigate their influence on the call completion rate during a transient overload. The important performance measure in this section is the number of calls supported by the network before the call completion rate deteriorates to a point beyond recovery. The objective of this analysis is to also determine suitable buffer threshold settings that allow the network to operate at optimum performance.

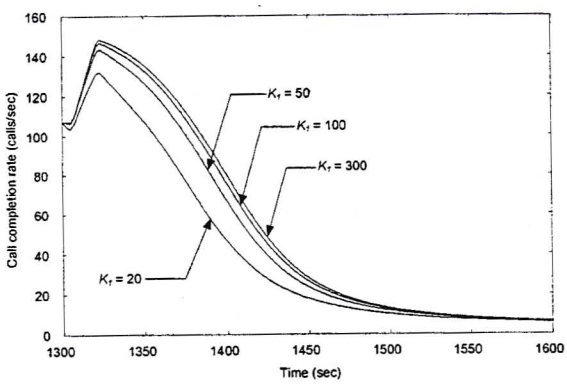


Figure 4-15. Call completion rate for various sizes of K_1, in a single priority scheme.

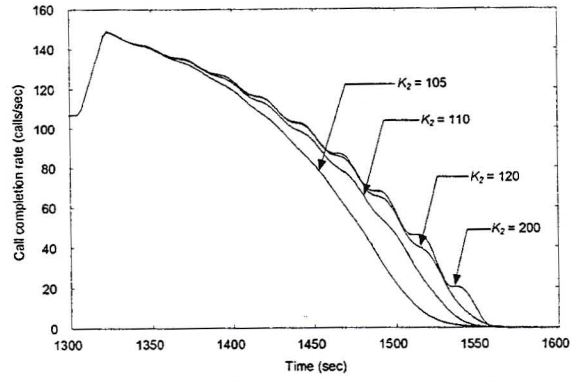


Figure 4-16. Call completion rate for various sizes of K_2, in a two priority scheme ($K_1 = 100$).

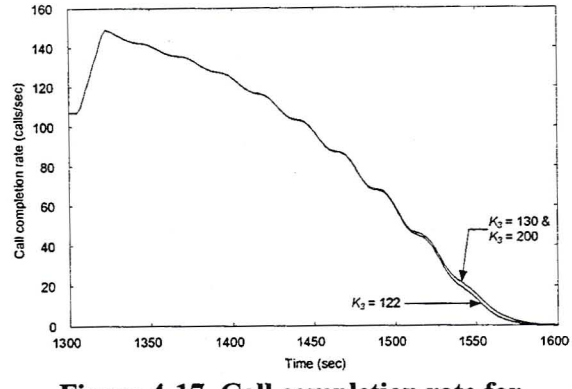


Figure 4-17. Call completion rate for various sizes of K_3, in a three priority scheme ($K_1 = 100$, $K_2 = 120$).

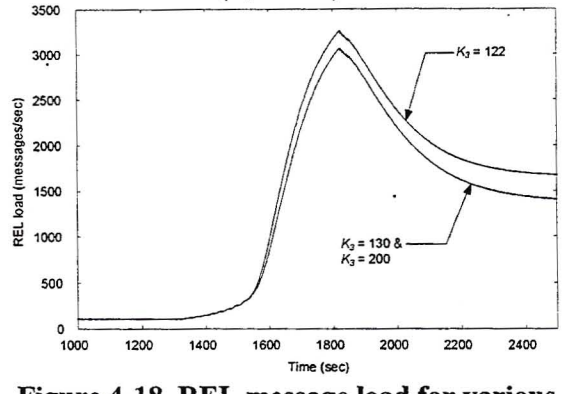


Figure 4-18. REL message load for various sizes of K_3, in a three priority scheme ($K_1 = 100$, $K_2 = 120$).

In the single message priority scheme the network is able support slightly more calls as the buffer threshold is increased, but increasing the buffer size above 100 messages does not produce a significant improvement in network performance (Figure 4-15). The very large buffers simply create excessive queuing delays with little improvement in the call completion rate or reduction in the network traffic volume. While, very small buffer sizes (e.g. $K_1 = 20$ messages) do not allow for effective utilisation of the STP processor. The small buffers are also filled to capacity within a small period of time. However, if the traffic arrival stream declines for a brief period, the queue will soon be emptied and the processor will be become idle. Small buffers are also sensitive to random fluctuations in the traffic arrival rate, and they could perceive a brief increase in the traffic arrival rate at the onset of congestion.

Figure 4-16 shows the call completion rate for a two priority (01111_{ISUP}) scheme when $K_1 = 100$ messages and K_2 is varied. In this scheme significantly more calls are successful than in the single priority scheme, before the call completion rate drops to zero. The performance improvement observed here is due to the higher throughput of ANM messages. The step-wise decrease in the call completion rate, with period T_{IAM}, becomes more prominent when very large level 2 thresholds are used. A larger value of K_2 will allow more REL messages to be admitted into the queue every T_{IAM} seconds, and therefore the queue length fluctuates less frequently about K_1. The REL messages also have a higher success rate and therefore fewer release reattempts occur. However, once the queue length drops to K_1, more IAM messages are admitted into the queue, as there are fewer REL messages to compete with. Hence, more calls are successful. Once the network has stabilised in a congested state, a large level 2 buffer threshold does not provide any additional benefit and it only increases the end-to-end message transfer delay. A level 2 buffer threshold of 120 or 130 messages is therefore suitable. In a system where the throughput of network management and status messages is important, very large queuing delays are undesirable since the time taken to relay congestion information back

to the source nodes is increased. When feedback controls are utilised, large message transfer delays could result in oscillatory traffic [Smith, 1994], as the traffic sources would be slow to respond to changes in the network state.

The performance improvement obtained with a very large level 3 buffer threshold in a 01212_{ISUP} message priority scheme is negligible (Figure 4-17). However, increasing K_3 above 122 messages results in fewer RLC messages being discarded and therefore the REL message load is reduced (Figure 4-18). Very large level 3 buffer thresholds that are greater than 130 messages are not necessary, since the queue length rarely exceeds 130 messages in the simulations and the remaining resources are largely under utilised.

4.3.1.5 The influence of call-holding times on network performance

Traffic in the signalling network is modulated by the subscriber characteristics [Bolotin, 1994]; e.g. the call holding time determines how soon after a call is established a release message will arrive. Short duration calls (e.g. to voice-mail services and busy indications) have highly correlated message streams since signalling messages arrive within quick succession of each other, while long duration calls (e.g. to internet services) are more lenient on the signalling network, but they seize more trunk circuits during overloads.

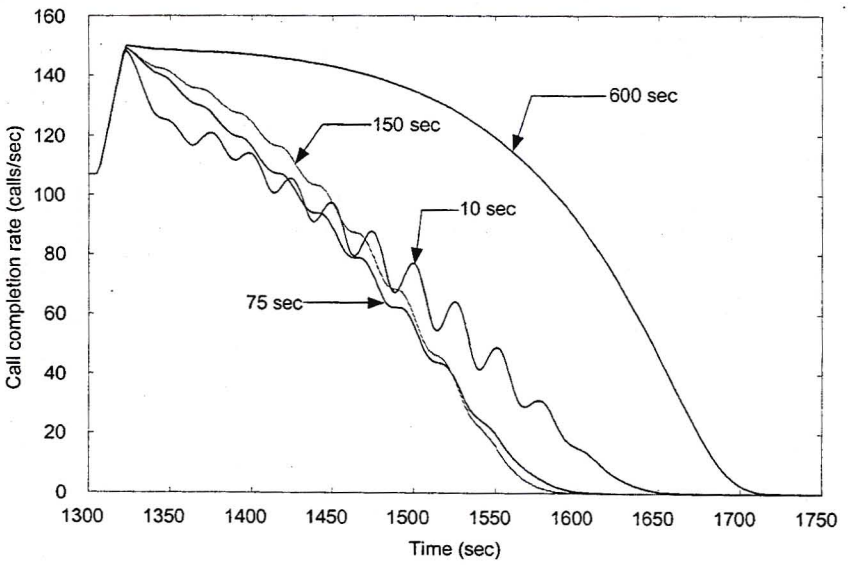


Figure 4-19. Call completion rate for different mean call holding times ($K_1 = 100, K_2 = 120, K_3 = 130$).

Figure 4-19 illustrates the influence of call holding times on the call completion rate, during a network wide overload. Very short duration calls (in the order of 15 seconds or less) generate a large number of REL messages soon after the overload is introduced, hence the call completion rate drops rapidly soon after the onset of congestion. But, since some of the REL messages were successful at the early stages of the overload, fewer REL messages are created when the network becomes congested and therefore more IAM messages are admitted into the network. The outcome of a lower REL message load and a higher IAM message load at the initial stages of congestion is a more pronounced oscillation every T_{IAM} seconds. If, in addition to the short call holding time, the call ringing time and customer reattempt period are also reduced then the oscillations become more pronounced as the message arrival streams become more bursty in nature.

In networks with longer call holding times (in the order of 75 to 300 seconds), the oscillations are smaller since fewer IAM messages are discarded at the early stages of the overload. Nevertheless, the REL load eventually grows very large once the bulk of the REL messages

from the calls that were initially successful arrive into the network. The call completion rate therefore drops to zero sooner than in the network carrying the very short duration calls.

Lastly, networks with very large call holding times (in the order of 600 seconds or longer) are able to maintain a high call completion rate for a longer period. However, network performance does deteriorate, but it is more gradual than in networks carrying calls with shorter holding times. Here the network is initially able to support all the traffic originating from the IAM, ACM and ANM message streams of the overload traffic. The call completion rate begins to decline once the REL load grows to a substantial volume that exceeds that network's message handling capability. The traffic in this network does not oscillate, since the increase in the REL message load is gradual and the STP queue lengths do not regularly fluctuate about K_1. In all the scenarios analysed, the traffic arrival rates are identical once the network reaches equilibrium.

4.3.2 *Network Failures*

This section examines a network wide overload scenario, when STPs 1 and 3 have failed. The network has a background traffic load of 50 calls/sec from each region. At t = 1300 seconds the first offered load from each region is increased to 70 calls/sec and maintained at this level for a period of 500 seconds and then returned to 50 calls/sec. The following solutions were obtained by modifying the routing matrix to consider the routes followed after the failures.

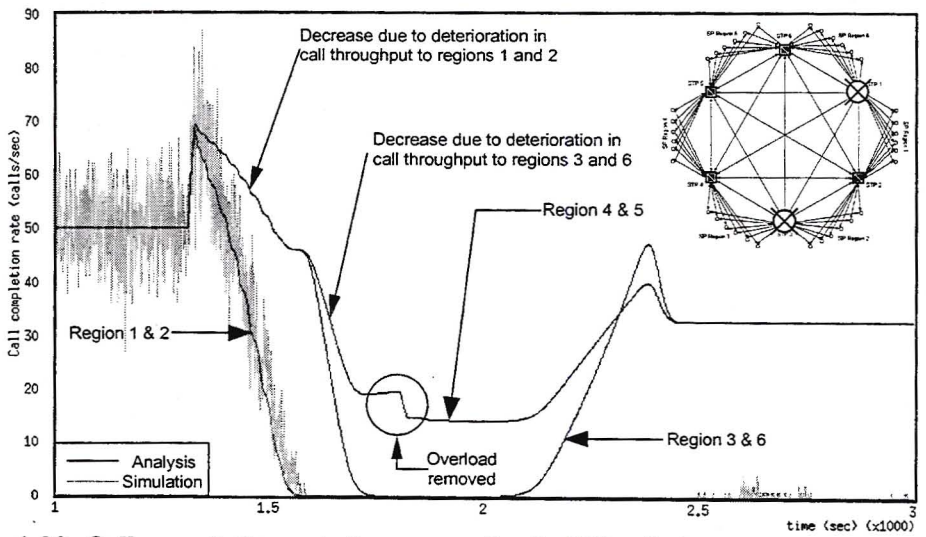


Figure 4-20. Call completion rate for an overload of 20 calls/sec with a duration of 500 seconds when STPs 1 and 3 have failed (01212_{ISUP} priority scheme with $K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

Figure 4-20 shows the impact of the overload on the call completion rate of each SP region. All the traffic to and from regions 1 and 2 is directed to STP-2, which is also the STP with the highest traffic load. STP-2 is therefore the first node in the network to become congested. Failed call attempts to regions 1 and 2 result in an increase in the REL traffic load from the other regions. Figure 4-20 also indicates where the decrease in the call completion rate is due to a deterioration in call throughput to the affected regions. All the calls between regions 3, 4, 5 and 6 are successful until the call completion rate to and from regions 1 and 2 drops to zero. Thereafter, the congestion propagates to STPs 4 and 6 and hence the call completion rates for regions 3 and 6 also drops to zero. Eventually, only calls between regions 4 and 5 are successful and half of the calls originating from SPs in regions 4 and 5 and destined to other SPs in the same region are also successful. STP-5 does not become congested as it carries the smallest traffic load. When the overload is removed, network traffic decreases and STPs 4 and 6 return to the uncongested state. The system eventually reaches equilibrium in a state where only STP-

1 is congested. In the simulation, a few calls from regions 1 and 2 are successful when the system reaches equilibrium. This indicates that the queue length of STP-2 does occasionally drift below K_1.

The above results demonstrate how localised congestion in a non-symmetrical network can propagate to other regions (STP-5 would also have become congested if a higher overload was used in the analysis). The first offered load of 70 calls/sec used to invoke congestion is also well below the network's maximum call handling capability of 130 calls/sec, when no failures are present. It is therefore possible for the entire network or part of the network to become congested in the event of one or more STP failures.

4.3.3 *Focused Overloads*

This section examines a focused overload scenario, when STPs 1 and 3 have failed. The network has a background traffic load of 45 calls/sec from each region. At $t = 1300$ seconds a focused overload of 10 calls/sec to region-1 is introduced. The overload is maintained for a period of 1200 seconds.

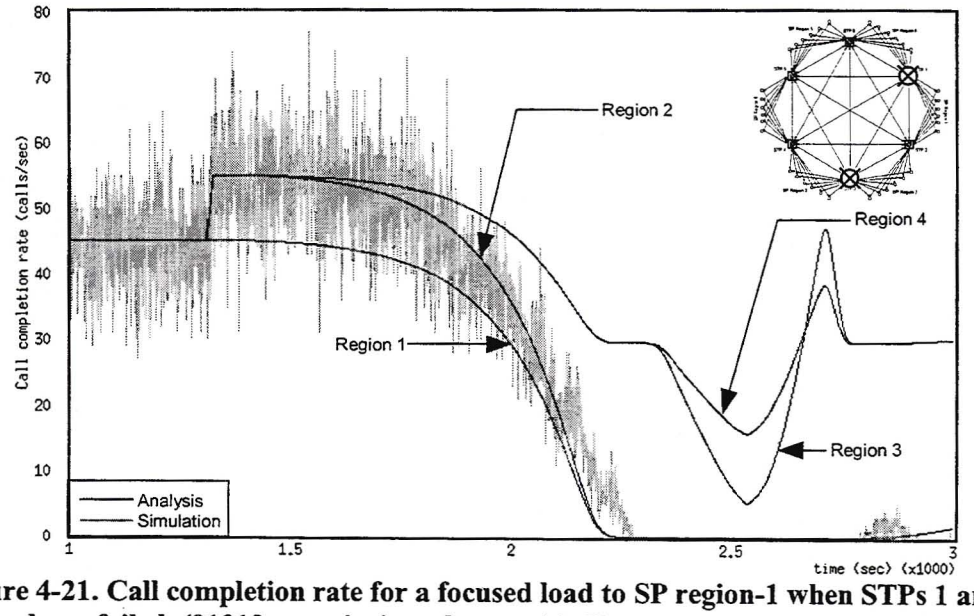


Figure 4-21. Call completion rate for a focused load to SP region-1 when STPs 1 and 3 have failed (01212_{ISUP} priority scheme with $K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

Figure 4-21 shows the impact of the overload on regions 1, 2, 3 and 4. As the overload is small, the deterioration in the call completion rates is gradual. After $t = 2300$ seconds, no calls to or from regions 1 and 2 are successful and only calls between the other regions are successful. However, the congestion eventually propagates to STPs 4 and 6, and consequently the number of successful calls to and from regions 3 and 6 decreases. At $t = 2500$ seconds the overload is removed, and the STPs 4 and 6 return to the uncongested state. Soon, thereafter the load to STP-1 decreases and some of the calls from regions 1 and 2 are eventually successful.

As in the previous section, congestion due to a focused overload can also propagate throughout the network. Propagation of congestion occurs when reattempts towards the congested region overwhelm the uncongested STPs. In this type of scenario, automatic call gapping mechanisms that limit the number of calls to high volume destinations could help to alleviate the congestion situation, and complement the actions of ISUP congestion control mechanisms.

4.3.4 *Sensitivity of the meta-stable region to changes in the system parameters*

Chapter 3 showed that three possible solutions are possible for some of the first offered loads. The previous sections also showed that it is possible for a network to remain congested and in equilibrium even after the overload traffic is removed. This section demonstrates the sensitivity of a network operating in the meta-stable region to very small changes in the system parameters.

A single STP network is examined in this section, since fewer system parameters are required. In a single STP network, all the SPs belong to a single SP region and therefore references to the message arrival rate at the STP only relate to messages from one region. Routing is also simplified since all the traffic is directed through a single STP. The other parameters in the system are identical to those used in Section 4.3.1 (e.g. STP's message processing rate, call holding time, etc.). It should be noted that a single STP network with these parameters has a call handling capability of ~200 calls/sec. When the single discard threshold scheme is used with K_1 = 100 messages, the region with three solutions extends from a first offered load of ~124 calls/sec to ~200 calls/sec. A detailed analysis of a single STP network is given in [Rumsewicz, 1993]. However, Rumsewicz confirms but does not present results to illustrate the sensitivity of the meta-stable region to small changes in the system parameters.

To demonstrate equilibrium in the meta-stable region one has to be able accurately represent the message arrival rates at the STP. For a message arrival rate of 2000 messages/sec to the STP, the probability of discard is 0.5 (a rational number). Likewise, one is able to calculate and represent the arrival rates of the individual message streams with rational numbers. The first offered load in this case is 138.1818 (or 5525/8) calls/sec. This precise representation of the message arrival rates and the probability of discard are necessary to accurately specify the initial conditions of a system in the meta-stable region. The sensitivity of network operation in the meta-stable region to very small errors in the initial parameters will become apparent in the following results.

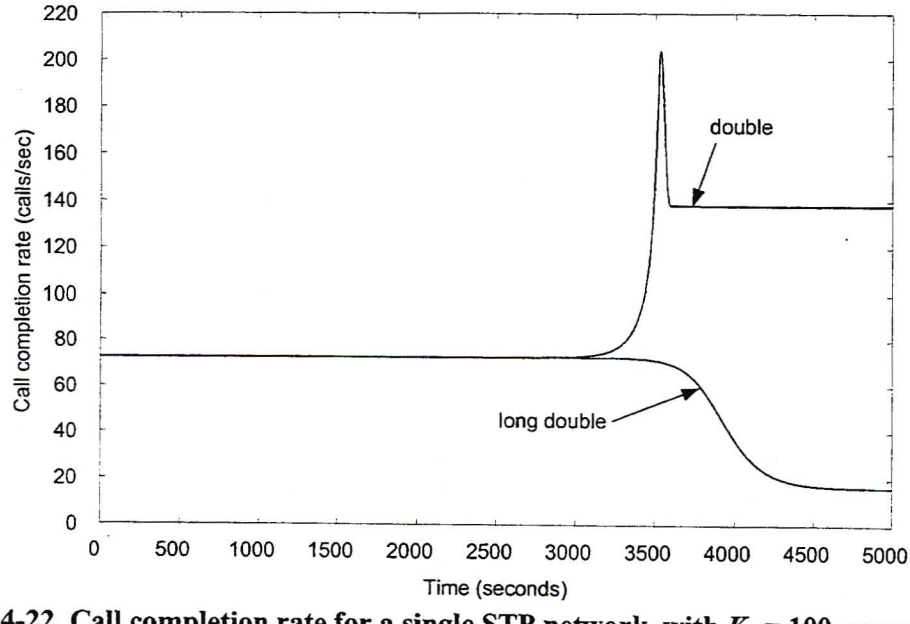


Figure 4-22. Call completion rate for a single STP network, with K_1 = 100, commencing from the meta-stable region. Different call completion rates are obtained depending on the type of variables used in the program code.

The computer programs used to solve the system of equations for the transient analysis were written in ANSI C code and compiled with gcc (version 2.7.2.1) on a Pentium II computer with

a Linux (kernel version 2.0.32) operating system. Note; computers are only able to represent floating-point numbers (i.e. real numbers) with a finite precision. The floating-point representation of the message arrival rates, in a computer, is therefore not 100% accurate. As a result, computations that iteratively solve a system of equations are susceptible to produce large errors if small errors exist in the initial parameters.

Two computer programs were written to investigate equilibrium in the meta-stable region. In the first program all the floating-point variables were defined as type *double*[1] and in the second program all the floating-point variables were defined as type *long double*[2]. Figure 4-22 shows the call completion rates obtained from both programs. In both cases the message arrival rates were accurate to within 10^{-13} at time $t = 0$ seconds. A much higher precision was not possible with the *long double* variables since some of the ANSI C mathematical functions are only able to return values of type *double*. Since the state of the system at some point in time depends on previous events in the system, the network's operating point eventually drifts towards either the uncongested region or the congested region. The state in which the network eventually settles appears to depend on the total message arrival rate at time $t = 0$ seconds. The total message arrival rate is not exactly 2000 messages/sec due to floating-point errors acquired in the representation of the individual message streams. If the message arrival rate to the STP is greater than 2000 messages/sec the network settles in the congested region or else it drifts towards the uncongested region. In a real-world network or a simulation, a network operating in the meta-stable region would quickly drift towards one of the stable regions, as the message arrival rates are not constant but random.

4.4 Summary

Most analytical models that examine the performance of packet switched networks concentrate on the system's performance during steady state conditions, since steady state models are easier to abstract and solve. The transient analytical models require explicit consideration of the underlying stochastic processes and are usually more difficult to solve with standard numerical analysis techniques. The few publications that have developed transient analytical models for SS7 networks, concentrate on the time varying performance of the congested entity. These models cannot be used to investigate how congestion manifests itself in a larger network environment, nor can they be used to examine how congestion propagates throughout a network.

The transient model developed in this chapter complements the steady state equilibrium model developed in the previous chapter. The results presented here are also able to provide an intuitive perspective on the realisation of sustained congestion after brief periods of overload. Like the previous chapter, the model developed here also considers the network configuration, routing and allows for the analysis of different message priority schemes, failure scenarios and different traffic loads from each region.

The results obtained show that it is possible for a network to become congested after a few hundred seconds of overload and thereafter remain in a congested state. The queue lengths of the processing buffers in the congested STPs are filled to their maximum capacity (or the level 1 discard threshold in the multiple discard threshold scheme) immediately after the overload traffic is introduced. Message discarding in the congested STPs leads to message reattempts and call reattempts a few seconds later. These reattempts aggravate congestion at the affected STPs and eventually throughput of the initial call set-up messages suffers. Consequently the call

[1] A variable of type *double* is stored as a 64-bit binary number and has 15 decimal digits of precision.
[2] A variable of type *long double* is stored as a 96-bit binary number and has 18 decimal digits of precision.

completion rate also deteriorates. Comparisons between the analytical results and simulation results show that the analysis provides a good approximation of the network's behaviour.

When the 01212_{ISUP} message priority scheme is used the rate of decrease in the call completion rate is lower than that observed in the single discard threshold scheme. More calls are therefore successful during the first 300 seconds of the overload, before the call completion rate drops to zero. A higher call completion rate is obtained since ANM messages that originate from IAM messages that were previously successful experience minimal discarding within the network. Once congested, the queue lengths in the affected STPs remain continuously above the level one threshold and therefore no IAM messages are admitted into the affected queues. In the 01212_{ISUP} message priority scheme the periodic surge in the REL message load every T_{IAM} seconds results in a step-wise decrease in the mean call completion rate and a corresponding step-wise increase of the mean queue length in the STPs. This burst in the REL message load is much larger and more prominent when very small call holding times are used. However when very large call holding times are present, the network is able to sustain more calls before the call completion rate drops to zero.

An investigation of whether buffer threshold settings can improve network performance in the early stages of congestion indicates that very large buffer sizes do not significantly increase the call completion rate nor do they significantly reduce the rate at which the call completion rate deteriorates. However, large buffers can negatively impact network performance by increasing the end-to-end message transfer delay during overloads.

In networks with unbalanced traffic loads, node failures or focused overloads, congestion commences in the STP that is carrying the largest traffic load. Once congested, a large number of release attempts (due to call failures) are initiated. These REL messages aggravate congestion at the already congested STP, but they also increase the traffic load carried by the intermediate STPs. Congestion therefore propagates throughout the network when the REL load to the congested region overwhelms the uncongested STPs.

5. Steady State Equilibrium Analysis of SS7 Network Performance in the PLMN

5.1 Introduction

Previous research on SS7 network performance has concentrated on ISUP signalling [Smith, 1994] and intelligent network transactions in a PSTN environment [Zepf & Rufa, 1994]. While, studies on signalling in mobile networks have generally used simple techniques to estimate the increase in signalling load due to mobility management functions (e.g. [Pollini et al., 1996] and [Tabbane, 1998]) or have represented the signalling load associated with various transactions with a cost value [Jain & Lin, 1995]. The major shortcoming of these studies is their inability to provide methods of determining the effect of signalling network congestion and failures on application level protocols and the effect of this congestion on the quality of service provided to the customer.

In Section 5.2 analytical models are developed to examine the influence of GSM mobility management protocols on signalling load and network performance during STP congestion. Equations are derived for the GSM MAP Location Management protocol, the GSM MAP Call Delivery protocol and various other mobility management protocols that have been proposed for 3G networks. Mobile network performance is determined by evaluating the location update success rate, the HLRs' database accuracy, the location cancellation success rate and the effectiveness of the MAP call delivery procedure. In addition, the equations derived in Chapter 3 are used to evaluate the ISUP call completion rate.

5.2 Steady State Equilibrium Analysis

The mobility management procedures described in Section 1.8 and illustrated in Figures 1-12 and 1-13 are used to calculate the traffic load attributed to each MAP message. The following assumptions are made in the analysis:

- As in present day implementations, each MSC has an integrated VLR. Each MSC/VLR controls only one location area.
- Final global title translation is assumed to be performed by the signalling points and not the STPs for messages between nodes within a mobile network. This is equivalent to the approach used by a number of mobile operators. Performing the global title translation and relay functions on the STPs is known to impede the relaying of end-to-end congestion information ([ITU-T Recommendation Q.715] and [Zepf & Rufa, 1994]).
- Location updates are due to subscriber mobility and are initiated when a mobile terminal enters the radio coverage of a new location area.
- The occurrence of periodic location updates is assumed to be negligible, since their frequency is configurable and the time period between subsequent periodic location updates is normally very large to conserve the scarce radio resources.
- To simplify the analysis the load attributed to inter-MSC handovers is assumed to be negligible. Intra-MSC handovers do not generate signalling traffic between MSCs.
- The signalling load attributed to GSM authentication procedures between the VLR and HLR is assumed to be negligible or zero, since the frequency of authentication and the lifetime of a ciphering key are configurable parameters, or authentication could be deactivated. Furthermore, if authentication is active each authentication procedure results in the transfer of multiple authentication vectors from the HLR to the VLR. The additional vectors are cached in the VLR. The frequency of authentication vector transfers from a HLR to a VLR relative to the number of location updates and calls made by a subscriber is

normally very small to conserve the scarce radio resources and to reduce signalling load.

The analysis below does not explicitly model the subscriber population densities within each location area when defining the mobility matrix. However, the analysis does not restrict one from using the gravity mobility model described by Lam et al. [1997] and Wong & Leung [2000] when defining the mobility matrix. This model is often used to analyse aggregate movement behaviour, typical of a scenario where a large number of subscribers are migrating from one location area to another. The amount of subscribers, $T_{i,j}$, moving from region i to region j is described by

$$T_{i,j} = K_{i,j} P_i P_j ,$$

where P_i is the population in region i, and $\{K_{i,j}\}$ are parameters that have to be determined for all possible regional pairs (i,j) [Wong & Leung, 2000]. $T_{i,j}$ can then be used to determine the number of location updates triggered in region i, due to subscribers leaving region j.

The notation used in this section is identical to that used in Chapter 3, but with the following modifications and additions:

i) α is a $M \times M \times M$ matrix of new call traffic where α_{ijk} is the number of new call arrivals per second in region i that are destined to subscribers who are roaming in region j, where the called subscribers have their profiles stored on a HLR in region k. The first offered load attributed to new call attempts from region i is therefore given by

$$\alpha_i = \sum_{j=1}^{M} \sum_{k=1}^{M} \alpha_{ijk} .$$

ii) α' is a $M \times M \times M$ matrix of new call traffic where α'_{ijk} is the number of new call arrivals per second including reattempts, where i denotes the originating region of the calls, j denotes the terminating region and k is the region in which the HLR of the called subscribers is located.

iii) β is a $M \times M \times M$ mobility matrix where β_{ijk} is the number of new location update attempts per second in region i, due to subscribers who were last registered on a VLR in region j and have their profiles stored on a HLR in region k. The first offered load attributed to new location update attempts from region i is given by

$$\beta_i = \sum_{j=1}^{M} \sum_{k=1}^{M} \beta_{ijk} .$$

iv) d denotes the number ISD messages required to transfer a subscriber's entire profile from a HLR to a VLR.

v) g is the maximum number of location update reattempts allowed. The attempt counter referred to in Section 1.8.1 is equal to $(g + 1)$. $g = 3$ for the GSM and UMTS circuit switched network and $g = 4$ for GPRS services.

vi) +ULA is used to denote ULA messages generated due to successfully completed insert subscriber data procedures. –ULA is used to denote ULA messages generated due to a timeout of the insert subscriber data procedures.

vii) +SRIA is used to denote SRIA messages containing a roaming number. –SRIA is used to denote SRIA messages generated due to a timeout of the provide roaming number procedures.

viii) SUL and SULA are used to denote the UL and ULA messages that are transferred in a Super-Charged network for a location update that does not require the transfer of a subscriber's profile.

ix) δ denotes the probability that a location update will require the transfer of a subscriber's profile from the HLR to the VLR in a Super-Charged network.

5.2.1 GSM Location Management

The total number of UL messages generated is equal to the total number of new location update (LU) attempts plus reattempts due to failures. The number of UL messages from a VLR in a region x to a HLR in region y is therefore given by the following:

$$\lambda_{ULxy} = \text{new LU attempts} + \text{1st LU reattempts} + \text{2nd LU reattempts} + \text{3rd LU reattempts} + \ldots \quad \textbf{(5-1)}$$

where

$$\text{new LU attempts} = \sum_{m=1}^{M} \beta_{xmy}$$

$$\text{1st LU reattempts} = (\text{new LU attempts}) \cdot X = \sum_{m=1}^{M} \beta_{xmy} X$$

$$\text{2nd LU reattempts} = (\text{1st LU reattempts}) \cdot X = \sum_{m=1}^{M} \beta_{xmy} X^2, \quad \text{and}$$

$$\text{3rd LU reattempts} = (\text{2nd LU reattempts}) \cdot X = \sum_{m=1}^{M} \beta_{xmy} X^3.$$

In the above expressions the number of location updates generated due to reattempts is equal to the number of location updates from the previous attempts multiplied by the probability (X) that the location updates were unsuccessful.

Equation (5-1) can be rewritten as

$$\lambda_{ULxy} = \sum_{m=1}^{M} \beta_{xmy} \sum_{n=0}^{g} X^n \qquad \textbf{(5-2)}$$

where the summation of terms from $n = 0$ to g is used to determine the number of times failed location updates may be reattempted. To determine the probability of a reattempt (X), the probability that a UL, ISD, ISDA or ULA message may be discarded along each route between the VLRs' region and the HLRs' region is calculated in (5-3). Each term within the round brackets of equation (5-3) calculates the following:

- the probability that a UL message may be discarded at a node along each route between the VLR and HLR,
- the probability that the UL message was successful multiplied by the probability that the 1st, 2nd, 3rd ... or dth ISD message may be discarded between the HLR and VLR,
- the probability that the UL message and all the ISD messages were successful multiplied by the probability that the 1st, 2nd, 3rd ... or dth ISDA message may be discarded between the VLR and HLR, and
- the probability that UL message and messages of the insert subscriber data procedure were successful multiplied by the probability that the resulting +ULA message may be discarded at a node along each route between the HLR and VLR.

Within equation (5-3) it is only necessary to evaluate the probability of one of the ISD or ISDA messages being discarded to consider the insert subscriber data procedure as being unsuccessful, while the procedure is only successful if all the ISD and ISDA messages (i.e. d messages) are successfully transferred between the VLR and HLR.

$$
\begin{aligned}
X = \sum_{i=1}^{R_{xy}} \Bigg(& \sum_{j=1}^{Q_{F\,xyi}} \frac{1}{R_{xy}} P_D(\phi_{UL},0,x,y,i,j) \\
& + \sum_{j=1}^{Q_{R\,xyi}} \frac{1}{R_{xy}} P_S(\phi_{UL},0,x,y,i,Q_{F\,xyi}) \sum_{k=1}^{d} P_S(\phi_{ISD},1,x,y,i,Q_{R\,xyi})^{k-1} P_D(\phi_{ISD},1,x,y,i,j) \\
& + \sum_{j=1}^{Q_{F\,xyi}} \frac{1}{R_{xy}} P_S(\phi_{UL},0,x,y,i,Q_{F\,xyi}) P_S(\phi_{ISD},1,x,y,i,Q_{R\,xyi})^{d} \sum_{k=1}^{d} P_S(\phi_{ISDA},0,x,y,i,Q_{F\,xyi})^{k-1} P_D(\phi_{ISDA},0,x,y,i,j) \\
& + \sum_{j=1}^{Q_{R\,xyi}} \frac{1}{R_{xy}} P_S(\phi_{UL},0,x,y,i,Q_{F\,xyi}) P_S(\phi_{ISD},1,x,y,i,Q_{R\,xyi})^{d} P_S(\phi_{ISDA},0,x,y,i,Q_{F\,xyi})^{d} P_D(\phi_{ULA},1,x,y,i,j) \Bigg)
\end{aligned}
\tag{5-3}
$$

The UL traffic along each route, z, from region x to region y is thus given by

$$\lambda_{UL\,xyz} = (1/R_{xy})\lambda_{UL\,xy}, \qquad \text{where } 1 \le z \le R_{xy}.$$

For every UL message that arrives at a HLR, d ISD messages are sent to the VLR that originated the UL message. Let $z = \Gamma(y,x,i)$ where $1 \le i \le R_{xy}$, then

$$\lambda_{ISDxyz} = \frac{d}{R_{yx}} \lambda_{ULyx} P_S(\phi_{UL},0,y,x,i,Q_{F\,yxi}). \tag{5-4}$$

Each ISD message that is received by a VLR, generates an ISDA message to the concerned HLR. The rate at which ISDA messages are transmitted over each route is given by

$$\lambda_{ISDAxyz} = \lambda_{ISDyxi} P_S(\phi_{ISD},0,y,x,i,Q_{F\,yxi}) \tag{5-5}$$

where $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

If all the ISD messages for a particular subscriber are successfully delivered to the VLR and all the resulting ISDA messages arrive at the HLR then the rate at which +ULA messages are transmitted over each route is given by

$$\lambda_{+ULAxyz} = \frac{\lambda_{ISDxyz}}{d} P_S(\phi_{ISD},0,x,y,z,Q_{F\,xyz})^{d} P_S(\phi_{ISDA},0,y,x,i,Q_{F\,yxi})^{d} \tag{5-6}$$

where $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

If a congested STP discards only one ISD or ISDA messages then the insert subscriber data procedure is unsuccessful and the resulting ULA message contains a *user error* parameter. The rate at which –ULA messages are transmitted over each route is given by

$$
\begin{aligned}
\lambda_{-ULAxyz} = & \frac{\lambda_{ISDxyz}}{d} \sum_{j=1}^{Q_{F\,xyz}} \left(\sum_{k=1}^{d} P_S(\phi_{ISD},0,x,y,z,Q_{F\,xyz})^{k-1} P_D(\phi_{ISD},0,x,y,z,j) \right) \\
& + \frac{\lambda_{ISDAyxi}}{d} \sum_{j=1}^{Q_{F\,yxi}} \left(\sum_{k=1}^{d} P_S(\phi_{ISDA},0,y,x,i,Q_{F\,yxi})^{k-1} P_D(\phi_{ISDA},0,y,x,i,j) \right)
\end{aligned}
\tag{5-7}
$$

where $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

When the first UL message of a location update arrives at a HLR, the HLR sends a CL message to the previous VLR address that is stored in the subscriber's profile and then replaces it with the subscriber's new VLR address. If the location update procedure fails and another UL message from a reattempt arrives at the HLR then no CL message is generated since the VLR address in the UL message is now identical to the VLR address stored in the subscriber's profile. The rate at which CL messages are generated from HLRs in region x to VLRs in region y is therefore given by

$$\lambda_{CLxy} = \sum_{m=1}^{M} \beta_{myx} \sum_{n=0}^{g} \left[\sum_{i=1}^{R_{mx}} \left(\sum_{j=1}^{Q_{Fmxi}} \frac{1}{R_{mx}} P_D(\phi_{UL},0,m,x,i,j) \right) \right]^n \sum_{i=1}^{R_{mx}} \left(\frac{1}{R_{mx}} P_S(\phi_{UL},0,m,x,i,Q_{Fmxi}) \right) \quad \textbf{(5-8)}$$

In the above equation the mobility matrix is used to determine the rate at which subscribers' whose profiles are stored on HLRs in region x leave a VLR in region y. This value is multiplied by the probability that the first UL message or at least one of the subsequent reattempts is successfully received by a HLR in region x.

The CL traffic along each route, z, from region x to region y is thus given by

$$\lambda_{CL\,xyz} = (1/R_{xy})\lambda_{CL\,xy}, \qquad \text{where } 1 \le z \le R_{xy}.$$

A CLA message is generated in response to every CL message received by a VLR. The rate at which CLA messages are transmitted over each route is therefore given by

$$\lambda_{CLAxyz} = \frac{1}{R_{yx}} \lambda_{CLyx} P_S(\phi_{CL},0,y,x,i,Q_{Fyxi}) \quad \textbf{(5-9)}$$

where $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

The location update success rate for region x, $S_{LU\,x}$, is obtained by calculating the total number of +ULA messages that have been successfully received:

$$S_{LUx} = \sum_{i=1}^{M} \sum_{j=1}^{R_{ix}} \lambda_{+ULAixj} P_S(\phi_{ULA},0,i,x,j,Q_{Fixj}) \quad \textbf{(5-10)}$$

The location update success rate provides a measure of the level of service provided to a customer. From the customer's perspective it determines if the customer is or isn't allowed to roam on the network. The effectiveness of the cancel location procedure, $S_{LC\,x}$, is calculated from the following equation where the probability that CL messages are successfully delivered to VLRs in region x is calculated.

$$S_{LCx} = \sum_{i=1}^{M} \sum_{j=1}^{R_{ix}} \lambda_{CLixj} P_S(\phi_{CL},0,i,x,j,Q_{Fixj}) \quad \textbf{(5-11)}$$

Note, the success rate of CLA messages is not used because the subscriber's profile is deleted from the VLR irrespective of whether or not the CLA message is successfully delivered to the HLR.

As mentioned above, the HLR is updated and a CL message is generated when the first UL message arrives at the HLR. The accuracy of HLRs in region x, $S_{HLR\,x}$, is therefore given by the following

$$S_{HLRx} = \sum_{y=1}^{M}\sum_{m=1}^{M}\beta_{myx}\sum_{n=0}^{3}\left[\sum_{i=1}^{R_{mx}}\left(\sum_{j=1}^{Q_{F\,mxi}}\frac{1}{R_{mx}}P_D(\phi_{UL},0,m,x,i,j)\right)\right]^n\sum_{i=1}^{R_{mx}}\left(\frac{1}{R_{mx}}P_S(\phi_{UL},0,m,x,i,Q_{F\,mxi})\right) \quad \textbf{(5-12)}$$

5.2.2 *GSM Call Delivery*

When a call arrives at a MSC, the ISUP call set-up procedure is preceded by the MAP call delivery procedures to first locate the called party's MSC and to obtain a roaming number that can then be used to route the call. The total number of SRI queries generated is equal to the total number of new call arrivals plus call reattempts due to failures. The expression for the number of calls including reattempts from a MSC in region x to a MSC in region m where the called party's HLR resides in region y is written as follows:

$$\alpha'_{xmy} = \text{new call arrivals} + \text{first call reattempts} + \text{second call reattempts} + \ldots \quad \textbf{(5-13)}$$

where

$$\text{new call arrivals} = \alpha_{xmy},$$

$$\text{first call reattempts} = \alpha_{xmy}\cdot\left[\sum_{i=1}^{R_{xy}}\left(\sum_{j=1}^{Q_{F\,xyi}}\frac{1}{R_{xy}}P_D(\phi_{SRI},0,x,y,i,j)\right)+\sum_{i=1}^{R_{xy}}\left(\frac{1}{R_{xy}}P_S(\phi_{SRI},0,x,y,i,Q_{F\,xyi})\right)\cdot Y_1\right.$$

$$+\sum_{i=1}^{R_{xy}}\left(\sum_{j=1}^{Q_{R\,xyi}}\frac{1}{R_{xy}}P_S(\phi_{SRI},0,x,y,i,Q_{F\,xyi})P_D(\phi_{SRIA},1,x,y,i,j)\right)\cdot Y_2$$

$$\left.+Y_3\cdot\left(\sum_{i=1}^{R_{xm}}\left(\sum_{j=1}^{Q_{F\,xmi}}\frac{1}{R_{xm}}P_D(\phi_{IAM},0,x,m,i,j)\right)+\sum_{i=1}^{R_{xm}}\left(\sum_{j=1}^{Q_{R\,xmi}}\frac{1}{R_{xm}}P_S(\phi_{IAM},0,x,m,i,Q_{F\,xmi})P_D(\phi_{ANM},1,x,m,i,j)\right)\right)\right]\cdot c,$$

$$Y_1 = \sum_{i=1}^{R_{ym}}\left(\sum_{j=1}^{Q_{F\,ymi}}\frac{1}{R_{ym}}P_D(\phi_{PRN},0,y,m,i,j)+\sum_{j=1}^{Q_{R\,ymi}}\frac{1}{R_{ym}}P_S(\phi_{PRN},0,y,m,i,Q_{F\,ymi})P_D(\phi_{PRNA},1,y,m,i,j)\right),$$

$$Y_2 = \sum_{i=1}^{R_{ym}}\left(\frac{1}{R_{ym}}P_S(\phi_{PRN},0,y,m,i,Q_{F\,ymi})P_S(\phi_{PRNA},1,y,m,i,Q_{R\,ymi})\right), \quad \text{and}$$

$$Y_3 = \sum_{i=1}^{R_{xy}}\left(\frac{1}{R_{xy}}P_S(\phi_{SRI},0,x,y,i,Q_{F\,xyi})P_S(\phi_{SRIA},1,x,y,i,Q_{R\,xyi})\right)\cdot Y_2.$$

The probability of a reattempt in the above equations is calculated by determining the probability that an earlier SRI, PRN, PRNA, +SRIA, IAM or ANM message was discarded.

The geometric series in (5-13) can be rewritten as

$$\alpha'_{xmy} = \alpha_{xmy} \cdot \left[1 - \left(\sum_{i=1}^{R_{xy}} \left(\sum_{j=1}^{Q_{F\,xyi}} \frac{1}{R_{xy}} P_D(\phi_{SRI},0,x,y,i,j) \right) + \sum_{i=1}^{R_{xy}} \left(\frac{1}{R_{xy}} P_S(\phi_{SRI},0,x,y,i,Q_{F\,xyi}) \right) \cdot Y_1 \right.\right.$$
$$+ \sum_{i=1}^{R_{xy}} \left(\sum_{j=1}^{Q_{R\,xyi}} \frac{1}{R_{xy}} P_S(\phi_{SRI},0,x,y,i,Q_{F\,xyi}) P_D(\phi_{SRIA},1,x,y,i,j) \right) \cdot Y_2$$
$$\left.\left. + Y_3 \cdot \left(\sum_{i=1}^{R_{xm}} \left(\sum_{j=1}^{Q_{F\,xmi}} \frac{1}{R_{xm}} P_D(\phi_{IAM},0,x,m,i,j) \right) + \sum_{i=1}^{R_{xm}} \left(\sum_{j=1}^{Q_{R\,xmi}} \frac{1}{R_{xm}} P_S(\phi_{IAM},0,x,m,i,Q_{F\,xmi}) P_D(\phi_{ANM},1,x,m,i,j) \right) \right) \right) \cdot c \right]^{-1} \quad \textbf{(5-14)}$$

The number of SRI messages from MSCs in region x to HLRs in region y, is obtained from the following summation,

$$\lambda_{SRI\,xy} = \sum_{m=1}^{M} \alpha'_{xmy} \,. \quad \textbf{(5-15)}$$

The SRI traffic along each route, z, from region x to region y is then given by

$$\lambda_{SRI\,xyz} = (1/R_{xy})\lambda_{SRI\,xy}, \qquad \text{where } 1 \le z \le R_{xy}.$$

The total number of PRN messages generated by HLRs in region x to MSCs in region y, is given below where the probability of success of the preceding SRI messages is calculated.

$$\lambda_{PRN\,xy} = \sum_{m=1}^{M} \alpha'_{myx} \sum_{i=1}^{R_{mx}} \frac{1}{R_{mx}} P_S(\phi_{SRI},0,m,x,i,Q_{F\,mxi}) \quad \textbf{(5-16)}$$

The PRN traffic along each route, z, from region x to region y is given by

$$\lambda_{PRN\,xyz} = (1/R_{xy})\lambda_{PRN\,xy}, \qquad \text{where } 1 \le z \le R_{xy}.$$

Every successful PRN message generates a PRNA message containing a roaming number that has been temporarily assigned to the called subscriber. Let $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

$$\lambda_{PRNA\,xyz} = \frac{1}{R_{yx}} \lambda_{PRN\,yx} P_S(\phi_{PRN},0,y,x,i,Q_{F\,yxi}) \quad \textbf{(5-17)}$$

A +SRIA message containing a roaming number is returned to the MSC that originated the SRI message if the provide roaming number procedure is successfully completed. Let $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

$$\lambda_{+SRIA\,xyz} = \sum_{m=1}^{M} \frac{\alpha'_{ymx}}{R_{yx}} P_S(\phi_{SRI},0,y,x,i,Q_{F\,yxi})$$
$$\times \sum_{r=1}^{R_{xm}} \left(\frac{1}{R_{xm}} P_S(\phi_{PRN},0,x,m,r,Q_{F\,xmr}) P_S(\phi_{PRNA},1,x,m,r,Q_{R\,xmr}) \right) \quad \textbf{(5-18)}$$

If the provide roaming number procedure fails due to timer expiration a –SRIA message is returned to the calling party's MSC. The rate at which these messages are generated for each route is given by

$$\lambda_{-SRIAxyz} = \sum_{m=1}^{M} \frac{\alpha'_{ymx}}{R_{yx}} P_S(\phi_{SRI},0,y,x,i,Q_{Fyxi}) \cdot \left(\sum_{r=1}^{R_{xm}} \left(\sum_{j=1}^{Q_{Fxmr}} \frac{1}{R_{xm}} P_D(\phi_{PRN},0,x,m,r,j) + \sum_{j=1}^{Q_{Rxmr}} \frac{1}{R_{xm}} P_S(\phi_{PRN},0,x,m,r,Q_{Fxmr}) P_D(\phi_{PRNA},1,x,m,r,j) \right) \right) \quad \textbf{(5-19)}$$

where $z = \Gamma(y,x,i)$ and $1 \le i \le R_{xy}$.

The roaming number obtained from the +SRIA message is placed in the called party address field of the IAM message and is used to route the call to the destination MSC. The number of IAM messages from region x to region y is given by the following equation, which calculates the probability that all the MAP messages were successful.

$$\lambda_{IAMxy} = \sum_{m=1}^{M} \alpha'_{xym} \sum_{i=1}^{R_{xm}} \frac{1}{R_{xm}} P_S(\phi_{SRI},0,x,m,i,Q_{Fxmi}) P_S(\phi_{SRIA},1,x,m,i,Q_{Rxmi}) \times \sum_{i=1}^{R_{my}} \left(\frac{1}{R_{my}} P_S(\phi_{PRN},0,m,y,i,Q_{Fmyi}) P_S(\phi_{PRNA},1,m,y,i,Q_{Rmyi}) \right) \quad \textbf{(5-20)}$$

The arrival rate of ACM, ANM, REL and RLC messages can be obtained from equations (3-15) to (3-17).

The roaming number retrieval success rate or the success rate of the send routing information procedure, S_{RN}, for the MSCs in region x, is given by the following

$$S_{RNx} = \sum_{i=1}^{M} \sum_{j=1}^{R_{ix}} \lambda_{+SRIAixj} P_S(\phi_{SRIA},0,i,x,j,Q_{Fixj}) \quad \textbf{(5-21)}$$

and the call completion rate is given by (3-18).

Other performance parameters such as the message throughput and end-to-end mean delay for the GSM location management protocol, the GSM call delivery protocol and the protocols in the following sections can be obtained with equations (3-19) to (3-30).

5.2.3 *Simple Location Update Protocol*

In both GSM [ETSI GSM 09.02] and IS-41 [Lin & DeVries, 1995] the location update procedure is accompanied by a procedure that is used to transfer the subscriber's profile from the HLR to the VLR. Many previous studies on location management have either completely overlooked this procedure or have assumed that this procedure was combined with the location update procedure and therefore did not add any additional signalling or database transaction load (e.g. [La Porta et al., 1996], [Krishna et al., 1996] and [Lin, 2000]). Since this simplified location management technique has been widely used in literature, the signalling traffic generated by this technique is analysed below and compared with the GSM location management protocol.

In the Simple Location Update (SLU) protocol the insert subscriber data procedure is not implemented. Instead, subscribers are assumed to have a generic profile and additional subscription information is included within the location update response from the HLR. Equations (5-2) to (5-7) can therefore be simplified down to the following:

$$\lambda_{ULxy} = \sum_{m=1}^{M} \beta_{xmy} \sum_{n=0}^{g} \left[\sum_{i=1}^{R_{xy}} \left(\sum_{j=1}^{Q_{F\,xyi}} \frac{1}{R_{xy}} P_D(\phi_{UL},0,x,y,i,j) + \sum_{j=1}^{Q_{R\,xyi}} \frac{1}{R_{xy}} P_S(\phi_{UL},0,x,y,i,Q_{F\,xyi}) P_D(\phi_{ULA},1,x,y,i,j) \right) \right]^n \tag{5-22}$$

$$\lambda_{ISDAxyz} = \lambda_{ISDxyz} = 0, \tag{5-23}$$

$$\lambda_{+ULAxyz} = \frac{1}{R_{yx}} \lambda_{ULyx} P_S(\phi_{UL},0,y,x,i,Q_{F\,yxi}), \tag{5-24}$$

$$\lambda_{-ULAxyz} = 0. \tag{5-25}$$

5.2.4 *Super-Charged Location Update Protocol*

The Super-Charger mechanism is a method defined by 3GPP (Section 2.6.3.2) to help reduce the signalling traffic and database transaction load in GSM and 3G networks. Even though Super-Charging reduces the signalling load, modifications to a large number of subscriber profiles could lead to an increase in the signalling load as the profiles of the affected subscribers will have to be transferred whenever they register on a VLR with the old profile. Super-Charging is therefore analysed, below, and compared with the GSM location management protocol. Equation (5-2) is rewritten as

$$\lambda_{ULxy} = \sum_{m=1}^{M} \delta \beta_{xmy} \sum_{n=0}^{g} X^n, \tag{5-26}$$

which represents the arrival rate of UL messages for location updates that require the transfer of a subscriber's profile. The rate at which SUL messages are generated from MSCs in region x to HLRs in region y is given by

$$\lambda_{SULxy} = \sum_{m=1}^{M} (1-\delta) \beta_{xmy} \sum_{n=0}^{g} \left[\sum_{i=1}^{R_{xy}} \left(\sum_{j=1}^{Q_{F\,xyi}} \frac{1}{R_{xy}} P_D(\phi_{UL},0,x,y,i,j) + \sum_{j=1}^{Q_{R\,xyi}} \frac{1}{R_{xy}} P_S(\phi_{UL},0,x,y,i,Q_{F\,xyi}) P_D(\phi_{ULA},1,x,y,i,j) \right) \right]^n \tag{5-27}$$

and the arrival rate of SULA messages from each region is given by

$$\lambda_{SULAxyz} = \frac{1}{R_{yx}} \lambda_{SULyx} P_S(\phi_{UL},0,y,x,i,Q_{F\,yxi}). \tag{5-28}$$

The CL and CLA message arrival rates in equations (5-8) and (5-9) is rewritten as follows for Super-Charged networks:

$$\lambda_{CLAxy} = \lambda_{CLxy} = 0. \tag{5-29}$$

The location update success rate for region x, S_{LUx}, is now given by the following equation, which calculates the probability that the +ULA and SULA messages are successfully delivered:

$$S_{LUx} = \sum_{i=1}^{M}\sum_{j=1}^{R_{ix}} \lambda_{+ULAixj} P_S\left(\phi_{ULA},0,i,x,j,Q_{F\,ixj}\right) + \sum_{i=1}^{M}\sum_{j=1}^{R_{ix}} \lambda_{SULAixj} P_S\left(\phi_{ULA},0,i,x,j,Q_{F\,ixj}\right). \quad \textbf{(5-30)}$$

5.2.5 *Lightweight Location Lookup Protocol*

The Lightweight Location Lookup Protocol (LiLLP) (Section 2.6.2.4) is similar to the procedures used in GSM to locate a subscriber for the delivery of a short message or to activate a network initiated GPRS session. Like the Super-charging mechanism, the LiLLP could be easily integrated into networks that support the standard GSM and 3G mobility management protocols for call delivery. This could be achieved through the use of optional parameters in the UL and SRI messages to indicate that the end-points support LiLLP. In a LiLLP network a HLR would have to respond to a SRI with a SRIA message that includes the called party's *MSC address*, rather a roaming number. The originating MSC could then use the *MSC address* to route the call to the terminating MSC, together with the called party's number in an optional parameter field of the IAM message.

To obtain the message arrivals rates let $Y_1 = 0$ and $Y_2 = 1$ in equation (5-14) while equations (5-16) to (5-20) can be reduced down to the following:

$$\lambda_{PRNAxy} = \lambda_{PRNxy} = 0, \quad \textbf{(5-31)}$$

$$\lambda_{+SRIAxyz} = \sum_{m=1}^{M} \frac{\alpha'_{ymx}}{R_{yx}} P_S\left(\phi_{SRI},0,y,x,i,Q_{F\,yxi}\right), \quad \textbf{(5-32)}$$

$$\lambda_{-SRIAxyz} = 0, \quad \textbf{(5-33)}$$

$$\lambda_{IAMxy} = \sum_{m=1}^{M} \alpha'_{xym} \sum_{i=1}^{R_{xm}} \frac{1}{R_{xm}} P_S\left(\phi_{SRI},0,x,m,i,Q_{F\,xmi}\right) P_S\left(\phi_{SRIA},1,x,m,i,Q_{R\,xmi}\right). \quad \textbf{(5-34)}$$

5.3 Numerical Results and Discussion

This Section presents the numerical results obtained for the steady state equilibrium analysis of the network shown in Figure 5-1. It is similar to the network used in the previous chapters, but here each region consists of 6 MSCs (each with an integrated VLR) and a HLR. As in the previous chapters the shortest-path routing algorithm is used.

The parameters used here are identical to those used in Chapter 3, but with the following modifications and additions:

- All the signalling links have a transfer rate of μ_L = 2048 kb/s in order to transfer the large signalling messages present in mobile networks without becoming congested.
- The individual message lengths are σ_{UL} = 70 octets, σ_{ULA} = 57 octets, σ_{CL} = 55 octets, σ_{CLA} = 40 octets, σ_{ISD} = 279 octets, σ_{ISDA} = 80 octets, σ_{SRI} = 57 octets, σ_{SRIA} = 69 octets, σ_{PRN} = 70 octets and σ_{PRNA} = 59 octets. These values were obtained from [Meier-Hellstern & Alonso, 1991]. The maximum MSU size (including the SIF, SIO and MTP level 3 information) is used for ISD messages, as multiple ISD messages are normally required to transfer large subscriber profiles.
- The analysis considers location updates in the circuit switched domain where the attempt counter = 4 (i.e. $g = 3$).
- Unless otherwise stated the analysis assumes two ISD messages are required to transfer a subscriber profile (i.e. $d = 2$). This value is based on the author's own recent observations of signalling traffic in GSM networks.

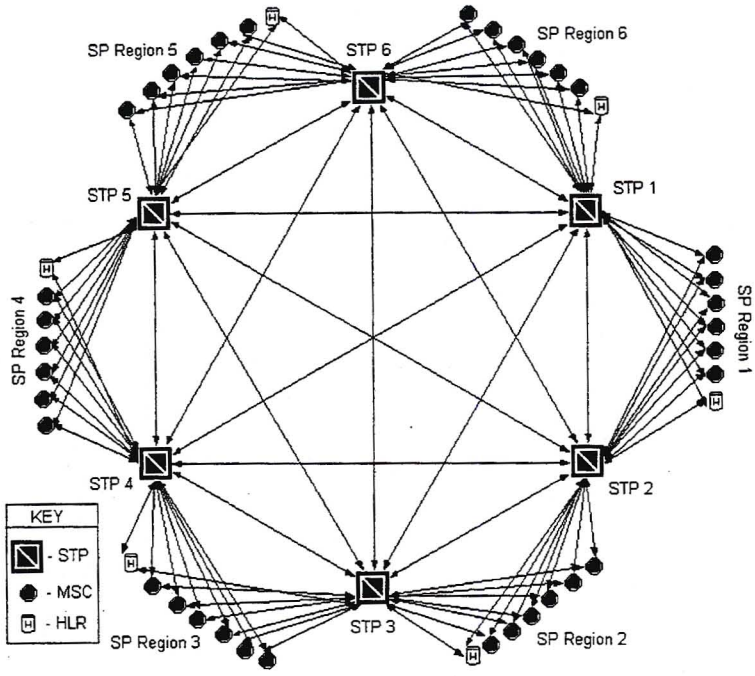


Figure 5-1. GSM signalling network architecture.

The first offered location update load from each region is assumed to be equal in the analysis. In the mobility model, fifty-percent of the new location update attempts in each VLR are due to subscribers who have arrived from VLRs in the same region. The other fifty-percent of the new location update attempts are due to subscribers who have arrived from the other regions - these subscribers have an equal probability of arriving from any of the other 30 VLRs. For example, if a region generates 60 LU attempts/sec, then 30 attempts/sec are due to subscribers who have moved between VLRs in the same region and 6 attempts/sec are due to subscribers who have arrived from each one of the other regions. Subscriber profiles are equally distributed across all the HLRs, therefore 10 UL messages/sec are sent to each HLR.

A call originating at a MSC has an equal probability of being destined to any of the other 35 MSCs in the network (similar to call distribution in Section 3.4.1). Since subscriber profiles are equally distributed across all the HLRs, if the first offered load is 60 calls/sec in each region then 10 SRI messages/sec are sent to each HLR. The overload scenarios analysed below are for a network wide overload.

5.3.1 Single Discard Threshold

In the single discard threshold scheme, excess messages arriving at the congested STPs are simply discarded when the maximum threshold of $K_1 = K_2 = K_3 = 100$ is reached. The results below are structured as follows:

- Section 5.3.1.1 examines the performance of the GSM location management protocol in a scenario where all the subscribers exhibit a very high mobility compared to the call arrival rate., which is considered to be negligible.
- Section 5.3.1.2 examines the performance of the GSM call delivery protocol and ISUP call set-up protocol in a scenario where all the subscribers exhibit very low mobility but they generate are a high volume of calls.
- Section 5.3.1.3 considers the interaction between the GSM location management, GSM call delivery and ISUP call set-up protocols during STP congestion.
- Section 5.3.1.4 analyses the influence of the location update attempt counter on network performance.

5.3.1.1 Location Management Performance

Figure 5-2 shows the location update success rate for various location update offered loads. An unexpected increase in the location update offered load may be attributed to various factors, including the failure of network elements (e.g. a VLR or BSC), an underestimated increase in the number of subscribers visiting holiday locations or an undesirable (yet planned) change to a location area boundary that results in a large increase in the location update rate. The location update success rate provides a measure of the fraction of mobile terminals that are allowed to roam on the network. The influence of different numbers of ISD messages per location update attempt is also shown in the figure. The curves are similar to those obtained in Chapter 3 for the ISUP call completion rate analysis. When one ISD message is required to transfer subscriber profiles, the location update success rate increases linearly until the success rate peaks at approximately 108 updates/sec and then the curve drops and bends back on itself twice to yield three possible solutions for some of the first offered loads. As is Chapter 3, the top and bottom sections of the curve correspond to stable regions where the network is either congested or uncongested, while the middle region represents a meta-stable region that is not realisable in simulations.

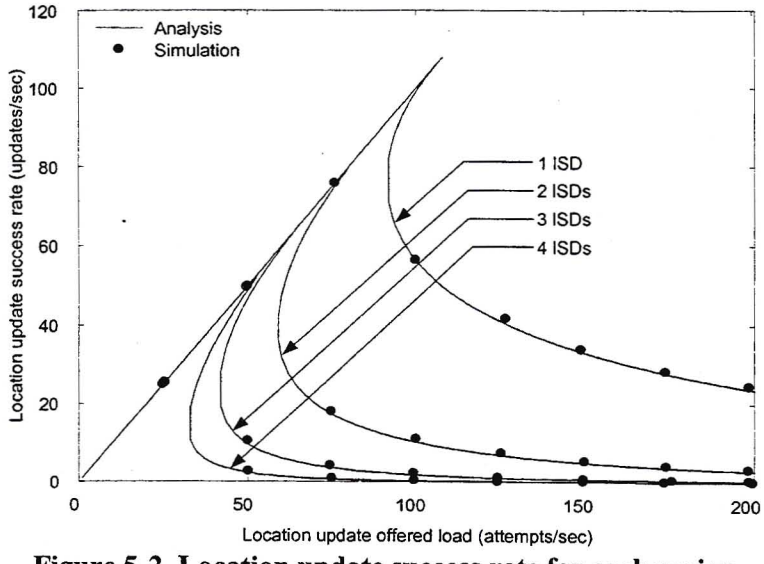


Figure 5-2. Location update success rate for each region.

In the scenario where one ISD message is required, the location update success rate in the congested region of the curve deteriorates gradually for higher loads to approximately 24 updates/sec when the offered load is 200 attempts/sec. When more ISD messages are required to transfer subscriber profiles the peak number of location updates supported by the network is reduced and the location update success rates drops to zero (or close to zero) for higher offered loads. The weaker success rate is due to the lower probability of all the messages involved in a location update being successful.

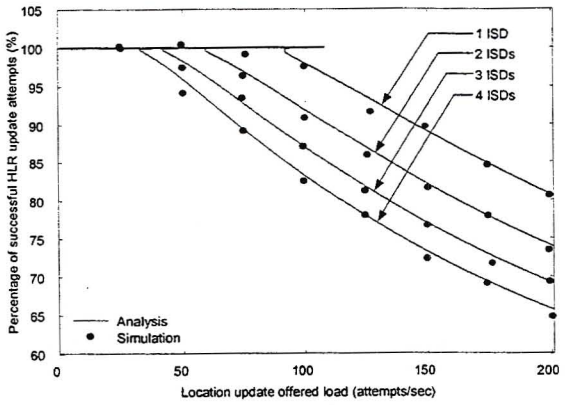


Figure 5-3. Accuracy of each HLR for various offered loads.

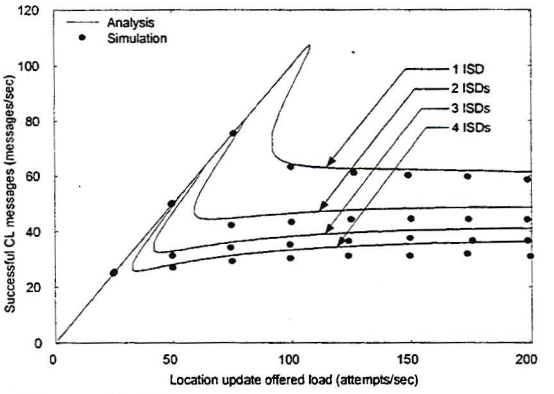


Figure 5-4. Location cancellation success rate for each region.

Maintaining an accurate HLR database is crucial in a mobile network, as the HLR provides the information that is necessary to route revenue generating services to a subscriber. Figure 5-3 shows the accuracy of the HLR's database for various loads. This graph was obtained by calculating the HLR accuracy obtained with equation (5-12) as a percentage of the location update offered load. The meta-stable region is not easily visible in these graphs since the first UL message (or one of the subsequent reattempts) is usually successfully delivered to the HLR and therefore a more than 99% accuracy is obtained. An interesting observation here, is that each HLR has stored the correct location for more than 65% of the subscribers who have moved to a new VLR when the offered load is 200 attempts/sec and 4 ISD messages are required per profile. Unfortunately, none of these subscribers are allowed to roam on the network since almost none of the location update procedures were successfully completed.

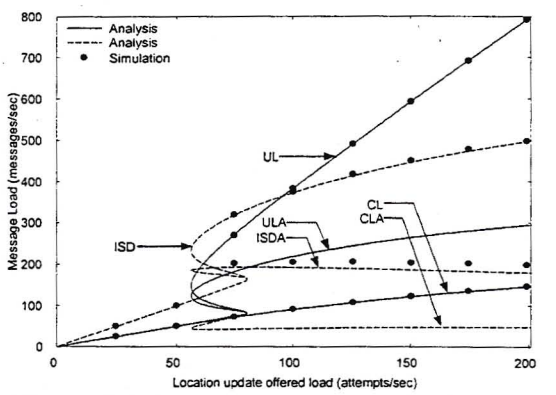


Figure 5-5. Individual message loads from each region.

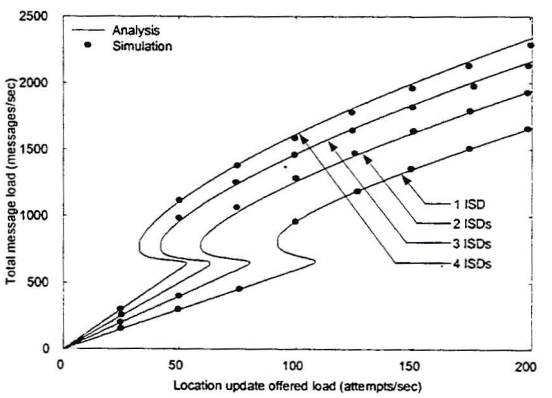


Figure 5-6. Total message load leaving each region.

The location cancellation success rate shows the effectiveness (or ineffectiveness) of the cancel location procedure at deleting subscriber entries from the PVLR. The loss of CL messages results in a depletion of the VLR's storage capacity as the profiles of subscribers who have moved into a new VLR's radio coverage area continues to consume resources in the PVLR. This could eventually lead to mobile terminals not being able to register in an affected VLR, even after the network has returned to the uncongested state. Figure 5-4 shows the number of CL messages that are successfully delivered to VLRs. When only one ISD is required per location update the location cancellation success rate decreases gradually for higher offered loads in the congested region, but it increases when two or more ISDs are required for a location update. This small increase is due to the higher probability of blocking experienced by the ISD messages at higher offered loads, which consequently leads to fewer ISDA messages being generated, as shown Figure 5-5. Figure 5-5 also shows that the signalling traffic load is dominated by UL and ISD messages during STP congestion. The volume of UL and ISD messages is sufficiently high to keep the network in a sustained congestion state even after the

overload traffic is removed. Figure 5-6 shows the total number of messages generated by each region for a various offered loads.

The use of STPs with higher processing capacities increases the maximum number of location update attempts sustainable by the network, but once congested the location update success rate will still drop close to zero in networks where more than one ISD message is required to transfer subscribers' profiles. The inevitable growth of subscriber profiles as new services are deployed in GSM and 3G networks will certainly result in an increase in the number of ISD messages required to transfer profiles from the HLR to the VLR. The above results illustrate the disadvantages of transferring subscriber profiles with a large number of messages, especially during congestion situations. However, high speed ATM signalling links [ITU-T Recommendation Q.2210] and SIGTRAN signalling links [IETF RFC 2719] could offer a solution. They not only allow for higher link capacities but they also support large MSU sizes. This capability can be used to transfer a single large ISD message from the HLR to the VLR, in order to reduce the load on the MTP level 3 routing processors and to increase the number of transactions sustainable by the network.

5.3.1.2 Call Delivery and Call Set-up Performance

Figure 5-7 shows the roaming number retrieval success rate and call completion rate for various call offered loads from each region. The roaming number retrieval success rate shows the effectiveness of the MAP procedures during congestion and call completion rate shows the call success rate as perceived by the customer. The curves are similar to those obtained in Chapter 3 and in the previous section, with the network supporting a peak call handling capability of 72 calls/sec. However, here the region with three solutions is not as prominent as in Chapter 3 with the network only remaining in a sustained congestion state for loads that are slightly below the peak throughput of the system. Network traffic in the congested region of the curve is mainly dominated by SRI messages from caller reattempts (Figure 5-8).

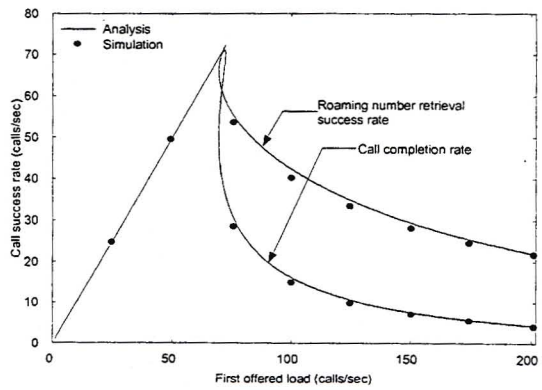


Figure 5-7. Call success rate for each region.

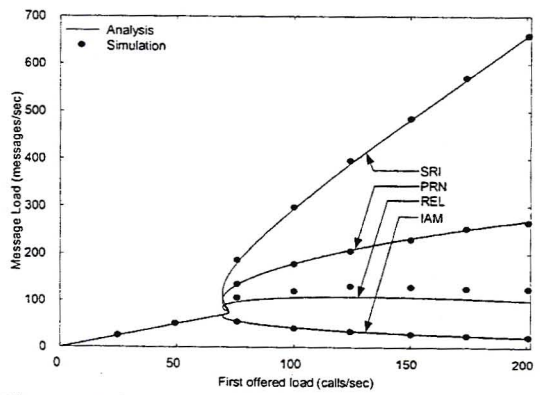


Figure 5-8. Message loads from each region.

5.3.1.3 Mobility Management and Call Set-up interaction

Figure 5-9 shows the influence of various call arrival rates on the location update success rate. The impact of increasing the number of calls is a decrease in the peak number of location update attempts sustainable by the network. Likewise, the location cancellation success rate and HLR accuracy are also reduced. Figure 5-10 shows the corresponding graphs for the roaming number retrieval success rate and call completion rate as a function of the location update offered load. Even though these graphs show that the call completion rate drops close to zero, it would be lower (possibly close to zero) if the inconsistencies in the HLRs' database were also considered in the analysis. Graphs for the location update success rate as a function of the call arrival rate produce analogous results.

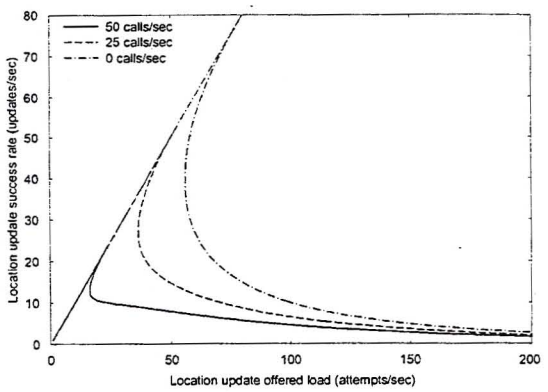


Figure 5-9. Location update success rate for various location update offered loads and call arrival rates.

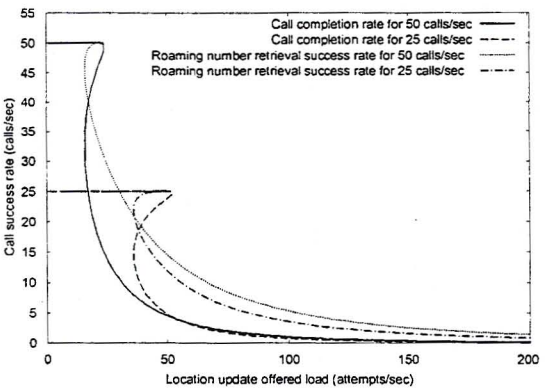


Figure 5-10. Call completion rate and roaming number retrieval success rate for various location update offered loads.

5.3.1.4 The impact of location update reattempts

Congestion in mobile networks, even when no calls are in progress, is aggravated by location update reattempts. Figure 5-11 illustrates the impact of different values of the location update attempt counter on the location update success rate. One notices that the network is susceptible to sustained congestion even when one reattempt is allowed. This is due to UL and ISD messages from reattempts, which are sufficient to keep the network in a congested state after the overload is removed. Even though the location management procedures defined for GPRS are equivalent to those defined for the circuit switched GSM domain, the GPRS attempt counter allows for up to 5 attempts if previous attempts have failed [3GPP TS 24.008 and 29.002]. From Figure 5-11 it is evident that congestion triggered by the GPRS location management procedures will extend the sustained congestion region to a lower value than in GSM. Future work in this area needs to focus on other reattempt strategies, for example exponential backoff algorithms in the event of a location update failure may help to alleviate congestion, especially during network element failures.

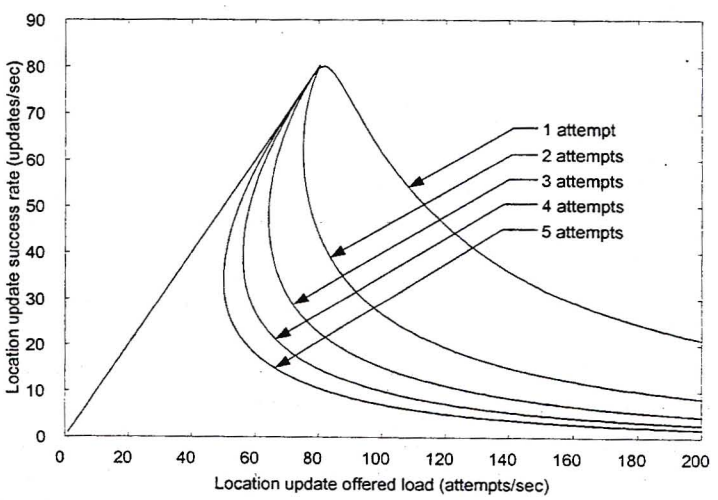


Figure 5-11. Location update success rate for various attempt counter values.

5.3.2 Multiple Discard Thresholds

Previous studies of SS7 congestion and flow control have only examined the effectiveness of multiple discard threshold schemes with ISUP traffic. This section examines how message priorities and selective discarding can help to improve the performance of application level procedures in mobile networks.

Results on the influence of different buffer threshold settings on the location management and call delivery protocols yields conclusions equivalent to those obtained in Section 3.4.1.3. The results are therefore not repeated, instead thresholds of $K_1 = 100$, $K_2 = 120$ and $K_3 = 130$ messages are used in the analysis.

5.3.2.1 Location Management Performance

Section 5.3.1.1 showed that signalling traffic load is dominated by UL and ISD messages during congestion. The high UL message load is a consequence of the reattempts due to previous failures; therefore the UL message load can be reduced by assigning the highest priority (2) to ULA messages and the lowest priority (0) to UL messages. Various combinations of message priority schemes were investigated and of these the performance results obtained for four of the priority schemes are shown in the figures below.

To easily make reference to the different priority assignment schemes a six digit notation of the type $abcdef_{LM}$ is used where

- a is the priority of the UL messages,
- b is the priority of the ISD messages,
- c is the priority of the ISDA messages,
- d is the priority of the ULA messages,
- e is the priority of the CL messages, and
- f is the priority of the CLA messages

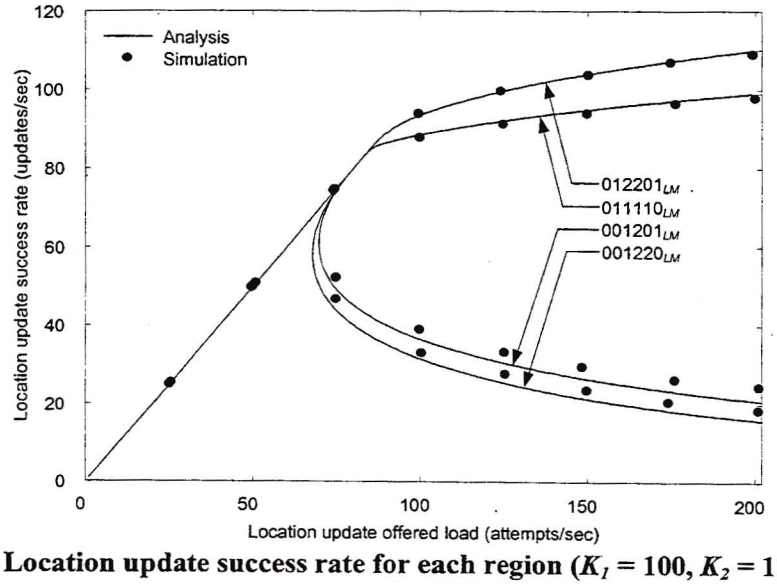


Figure 5-12. Location update success rate for each region ($K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

The following conclusions were obtained from the analysis:

- In the priority schemes were UL messages are assigned a priority of zero and the ISD, ISDA and ULA messages are assigned a non-zero priority the location update success rate does not deteriorate for offered loads that exceed the network's capacity of 80 updates/sec - instead the location update success rate increases gradually. Furthermore, since the ISD and ISDA messages experience a very low probability of discarding very few –ULA messages are generated. However, the location cancellation success rate and the accuracy of the HLR databases are generally lower than that obtained with other priority schemes (Figures 5-12 and 5-13). Nevertheless, this approach is more effective as more subscribers are allowed to roam on the network. Successful HLR updates from subscribers who are not allowed to roam, due to failed location updates, are of no benefit to the network.
- The expected number of messages in the STP's MTP level 3 queue during congestion is approximately 100 messages and therefore messages with a non-zero priority are less likely

to be discarded. The non-zero priority value assigned to the messages has a negligible impact on performance; i.e. the performance difference between the priority schemes 012201_{LM}, 021201_{LM} and 022202_{LM} is very small.

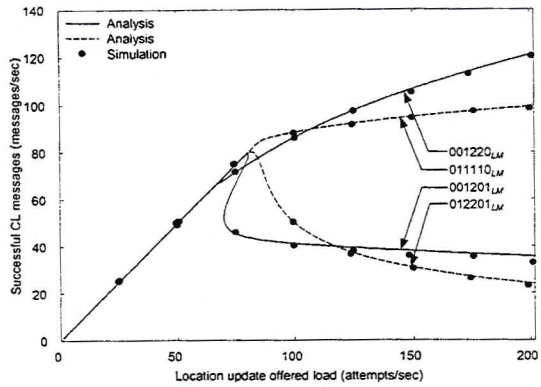


Figure 5-13. Location cancellation success rate for each region ($K_1 = 100, K_2 = 120, K_3 = 130$).

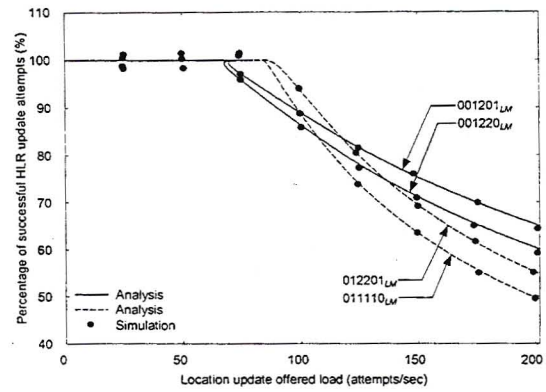


Figure 5-14. HLR accuracy in each region ($K_1 = 100, K_2 = 120, K_3 = 130$).

5.3.2.2 Call Delivery and Call Set-up Performance

From Section 5.3.1.2 it is evident that call reattempts during congestion lead to a deterioration in the call completion rate at very high loads. This section examines how message priorities can influence the call completion rate in a mobile network. The SRI messages are assigned the lowest priority (0) to control the number of calls admitted during congestion. In the following results, PRN and PRNA messages are assigned an intermediate priority (1) and SRIA messages are assigned the highest priority (2). Results are not shown for the cases where various different priorities, of 1 and 2, are assigned to the PRN, PRNA and SRIA messages as the performance differences were found to be negligible. However network performance was found to be very sensitive to the priorities assigned to ISUP messages. As in Chapter 3, a five-digit notation of the type $abcde_{ISUP}$ is used to simplify the representation of the different priority assignment schemes.

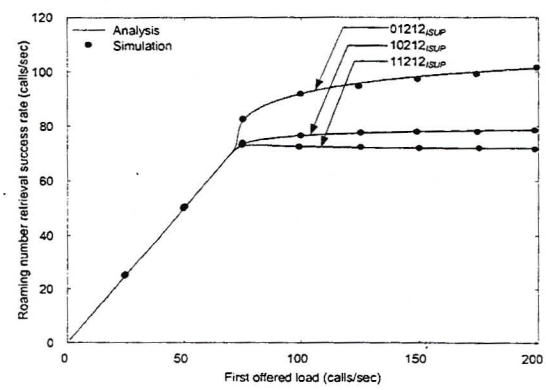


Figure 5-15. Roaming number retrieval success rate at each region for different priority schemes ($K_1 = 100, K_2 = 120, K_3 = 130$).

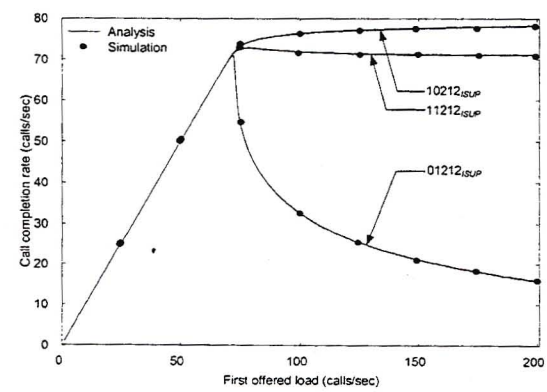


Figure 5-16. Call completion rate at each region for different priority schemes ($K_1 = 100, K_2 = 120, K_3 = 130$).

Figures 5-15 and 5-16 show the roaming number retrieval success rate and the call completion rate for three ISUP priority schemes (01212_{ISUP}, 10212_{ISUP} and 11212_{ISUP}). The main observations and conclusions obtained from these results are as follows:

- Unlike a PSTN network where a discarded IAM message generates a REL and RLC to clear the seized circuit, a discarded UL or SRI message does produce any subsequent messages

(other than reattempts). Furthermore, in the priority schemes where only the SRI message is assigned the lowest priority the number of messages of each type that are admitted into the network is almost constant and therefore the roaming number retrieval success rate and the call completion rate is almost constant.

- When the IAM messages are assigned a priority of zero, the competition for buffer resources at the congested STPs leads to more IAM messages being discarded and therefore the call completion rate deteriorates. Consequently, an increase in call reattempts results in a higher roaming number retrieval success rate.
- In all the location management and call delivery priority schemes examined here the expected number of messages in the STPs' MTP level 3 processing buffers is slightly greater than or less than K_1 since the dominant UL and SRI messages are assigned the lowest priority. As a result messages with a non-zero priority to subject to a very low probability of discard.
- In the cases where the PRN, PRNA and SRIA messages are assigned a non-zero priority very few –SRIA messages are generated.
- Unlike PSTN congestion scenarios, if suitable priorities are assigned to the MAP and ISUP messages in a mobile network, it is possible to maintain the network's performance at an optimum level for both user parts with a minimal degradation in the effective throughput.

GSM networks do not implement message prioritisation with selective discarding and there have been no previous studies that examined the influence of congestion on GSM MAP level procedures and ISUP call set-up procedures. The results obtained in this section strongly suggest that it is possible to maintain the performance of a GSM or a 3G network at an optimum level during periods of overload with a simple selective discard scheme. When the interaction between the mobility management and call set-up protocols is considered the conclusions are equivalent to those obtain in Section 5.3.1.3, i.e. the call completion rate and location update success rate sustainable is reduced, but the curves have similar characteristics of those obtained in this section.

5.3.3 Performance of the Simple Location Update and Super-Charged Location Update Protocols

Figure 5-17 shows the location update success rate of the GSM, SLU and Super-Charged Location Update protocols when 2 ISD messages are required to transfer a subscriber profile. The graphs also show the location update success rate for different values of δ in a Super-Charged network. For the Super-Charged scenario where no ISD messages are exchanged ($\delta = 0$), the network is able to support up to 326 updates/sec (a more than 300% increase when compared to GSM). In the extreme case where $\delta = 1$ the location update success rate is slightly higher than that of GSM since the cancel location procedure is not performed. It is important to note that the value of δ is not constant, and it can vary depending on how often subscribers' profiles change. Likewise, VLR database management schemes and VLR or HLR failures can also influence δ. For example, in a Super-Charged network where $\delta = 0.25$ and the location update arrival rate is 250 attempts/sec all the STPs would be uncongested. However, the modification of a large number of subscriber profiles could increase δ to 0.5 and the resulting signalling traffic load would invoke STP congestion. Sufficient reserve signalling capacity is therefore required in Super-Charged networks, to avoid congestion due to an increase in δ.

A network with the SLU protocol has a peak throughput of 171 updates/sec (58% more than a GSM network that requires only one ISD message to transfer subscriber profiles). The curve obtained for the location update success rate also differs from that obtained for the GSM and Super-Charged location updates. With the SLU protocol, the location update success rate decreases gradually for very high offered loads and the system does not have a region with three solutions where the network will remain in a sustained congestion state after a brief overload.

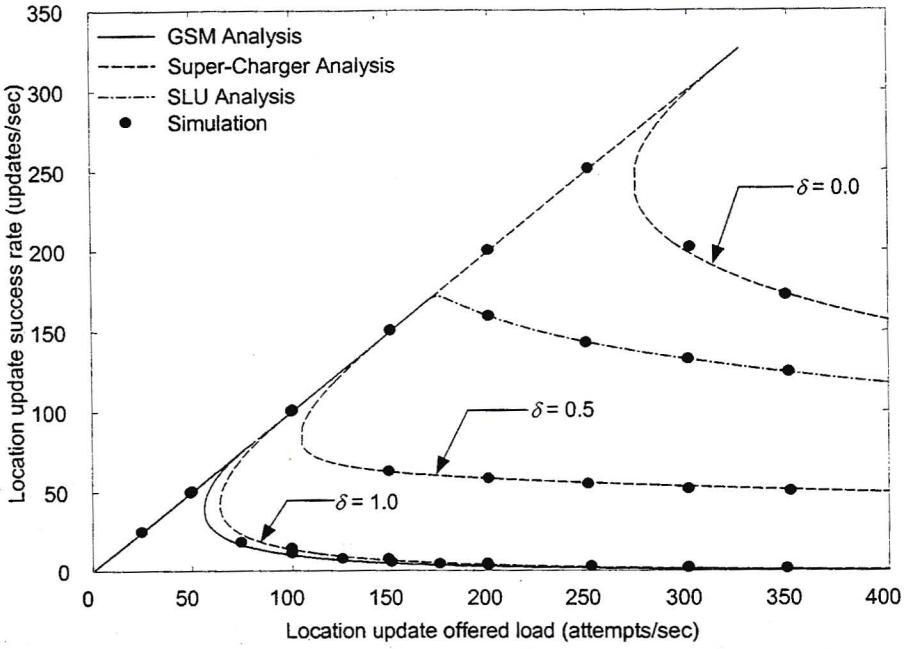


Figure 5-17. Location update success rate comparison between GSM, the Simple Location Update and Super-Charged Location Update Protocols.

While the difference in the performance of the GSM and SLU protocols serves to illustrate the inaccuracy of models used by previous studies on GSM and IS-41 mobility management (for example in [La Porta et al., 1996], [Krishna et al., 1996] and [Lin, 2000]), the SLU protocol may be realisable in future mobile networks. High speed ATM signalling links and SIGTRAN signalling links, which support large MSU sizes, could allow the HLR to include the subscriber profile within the ULA message and thus eliminate need for the insert subscriber data procedure during a location update.

5.3.4 The Lightweight Location Lookup Protocol (LiLLP)

Figure 5-18 shows a comparison between the GSM and LiLLP protocols. The simplified nature of the LiLLP protocol allows for a higher roaming number retrieval success rate to be achieved and consequently a slightly higher call completion rate is also obtained. The curve obtained for the roaming number retrieval rate differs from call completion rate curve, in that the meta-stable region and the congested region of the curve (where three solutions exist) are higher than the uncongested region. The roaming number retrieval success rate is greater than the actual offered load due to call reattempts since the failure of a ISUP call set-up procedure may result in more than one successful send routing information procedure. The discrepancy in the analytical results and simulation results of the LiLLP in Figure 5-18 is due to the correlated nature of REL traffic streams, which are significantly higher than in GSM.

The high roaming number retrieval success rate generates a large number of IAM messages, however the ISUP call completion rate is significantly lower due to competition for STP resources with the large number of SRI and REL messages that are present in the network. Nevertheless, network performance for the LiLLP can be maintained at an optimum level if message priorities and multiple discard thresholds are implemented at the STPs.

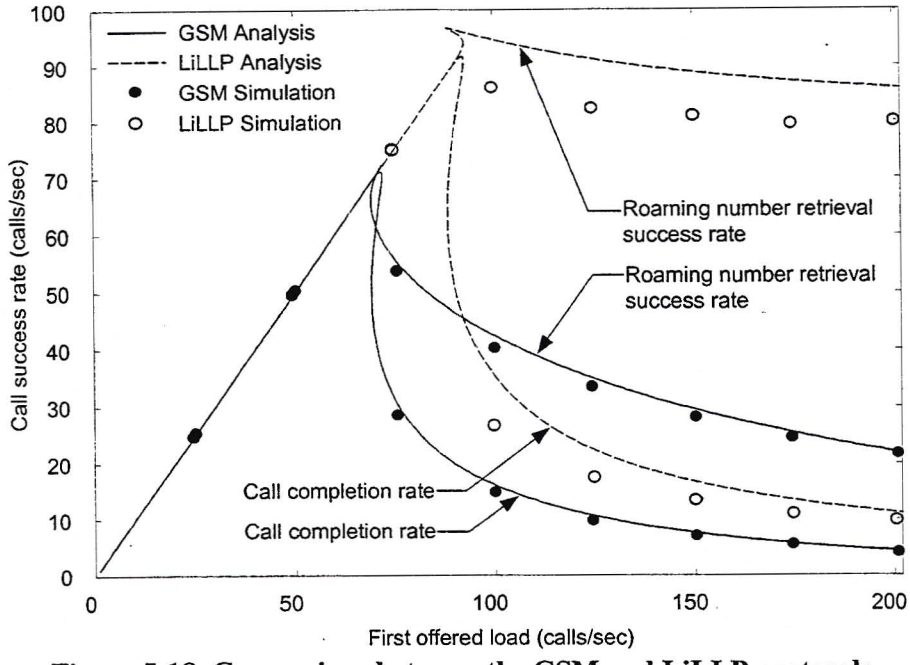


Figure 5-18. Comparison between the GSM and LiLLP protocols.

5.4 Summary

Previous work on SS7 network performance has concentrated on signalling in a PSTN environment, while studies of signalling in mobile networks have used simple techniques of estimating the load attributed to mobility management functions. This chapter extends the mathematical models derived in Chapter 3 to allow one to analyse the performance of signalling protocols in present day and future mobile networks. The results examine the impact of STP congestion on MAP location management and the MAP call delivery protocols as well as the ISUP call set-up procedures. In addition to the GSM mobility management procedures the following were also considered:

- The Simple Location Update (SLU) Protocol
- The Super-Charged Location Update Protocol
- The Lightweight Location Lookup Protocol (LiLLP)

The results obtained for STPs with a single discard threshold are characteristic of those obtained in Chapter 3 for the ISUP call set-up procedure. A burst of location updates or calls can result in a sustained overload state after the overload traffic has been removed. The location update reattempt procedure and call reattempts are primarily responsible for maintaining the sustained overload. Network traffic is therefore dominated by UL and SRI messages. Together with the decrease in the location update success rate for higher offered loads there is an equivalent deterioration in the accuracy of the HLRs' databases and in the location cancellation success rate.

The number of ISD messages required to transfer a subscriber's profile also significantly influences network performance. When two or more ISD messages are required the peak message handling capability of the network is reduced and the location update success rate drops close to zero more rapidly. These results highlight the importance of using the large message size capability of high speed ATM signalling links and SIGTRAN signalling links to transfer the entire subscriber profile in a single ISD message. Alternatively the subscriber profile could be transferred in a single large ULA message, thus eliminating the need for the insert subscriber data procedure.

Unlike PSTN congestion scenarios, if suitable priorities are assigned to the MAP and ISUP messages in a mobile network, it is possible to maintain the network's performance at an optimum level for both users with a minimal degradation in the effective throughput of the system. Network performance is found to be effective if the lowest priority is assigned to the UL and SRI message and the higher priorities are assigned to the other messages (including IAM messages).

A comparison between GSM and SLU protocol (which is commonly used by other researchers) shows that the SLU protocol is not adequate at modelling the performance and behaviour of GSM. Not only does the SLU protocol support a higher location update success rate, since no ISD and ISDA messages are used, but unlike GSM network performance degrades gradually for higher offered loads during congestion. Furthermore, with the SLU protocol the network does not remain congested once the overload traffic is removed. The Super-Charger mechanism is found to significantly reduce signalling load and it can increase throughput by up to 300% when compared to the GSM scenario that requires the transfer of two ISD messages. The LiLLP protocol also offers some relief to the signalling network by reducing the number of messages generated by the MAP call delivery procedures, but the ISUP call completion rate is only marginally improved at very high loads.

6. Transient Analysis of SS7 Network Performance in the PLMN

6.1 Introduction

While a number of previous studies have examined the performance of mobility management protocols and have also proposed alternative approaches to mobility management they have usually focused on the steady state behaviour of mobile networks (e.g. [Jain & Lin, 1995], [Schopp, 2000] and [Mayer, 2000]). Studies that do consider the time varying behaviour of mobile networks have instead used simulation models as part of their analysis (e.g. [Lam et al., 1997] and [Cui et al., 1998]). Yet, they do not explicitly model the SS7 message flows, mobility management timers and procedures, or the impact of congestion and failure scenarios on protocol performance.

As in fixed-line networks, mobile networks also experience call overloads due to media stimulated events. Additionally mobile networks can also experience a surge in mobility management traffic, possibly due to one of the following events:

- The failure of a BTS or BSC that provides radio coverage on the edge of a location area can trigger a surge in the location update traffic due to mobile terminals registering on the adjacent VLR in areas where spill-over radio coverage is available from the adjacent location area.
- A surge in location updates can occur if there is a failure/recovery of a VLR or HLR.
- A large number of subscribers propagating to a particular area (e.g. a cricket match) can also trigger to an increase in the number of location updates and calls in that area.

In some cases the sudden increase in mobility management traffic can invoke signalling network congestion. In this chapter transient mathematical models are derived to evaluate the performance of the GSM mobility management protocol, the SLU protocol, the Super-Charged Location Update protocol and the LiLLP protocol.

6.2 Transient Mathematical Analysis

The notation used in this section is identical to that used in earlier chapters, but with the following modifications and additions:

i) $\alpha(t)$ is a $M \times M \times M$ call arrival rate matrix at time t where $\alpha_{ijk}(t)$ is the number of new call arrivals per second in region i that are destined to subscribers who are roaming in region j and have their profiles stored on a HLR in region k. The first offered load attributed to new call attempts from region i is given by

$$\alpha_i = \sum_{j=1}^{M} \sum_{k=1}^{M} \alpha_{ijk} .$$

ii) $\alpha'(t)$ is a $M \times M \times M$ call arrival rate matrix at time t where $\alpha'_{ijk}(t)$ is the number of call arrivals per second including reattempts, where i denotes the originating region of the calls, j denotes the terminating region and k is the region in which the HLR of the called subscribers is located.

iii) $\beta(t)$ is a $M \times M \times M$ mobility matrix at time t where $\beta_{ijk}(t)$ is the number of new location update attempts per second in region i, due to subscribers who were last registered on a VLR in region j and have their profiles stored on a HLR in region k. The first offered load attributed to new location update attempts in region i is given by

$$\beta_i = \sum_{j=1}^{M}\sum_{k=1}^{M}\beta_{ijk} .$$

iv) T_{3210} is the timer that is started when the mobile terminal initiates a location update. If a successful acknowledgement is not received before T_{3210} expires the location update is assumed to have failed.

v) T_{3211} is the time period before a location update is reattempted after a failure.

vi) T_{ISD} is the timer that is started when the insert subscriber data procedure is initiated. If T_{ISD} expires a –ULA message is returned to the VLR.

vii) T_{SRI} is the timer that is started when a SRI message is sent from a MSC to a HLR. If T_{SRI} expires the call is assumed to have failed.

viii) T_{PRN} is the timer that is started when the provide roaming number procedure is initiated. If T_{PRN} expires a –SRIA message is returned to the MSC.

ix) $T_{UL} = T_{3210} + T_{3211}$ is the time period between location update attempts if a UL or ULA message was discarded during the first attempt.

x) $T_{ULF} = T_{ISD} + T_{3211} + Q_{F\,mni}D + Q_{R\,mni}D$ determines the time period between location update attempts in equations (6-1) to (6-4) if an ISD or ISDA message was discarded during the first attempt and the resulting –ULA was successfully delivered to the MSC.

6.2.1 GSM Location Management

The transmission rate of UL messages from VLRs in region m to HLRs in region n at time t is given by the following expression,

$$\lambda_{UL\,mn}(t) = \sum_{k=1}^{M}\beta_{mkn}(t) + \sum_{k=1}^{M}\beta'_{mkn}(t) + \sum_{k=1}^{M}\beta''_{mkn}(t) + \sum_{k=1}^{M}\beta'''_{mkn}(t) \qquad \textbf{(6-1)}$$

where the first term denotes the arrival rate of new location update attempts, and the subsequent terms represent the arrival rates of the first, second and third reattempts respectively.

The following equations evaluate the arrival rates of the reattempts in (6-1) at time t.

$$\beta'_{mkn}(t) = \sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}\left[\beta_{mkn}(t-T_{UL})X_1 + \beta_{mkn}(t-T_{ULF})X_2\right]$$

$$\beta''_{mkn}(t) = \sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}\left[\beta'_{mkn}(t-T_{UL})X_1 + \beta'_{mkn}(t-T_{ULF})X_2\right]$$

$$\beta'''_{mkn}(t) = \sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}\left[\beta''_{mkn}(t-T_{UL})X_1 + \beta''_{mkn}(t-T_{ULF})X_2\right]$$

X_1 in the above equation determines the probability of a reattempt due to expiry of T_{3210}, and is calculated as follows

$$
\begin{aligned}
X_1 &= \sum_{j=1}^{Q_{F\,mni}} P_D(\phi_{UL},0,m,n,i,j,t-T_{UL}) \\
&+ \Big[P_S(\phi_{UL},0,m,n,i,Q_{F\,mni},t-T_{UL})P_S(\phi_{ISD},1,m,n,i,Q_{R\,mni},t-T_{UL}+Q_{F\,mni}D)^d \\
&P_S(\phi_{ISDA},0,m,n,i,Q_{F\,mni},t-T_{UL}+Q_{F\,mni}D+Q_{R\,mni}D)^d \\
&\sum_{j=1}^{Q_{R\,mni}} P_D(\phi_{ULA},1,m,n,i,j,t-T_{UL}+Q_{F\,mni}D+Q_{R\,mni}D+Q_{F\,mni}D)\Big] \\
&+ P_S(\phi_{UL},0,m,n,i,Q_{F\,mni},t-T_{UL})\sum_{j=1}^{Q_{R\,mni}} P_D(\phi_{ULA},1,m,n,i,j,t-T_{UL}+Q_{F\,mni}D+T_{ISD})\cdot X_3
\end{aligned}
\tag{6-2}
$$

where first two terms evaluate the probability that a UL or +ULA message was previously discarded. The third term evaluates the probability that an ISD or ISDA message was discarded and thereafter the subsequent –ULA message was also discarded. X_3 is given by the following

$$
\begin{aligned}
X_3 &= \Bigg[\sum_{h=1}^{d} P_S(\phi_{ISD},1,m,n,i,Q_{R\,mni},t-T_{UL}+Q_{F\,mni}D)^{h-1} \\
&\sum_{j=1}^{Q_{R\,mni}} P_D(\phi_{ISD},1,m,n,i,j,t-T_{UL}+Q_{F\,mni}D)\Bigg]+\Bigg[P_S(\phi_{ISD},1,m,n,i,Q_{R\,mni},t-T_{UL}+Q_{F\,mni}D)^d \\
&\sum_{h=1}^{d} P_S(\phi_{ISDA},0,m,n,i,Q_{F\,mni},t-T_{UL}+Q_{F\,mni}D+Q_{R\,mni}D)^{h-1} \\
&\sum_{j=1}^{Q_{F\,mni}} P_D(\phi_{ISDA},0,m,n,i,j,t-T_{UL}+Q_{F\,mni}D+Q_{R\,mni}D)\Bigg]
\end{aligned}
\tag{6-3}
$$

where each term calculates the probability that an ISD or ISDA message was discarded in (6-2). The probability that an ISD or ISDA message was discarded and the resulting –ULA was successfully delivered is given by the following

$$
\begin{aligned}
X_2 &= P_S(\phi_{UL},0,m,n,i,Q_{F\,mni},t-T_{ULF})P_S(\phi_{ULA},1,m,n,i,Q_{R\,mni},t-T_{ULF}+Q_{F\,mni}D+T_{ISD}) \\
&\cdot\Bigg[\sum_{j=1}^{Q_{R\,mni}} P_D(\phi_{ISD},1,m,n,i,j,t-T_{ULF}+Q_{F\,mni}D)\sum_{h=1}^{d} P_S(\phi_{ISD},1,m,n,i,Q_{R\,mni},t-T_{ULF}+Q_{F\,mni}D)^{h-1} \\
&+ P_S(\phi_{ISD},1,m,n,i,Q_{R\,mni},t-T_{ULF}+Q_{F\,mni}D)^d \\
&\sum_{j=1}^{Q_{F\,mni}} P_D(\phi_{ISDA},0,m,n,i,j,t-T_{ULF}+Q_{F\,mni}D+Q_{R\,mni}D) \\
&\sum_{h=1}^{d} P_S(\phi_{ISDA},0,m,n,i,Q_{F\,mni},t-T_{ULF}+Q_{F\,mni}D+Q_{R\,mni}D)^{h-1}\Bigg]
\end{aligned}
\tag{6-4}
$$

The transmission rate of ISD messages from HLRs is given below, where d ISD messages are generated for every UL message that is received. Let $z = \Gamma(n,m,i)$ where $1 \le i \le R_{mn}$.

$$
\lambda_{ISDmnz}(t) = \frac{d}{R_{nm}}\lambda_{ULnm}(t-Q_{F\,nmi}D)P_S(\phi_{UL},0,n,m,i,Q_{F\,nmi},t-Q_{F\,nmi}D) \tag{6-5}
$$

The transmission rate of ISDA messages at time t is given by

$$
\lambda_{ISDAmnz}(t) = \lambda_{ISDnmi}(t-Q_{F\,nmi}D)P_S(\phi_{ISD},0,n,m,i,Q_{F\,nmi},t-Q_{F\,nmi}D) \tag{6-6}
$$

where $z = \Gamma(n,m,i)$ and $1 \le i \le R_{mn}$ in (6-6) and the subsequent equations.

The transmission rate of +ULA messages indicating the successful completion of location updates is given by following equation.

$$\lambda_{+ULAmnz}(t) = \frac{\lambda_{ISDmnz}\left(t - Q_{F\,mnz}D - Q_{F\,nmi}D\right)}{d} P_S\left(\phi_{ISDA},0,n,m,i,Q_{F\,nmi},t - Q_{F\,nmi}D\right)^d \cdot P_S\left(\phi_{ISD},0,m,n,z,Q_{F\,mnz},t - Q_{F\,mnz}D - Q_{F\,nmi}D\right)^d \quad \textbf{(6-7)}$$

if all the ISD and ISDA messages were previously successful.

The transmission rate of –ULA messages indicating a location update failure due to the discarding of an earlier ISD or ISDA message is given by

$$\lambda_{-ULAmnz}(t) = \frac{\lambda_{ISDmnz}\left(t - T_{ISD}\right)}{d} \sum_{j=1}^{Q_{F\,mnz}} P_D\left(\phi_{ISD},0,m,n,z,j,t - T_{ISD}\right) \cdot \sum_{h=1}^{d} P_S\left(\phi_{ISD},0,m,n,z,Q_{F\,mnz},t - T_{ISD}\right)^{h-1} + \frac{\lambda_{ISDAnmi}\left(t - T_{ISD} + Q_{R\,nmi}D\right)}{d} \sum_{j=1}^{Q_{F\,nmi}} P_D\left(\phi_{ISDA},0,n,m,i,j,t - T_{ISD} + Q_{R\,nmi}D\right) \cdot \sum_{h=1}^{d} P_S\left(\phi_{ISDA},0,n,m,i,Q_{F\,nmi},t - T_{ISD} + Q_{R\,nmi}D\right)^{h-1} \quad \textbf{(6-8)}$$

A CL message is generated by a HLR on receipt of the first successful UL message from a location update. The transmission rate of CL messages from HLRs in region m to VLRs in region n at time t is given by

$$\lambda_{CLmn}(t) = \sum_{k=1}^{M} \sum_{x=0}^{3} \sum_{i=1}^{R_{km}} \left(\frac{1}{R_{km}} \beta_{knm}\left(t - Q_{F\,kmi}D - xT_{UL}\right) P_S\left(\phi_{UL},0,k,m,i,Q_{F\,kmi},t - Q_{F\,kmi}D\right) \right) \prod_{y=1}^{x} \sum_{j=1}^{Q_{F\,kmi}} P_D\left(\phi_{UL},0,k,m,i,j,t - Q_{F\,kmi}D - yT_{UL}\right) \quad \textbf{(6-9)}$$

and the number of CLA messages generated at time t is

$$\lambda_{CLAmnz}(t) = \frac{1}{R_{nm}} \lambda_{CLnm}\left(t - Q_{F\,nmi}D\right) P_S\left(\phi_{CL},0,n,m,i,Q_{F\,nmi},t - Q_{F\,nmi}D\right). \quad \textbf{(6-10)}$$

The performance measures for the location update success rate, the location cancellation success rate and the instantaneous HLR accuracy are given by the following equations respectively:

$$S_{LUm}(t) = \sum_{i=1}^{M} \sum_{j=1}^{R_{im}} \lambda_{+ULAimj}\left(t - Q_{F\,imj}D\right) P_S\left(\phi_{ULA},0,i,m,j,Q_{F\,imj},t - Q_{F\,imj}D\right) \quad \textbf{(6-11)}$$

$$S_{LCm}(t) = \sum_{i=1}^{M} \sum_{j=1}^{R_{im}} \lambda_{CLimj}\left(t - Q_{F\,imj}D\right) P_S\left(\phi_{CL},0,i,m,j,Q_{F\,imj},t - Q_{F\,imj}D\right) \quad \textbf{(6-12)}$$

$$S_{HLR\,m}(t)=\sum_{n=1}^{M}\sum_{k=1}^{M}\sum_{x=0}^{3}\left(\sum_{i=1}^{R_{km}}\left(\frac{1}{R_{km}}\beta_{knm}(t-Q_{F\,kmi}D-xT_{UL})P_S(\phi_{UL},0,k,m,i,Q_{F\,kmi},t-Q_{F\,kmi}D)\right)\right. \quad \textbf{(6-13)}$$
$$\left.\prod_{y=1}^{x}\sum_{j=1}^{Q_{F\,kmi}}P_D(\phi_{UL},0,k,m,i,j,t-Q_{F\,kmi}D-yT_{UL})\right)$$

6.2.2 GSM Call Delivery

The total number of calls originating from a region including reattempts can be evaluated from the following equation

$$\alpha'_{mnh}(t)=\alpha_{mnh}(t)+c\int_0^{\infty}[Y_1+Y_2+Y_3+Y_4+Y_5+Y_6+Y_7]w(x)dx \quad \textbf{(6-14)}$$

where terms Y_1 through to Y_7 represent the following:

- the probability that a previous SRI message was discarded,

$$Y_1=\frac{1}{R_{mh}}\sum_{i=1}^{R_{mh}}\alpha'_{mnh}(t-x-T_{SRI})\sum_{j=1}^{Q_{F\,mhi}}P_D(\phi_{SRI},0,m,h,i,j,t-x-T_{SRI})$$

- the probability that a previous PRN message and the subsequent –SRIA message were discarded,

$$Y_2=\frac{1}{R_{mh}}\sum_{i=1}^{R_{mh}}\alpha'_{mnh}(t-x-T_{SRI})P_S(\phi_{SRI},0,m,h,i,Q_{F\,mhi},t-x-T_{SRI})$$
$$\cdot\sum_{k=1}^{R_{hn}}\frac{1}{R_{hn}}\left(\sum_{j=1}^{Q_{F\,hnk}}P_D(\phi_{PRN},0,h,n,k,j,t-x-T_{SRI}+Q_{F\,mhi}D)\right)$$
$$\cdot\left(\sum_{j=1}^{Q_{R\,mhi}}P_D(\phi_{SRIA},1,m,h,i,j,t-x-T_{SRI}+T_{PRN})\right)$$

- the probability that a previous PRN message was discarded and the subsequent –SRIA message was successful,

$$Y_3=\frac{1}{R_{mh}}\sum_{i=1}^{R_{mh}}\alpha'_{mnh}(t-x-T_{PRN}-Q_{F\,mhi}D-Q_{R\,mhi}D)$$
$$\cdot P_S(\phi_{SRI},0,m,h,i,Q_{F\,mhi},t-x-T_{PRN}-Q_{F\,mhi}D-Q_{R\,mhi}D)$$
$$\cdot\sum_{k=1}^{R_{hn}}\frac{1}{R_{hn}}\sum_{j=1}^{Q_{F\,hnk}}P_D(\phi_{PRN},0,h,n,k,j,t-x-T_{PRN}-Q_{R\,mhi}D)$$
$$P_S(\phi_{SRIA},1,m,h,i,Q_{R\,mhi},t-x-Q_{R\,mhi}D)$$

- the probability that a previous PRNA message and the subsequent –SRIA message were discarded,

$$Y_4 = \frac{1}{R_{mh}} \sum_{i=1}^{R_{mh}} \alpha'_{mnh}(t-x-T_{SRI}) P_S(\phi_{SRI},0,m,h,i,Q_{F\,mhi},t-x-T_{SRI})$$
$$\frac{1}{R_{hn}} \sum_{k=1}^{R_{hn}} P_S(\phi_{PRN},0,h,n,k,Q_{F\,hnk},t-x-T_{SRI}+Q_{F\,mhi}D)$$
$$\cdot\left(\sum_{j=1}^{Q_{R\,hnk}} P_D(\phi_{PRNA},1,h,n,k,j,t-x-T_{SRI}+Q_{F\,mhi}D+Q_{F\,hnk}D)\right)$$
$$\cdot\left(\sum_{j=1}^{Q_{R\,mhi}} P_D(\phi_{SRIA},1,m,h,i,j,t-x-T_{SRI}+T_{PRN})\right)$$

- the probability that a previous PRNA message was discarded and the subsequent –SRIA message was successful,

$$Y_5 = \sum_{i=1}^{R_{mh}} \sum_{k=1}^{R_{hn}} \frac{1}{R_{mh}} \alpha'_{mnh}(t-x-T_{PRN}-Q_{F\,mhi}D-Q_{R\,mhi}D)$$
$$\cdot P_S(\phi_{SRI},0,m,h,i,Q_{F\,mhi},t-x-T_{PRN}-Q_{F\,mhi}D-Q_{R\,mhi}D)$$
$$\cdot\frac{1}{R_{hn}} P_S(\phi_{PRN},0,h,n,k,Q_{F\,hnk},t-x-T_{PRN}-Q_{R\,mhi}D)$$
$$\cdot\sum_{j=1}^{Q_{R\,hnk}} P_D(\phi_{PRNA},1,h,n,k,j,t-x-T_{PRN}-Q_{R\,mhi}D+Q_{F\,hnk}D)$$
$$\cdot P_S(\phi_{SRIA},1,m,h,i,Q_{R\,mhi},t-x-Q_{R\,mhi}D)$$

- the probability that a previous +SRIA message was discarded,

$$Y_6 = \sum_{i=1}^{R_{mh}} \frac{1}{R_{mh}} \alpha'_{mnh}(t-x-T_{SRI}) P_S(\phi_{SRI},0,m,h,i,Q_{F\,mhi},t-x-T_{SRI})$$
$$\cdot\sum_{k=1}^{R_{hn}} \frac{1}{R_{hn}} P_S(\phi_{PRN},0,h,n,k,Q_{F\,hnk},t-x-T_{SRI}+Q_{F\,mhi}D)$$
$$\cdot P_S(\phi_{PRNA},1,h,n,k,Q_{R\,hnk}D,t-x-T_{SRI}+Q_{F\,mhi}D+Q_{F\,hnk}D)$$
$$\cdot\sum_{j=1}^{Q_{R\,mhi}} P_D(\phi_{SRIA},1,m,h,i,j,t-x-T_{SRI}+Q_{F\,mhi}D+Q_{F\,hnk}D+Q_{R\,hnk}D)$$

- and the probability that all the MAP messages were successful but the subsequent IAM or ANM message was discarded.

$$
\begin{aligned}
Y_7 = &\left[\sum_{i=1}^{R_{mh}} \frac{1}{R_{mh}} \sum_{k=1}^{R_{hn}} \frac{1}{R_{hn}} \left[\alpha'_{mnh}\left(t - x - T_{IAM} - Q_{F\,mhi}D - Q_{F\,hnk}D - Q_{R\,hnk}D - Q_{R\,mhi}D\right)\right.\right. \\
&P_S\left(\phi_{SRI}, 0, m, h, i, Q_{F\,mhi}, t - x - T_{IAM} - Q_{F\,mhi}D - Q_{F\,hnk}D - Q_{R\,hnk}D - Q_{R\,mhi}D\right) \\
&P_S\left(\phi_{PRN}, 0, h, n, k, Q_{F\,hnk}, t - x - T_{IAM} - Q_{F\,hnk}D - Q_{R\,hnk}D - Q_{R\,mhi}D\right) \\
&P_S\left(\phi_{PRNA}, 1, h, n, k, Q_{R\,hnk}D, t - x - T_{IAM} - Q_{R\,hnk}D - Q_{R\,mhi}D\right) \\
&\left.\left.P_S\left(\phi_{SRIA}, 1, m, h, i, Q_{R\,mhi}, t - x - T_{IAM} - Q_{R\,mhi}D\right)\right]\right] \\
&\left[\sum_{i=1}^{R_{mn}} \sum_{j=1}^{Q_{F\,mni}} \frac{1}{R_{mn}} P_D\left(\phi_{IAM}, 0, m, n, i, j, t - x - T_{IAM}\right) + \sum_{i=1}^{R_{mn}} \frac{1}{R_{mn}} P_S\left(\phi_{IAM}, 0, m, n, i, Q_{F\,mni}, t - x - T_{IAM}\right)\right. \\
&\left.\sum_{j=1}^{Q_{R\,mni}} \int_0^{t-y} P_D\left(\phi_{ANM}, 1, m, n, i, j, t - x - T_{IAM} + y + Q_{F\,mni}D\right) q(y)\,dy\right]
\end{aligned}
$$

The transmission rate of SRI messages from region m to region n is then given by

$$\lambda_{SRI\,mn}(t) = \sum_{i=1}^{M} \alpha'_{min}(t) \qquad \textbf{(6-15)}$$

and $\lambda_{SRI\,mnz}(t) = (1/R_{mn})\lambda_{SRI\,mn}(t)$, where $1 \le z \le R_{mn}$.

The number of PRN messages generated at time t is

$$\lambda_{PRN\,mn}(t) = \sum_{k=1}^{M} \sum_{i=1}^{R_{km}} \frac{1}{R_{km}} \alpha'_{knm}\left(t - Q_{F\,kmi}D\right) P_S\left(\phi_{SRI}, 0, k, m, i, Q_{F\,kmi}, t - Q_{F\,kmi}D\right) \qquad \textbf{(6-16)}$$

and $\lambda_{PRN\,mnz}(t) = (1/R_{mn})\lambda_{PRN\,mn}(t)$, where $1 \le z \le R_{mn}$,

while the transmission rate of PRNA messages in response to successfully received PRN messages is given by

$$\lambda_{PRNA\,mnz}(t) = \frac{1}{R_{nm}} \lambda_{PRN\,nm}\left(t - Q_{F\,nmi}D\right) P_S\left(\phi_{PRN}, 0, n, m, i, Q_{F\,nmi}, t - Q_{F\,nmi}D\right) \qquad \textbf{(6-17)}$$

where $z = \Gamma(n,m,i)$ and $1 \le i \le R_{mn}$.

The transmission rate of +SRIA messages due to successfully completed provide roaming number procedures is given by

$$
\begin{aligned}
\lambda_{+SRIA\,mnz}(t) = &\sum_{j=1}^{M} \sum_{k=1}^{R_{mj}} \frac{1}{R_{nm}} \alpha'_{njm}\left(t - Q_{F\,nmi}D - Q_{F\,mjk}D - Q_{R\,mjk}D\right) \\
&\cdot P_S\left(\phi_{SRI}, 0, n, m, i, Q_{F\,nmi}, t - Q_{F\,nmi}D - Q_{F\,mjk}D - Q_{R\,mjk}D\right) \\
&\cdot \left(\frac{1}{R_{mj}} P_S\left(\phi_{PRN}, 0, m, j, k, Q_{F\,mjk}, t - Q_{F\,mjk}D - Q_{R\,mjk}D\right) P_S\left(\phi_{PRNA}, 1, m, j, k, Q_{R\,mjk}, t - Q_{R\,mjk}D\right)\right)
\end{aligned}
\qquad \textbf{(6-18)}
$$

while the transmission rate –SRIA messages due to the expiry of T_{PRN} is given by

$$\lambda_{-SRIAmnz}(t)=\sum_{j=1}^{M}\sum_{k=1}^{R_{mj}}\frac{1}{R_{nm}}\alpha'_{njm}\left(t-Q_{F\,nmi}D-T_{PRN}\right)P_S\left(\phi_{SRI},0,n,m,i,Q_{F\,nmi},t-Q_{F\,nmi}D-T_{PRN}\right)$$
$$\times\left(\sum_{l=1}^{Q_{F\,mjk}}\frac{1}{R_{mj}}P_D\left(\phi_{PRN},0,m,j,k,l,t-T_{PRN}\right)\right.$$
$$\left.+\frac{1}{R_{mj}}P_S\left(\phi_{PRN},0,m,j,k,Q_{F\,mjk},t-T_{PRN}\right)\sum_{l=1}^{Q_{R\,mjk}}P_D\left(\phi_{PRNA},1,m,j,k,l,t-T_{PRN}+Q_{F\,mjk}D\right)\right) \quad \textbf{(6-19)}$$

where $z = \Gamma(n,m,i)$ and $1 \le i \le R_{mn}$ in the above equations.

The transmission rate of IAM messages between the originating MSC and terminating MSC is given by

$$\lambda_{IAM\,mn}(t)=\sum_{i=1}^{M}\sum_{j=1}^{R_{mi}}\sum_{k=1}^{R_{in}}\frac{1}{R_{mi}R_{in}}\alpha'_{mni}\left(t-Q_{F\,mij}D-Q_{F\,ink}D-Q_{R\,ink}D-Q_{R\,mij}D\right)$$
$$P_S\left(\phi_{SRI},0,m,i,j,Q_{F\,mij},t-Q_{F\,mij}D-Q_{F\,ink}D-Q_{R\,ink}D-Q_{R\,mij}D\right)$$
$$P_S\left(\phi_{SRIA},1,m,i,j,Q_{R\,mij},t-Q_{R\,mij}D\right) \quad \textbf{(6-20)}$$
$$P_S\left(\phi_{PRN},0,i,n,k,Q_{F\,ink},t-Q_{F\,ink}D-Q_{R\,ink}D-Q_{R\,mij}D\right)$$
$$P_S\left(\phi_{PRNA},1,i,n,k,Q_{R\,ink},t-Q_{R\,ink}D-Q_{R\,mij}D\right)$$

and $$\lambda_{IAM\,mni}(t)=\frac{1}{R_{mn}}\lambda_{IAM\,mn}(t) \qquad \text{where } 1 \le z \le R_{xy}.$$

The transmission rates of ACM, ANM, REL and RLC messages at time t can be obtained from equations (4-4) to (4-7).

The success rate of the send routing information procedure or the roaming number retrieval success rate for MSCs in region m is given by

$$S_{RN\,m}(t)=\sum_{i=1}^{M}\sum_{j=1}^{R_{im}}\lambda_{+SRIAimj}\left(t-Q_{F\,imj}D\right)P_S\left(\phi_{SRIA},0,i,m,j,Q_{F\,imj},t-Q_{F\,imj}D\right) \quad \textbf{(6-21)}$$

while the call completion rate can be obtained from (4-9).

6.2.3 Simple Location Update Protocol

In the Simple Location Update protocol no ISD, ISDA or –ULA messages are generated. The arrival rate of reattempts in (6-1) is therefore given by the following:

$$\beta'_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}[\beta_{mkn}(t-T_{UL})X]$$
$$\beta''_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}[\beta'_{mkn}(t-T_{UL})X].$$
$$\beta'''_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}[\beta''_{mkn}(t-T_{UL})X]$$

where

$$X=\sum_{j=1}^{Q_{F\,mni}}P_D(\phi_{UL},0,m,n,i,j,t-T_{UL})+P_S(\phi_{UL},0,m,n,i,Q_{F\,mni},t-T_{UL})$$
$$\cdot\sum_{j=1}^{Q_{R\,mni}}P_D(\phi_{ULA},1,m,n,i,j,t-T_{UL}+Q_{F\,mni}D+Q_{R\,mni}D+Q_{F\,mni}D)$$

Equations (6-5) to (6-8) are also reduced to the following

$$\lambda_{ISDAmnz}(t)=\lambda_{ISDnmi}(t)=0 \qquad \textbf{(6-22)}$$

$$\lambda_{+ULAmnz}(t)=\frac{1}{R_{nm}}\lambda_{ULnm}(t-Q_{F\,nmi}D)P_S(\phi_{UL},0,n,m,i,Q_{F\,nmi},t-Q_{F\,nmi}D) \qquad \textbf{(6-23)}$$

$$\lambda_{-ULAmnz}(t)=0 \qquad \textbf{(6-24)}$$

6.2.4 *Super-Charged Location Update Protocol*

In a Super-Charged network the arrival rate of UL messages that require the transfer of subscriber profiles to VLRs in region *m* from HLRs in region *n* is given by

$$\lambda_{UL\,mn}(t)=\delta\sum_{k=1}^{M}\beta_{mkn}(t)+\sum_{k=1}^{M}\beta'_{mkn}(t)+\sum_{k=1}^{M}\beta''_{mkn}(t)+\sum_{k=1}^{M}\beta'''_{mkn}(t) \qquad \textbf{(6-25)}$$

where the arrival rate of reattempts in (6-25) is given by

$$\beta'_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{\delta}{R_{mn}}[\beta_{mkn}(t-T_{UL})X_1+\beta_{mkn}(t-T_{ISD}-T_{3211}-Q_{F\,mni}D-Q_{R\,mni}D)X_2]$$
$$\beta''_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}[\beta'_{mkn}(t-T_{UL})X_1+\beta'_{mkn}(t-T_{ISD}-T_{3211}-Q_{F\,mni}D-Q_{R\,mni}D)X_2]$$
$$\beta'''_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}[\beta''_{mkn}(t-T_{UL})X_1+\beta''_{mkn}(t-T_{ISD}-T_{3211}-Q_{F\,mni}D-Q_{R\,mni}D)X_2]$$

and X_1 and X_2 are obtained from equations (6-2) and (6-4), respectively. The arrival rate of the subsequent messages is given by equations (6-5) to (6-8).

The arrival rate of SUL messages from VLRs in region *m* to HLRs in region *n* is given by

$$\lambda_{SUL\,mn}(t)=(1-\delta)\sum_{k=1}^{M}\beta_{mkn}(t)+\sum_{k=1}^{M}\hat{\beta}'_{mkn}(t)+\sum_{k=1}^{M}\hat{\beta}''_{mkn}(t)+\sum_{k=1}^{M}\hat{\beta}'''_{mkn}(t) \qquad \textbf{(6-26)}$$

where

$$\hat{\beta}'_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}\left[(1-\delta)\beta_{mkn}(t-T_{UL})\hat{X}\right]$$
$$\hat{\beta}''_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}\left[\hat{\beta}'_{mkn}(t-T_{UL})\hat{X}\right]$$
$$\hat{\beta}'''_{mkn}(t)=\sum_{i=1}^{R_{mn}}\frac{1}{R_{mn}}\left[\hat{\beta}''_{mkn}(t-T_{UL})\hat{X}\right]$$

and

$$\hat{X}=\sum_{j=1}^{Q_{F\,mni}}P_D(\phi_{UL},0,m,n,i,j,t-T_{UL})+P_S(\phi_{UL},0,m,n,i,Q_{F\,mni},t-T_{UL})$$
$$\cdot\sum_{j=1}^{Q_{R\,mni}}P_D(\phi_{ULA},1,m,n,i,j,t-T_{UL}+Q_{F\,mni}D+Q_{R\,mni}D+Q_{F\,mni}D) \quad \textbf{(6-27)}$$

while the arrival rate of SULA messages is given by

$$\lambda_{SULAmnz}(t)=\frac{1}{R_{nm}}\lambda_{SULnm}(t-Q_{F\,nmi}D)P_S(\phi_{UL},0,n,m,i,Q_{F\,nmi},t-Q_{F\,nmi}D). \quad \textbf{(6-28)}$$

Since no CL and CLA messages are transferred in Super-Charged networks the arrival rates of these messages is set to zero.

$$\lambda_{CLAmnz}(t)=\lambda_{CLmnz}(t)=0 \quad \textbf{(6-29)}$$

The location update success rate in a Super-Charged network is now given by the following where the probability of successful delivery of +ULA and SULA messages is determined:

$$S_{LUm}(t)=\sum_{i=1}^{M}\sum_{j=1}^{R_{im}}\lambda_{+ULAimj}(t-Q_{F\,imj}D)P_S(\phi_{ULA},0,i,m,j,Q_{F\,imj},t-Q_{F\,imj}D)$$
$$+\sum_{i=1}^{M}\sum_{j=1}^{R_{im}}\lambda_{SULAimj}(t-Q_{F\,imj}D)P_S(\phi_{ULA},0,i,m,j,Q_{F\,imj},t-Q_{F\,imj}D) \quad \textbf{(6-30)}$$

6.2.5 *Lightweight Location Lookup Protocol*

The simplified nature of the LiLLP protocol allows the GSM call delivery equations to be greatly simplified. The arrival rate of calls including reattempts is given by

$$\alpha'_{mnh}(t)=\alpha_{mnh}(t)+c\int_0^{\infty}[Y_1+Y_2+Y_3]w(x)dx \quad \textbf{(6-31)}$$

where Y_1 to Y_3 represent the following

- the probability that a previous SRI message was discarded,

$$Y_1=\frac{1}{R_{mh}}\sum_{i=1}^{R_{mh}}\alpha'_{mnh}(t-x-T_{SRI})\sum_{j=1}^{Q_{F\,mhi}}P_D(\phi_{SRI},0,m,h,i,j,t-x-T_{SRI})$$

- the probability that a previous +SRIA message was discarded,

$$Y_2 = \sum_{i=1}^{R_{mh}} \frac{1}{R_{mh}} \alpha'_{mnh}(t-x-T_{SRI}) P_S(\phi_{SRI},0,m,h,i,Q_{F\,mhi},t-x-T_{SRI})$$
$$\cdot \sum_{j=1}^{Q_{R\,mhi}} P_D(\phi_{SRIA},1,m,h,i,j,t-x-T_{SRI}+Q_{F\,mhi}D)$$

- and the probability that the SRI and +SRIA messages were successful but the IAM or ANM message was discarded.

$$Y_3 = \left[\sum_{i=1}^{R_{mh}} \frac{1}{R_{mh}} \left[\alpha'_{mnh}(t-x-T_{IAM}-Q_{F\,mhi}D-Q_{R\,mhi}D)\right.\right.$$
$$P_S(\phi_{SRI},0,m,h,i,Q_{F\,mhi},t-x-T_{IAM}-Q_{F\,mhi}D-Q_{R\,mhi}D)$$
$$\left.\left.P_S(\phi_{SRIA},1,m,h,i,Q_{R\,mhi},t-x-T_{IAM}-Q_{R\,mhi}D)\right]\right]$$
$$\left[\sum_{i=1}^{R_{mn}} \sum_{j=1}^{Q_{F\,mni}} \frac{1}{R_{mn}} P_D(\phi_{IAM},0,m,n,i,j,t-x-T_{IAM}) + \sum_{i=1}^{R_{mn}} \frac{1}{R_{mn}} P_S(\phi_{IAM},0,m,n,i,Q_{F\,mni},t-x-T_{IAM})\right.$$
$$\left.\sum_{j=1}^{Q_{R\,mni}} \int_0^{t-y} P_D(\phi_{ANM},1,m,n,i,j,t-x-T_{IAM}+y+Q_{F\,mni}D) q(y)dy\right]$$

Equations (6-16) to (6-20) are simplified to the following:

$$\lambda_{PRNAmnz}(t) = \lambda_{PRN\,mnz}(t) = 0, \tag{6-32}$$

$$\lambda_{+SRIAmnz}(t) = \sum_{j=1}^{M} \frac{1}{R_{nm}} \alpha'_{njm}(t-Q_{F\,nmi}D) P_S(\phi_{SRI},0,n,m,i,Q_{F\,nmi},t-Q_{F\,nmi}D), \tag{6-33}$$

$$\lambda_{-SRIAmnz}(t) = 0, \text{ and} \tag{6-34}$$

$$\lambda_{IAM\,mn}(t) = \sum_{i=1}^{M} \sum_{j=1}^{R_{mi}} \frac{1}{R_{mi}} \alpha'_{mni}(t-Q_{F\,mij}D-Q_{R\,mij}D)$$
$$\cdot P_S(\phi_{SRI},0,m,i,j,Q_{F\,mij},t-Q_{F\,mij}D-Q_{R\,mij}D)$$
$$\cdot P_S(\phi_{SRIA},1,m,i,j,Q_{R\,mij},t-Q_{R\,mij}D) \tag{6-35}$$

6.3 Numerical Results and Discussion

This section presents the results obtained for the transient analysis of the network shown in Figure 5.1. The system parameters used here are identical to those used in the earlier chapters but with the following additions:

- The location update request timer T_{3210} = 20 seconds, as defined in [ETSI GSM 04.08].
- The location update reattempt timer T_{3211} = 15 seconds, as defined in [ETSI GSM 04.08].
- The MAP insert subscriber data timer T_{ISD} = 15 seconds. The minimum allowed MAP timer value defined in [ETSI GSM 09.02] is used. Investigations of different timer values, which are less than T_{3210}, were found to have a negligible impact on the results obtained.
- The MAP provide roaming number timer T_{PRN} = 15 seconds. The minimum allowed MAP timer value defined in [ETSI GSM 09.02] is used, to minimise the time taken to establish a call.

- The MAP send routing information timer T_{SRI} = 20 seconds. Here a value that is slightly larger than T_{PRN} is used to allow for processing delays and propagation delays that may exist between the sending of a SRI message and the receipt of a –SRIA message. An analysis of the effect of larger T_{SRI} and T_{PRN} values on network performance was found to be small.

6.3.1 Single Discard Threshold

The simulation and analysis are started from a state of rest when no calls or location updates are present in the network. As in Section 5.3, fifty-percent of the new location update attempts from each VLR are due to subscribers who have arrived from VLRs in the same region and the other location updates are due to subscribers arriving from the other regions. In the analysis of the location management procedures, at $t = 0$ seconds new location update attempts are introduced at a rate of 60 attempts/sec in each region and the offered load is increased to 120 attempts/sec at $t = 1300$ seconds for a period of 500 seconds. In the analysis of the call delivery and call set-up procedures, at $t = 0$ seconds new calls are introduced at a rate of 70 calls/sec from each region and an overload of 50 calls/sec is introduced at $t = 1300$ seconds for a period of 500 seconds.

6.3.1.1 Location Management Performance

Figure 6-1 shows the predicted and simulated location update success rates for various numbers of ISD messages per location update attempt. When only one ISD message is required per location update attempt, the location update success rate peaks soon after the overload is introduced and again after the overload is removed. In all cases the location update success rate decreases as the volume of traffic due to reattempts increases. Eventually the aggressive nature of the location update reattempt procedure leads to the network traffic being dominated by UL messages. Likewise, the location cancellation success rate is also negatively impacted and the throughput of CL messages decreases after an initial peak (Figure 6-2). The rate of decrease in network performance is also sensitive to the number of ISD messages generated as the network becomes congested much sooner when more ISD messages are present and is more likely to remain congested after smaller periods of overload. These results also suggest that the use of high-speed ATM signalling links and SIGTRAN signalling links to transfer large subscriber profiles in a single ISD message would reduce the rate at which the network progresses towards the congested state, and possibly allow the network to return to the uncongested state after a brief overload. When the overload traffic is removed the location cancellation success rate decreases. This is due to the large number of ISD, ISDA and ULA messages that are generated by the backlogged location updates. The load generated by these messages is significantly greater than the volume of CL and CLA messages present and therefore a smaller number of CL messages are admitted into the STPs.

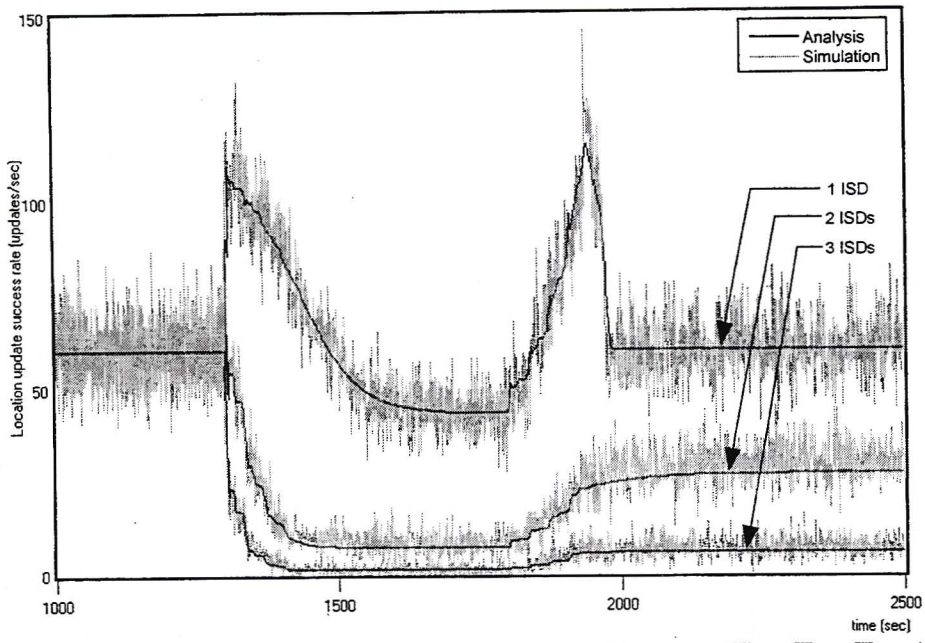


Figure 6-1. Location update success rate for each SP region ($K_1 = K_2 = K_3 = 100$).

The difference between the location update offered load and the location cancellation success rate can be used to determine how many subscriber profiles were not successfully deleted from the PVLR, as a function of time. This value can then be used to predict when a VLR's database utilisation will reach its maximum capacity.

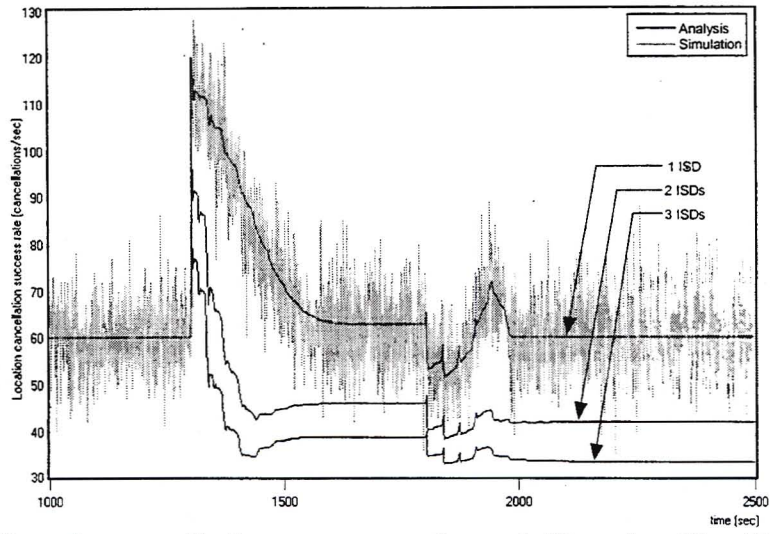


Figure 6-2. Location cancellation success rate for each SP region ($K_1 = K_2 = K_3 = 100$).

Figure 6-3 illustrates the rate at which the HLR is updated during the overload. The database update rate peaks at $t = 1300$ seconds but decreases soon thereafter as the resulting ISD and ISDA messages suppress the throughput of new UL messages. The net-effect is a reduction in the subsequent ISD load and an increase in the number of location update reattempts. When the overload traffic is removed a step-wise decrease in the HLR update rate observed in Figure 6-3. The step-wise decrease has a period of slightly greater than 30 seconds, which is the reattempt period between location update reattempts when the insert subscriber data procedure is unsuccessful and the –ULA message is successfully delivered to the originating VLR. Eventually, all the HLR updates are successful when only one ISD message is required per location update attempt and the location of between 1% and 5% of the subscribers is not known when 2 or 3 ISD messages are required per location update.

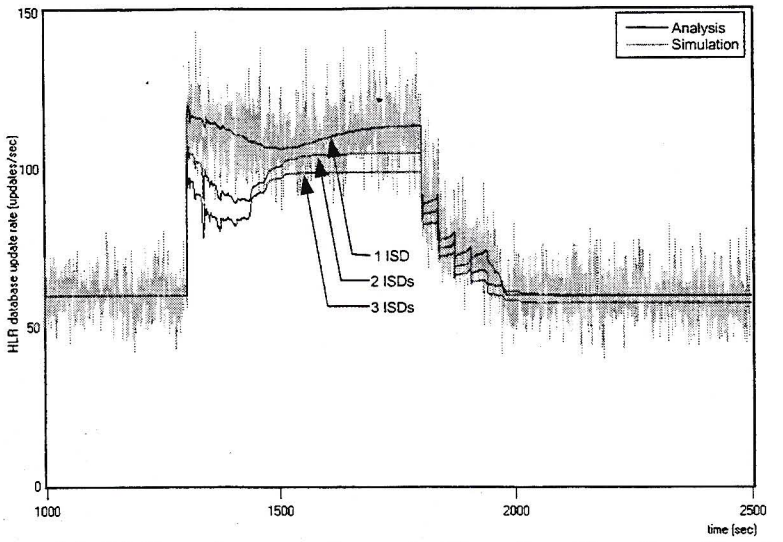


Figure 6-3. HLR update rate for each SP region ($K_1 = K_2 = K_3 = 100$).

6.3.1.2 Call Delivery and Call Set-up Performance

Figure 6-4 shows the predicted and simulated roaming number retrieval success rate and the call completion rate when a call overload is introduced into the network. The results are similar to those observed in Chapter 4 where the network remains in a sustained congestion state after experiencing a brief burst of calls. However, here signalling traffic is dominated by SRI messages rather than REL messages and the call completion rate is significantly higher once the overload traffic is removed.

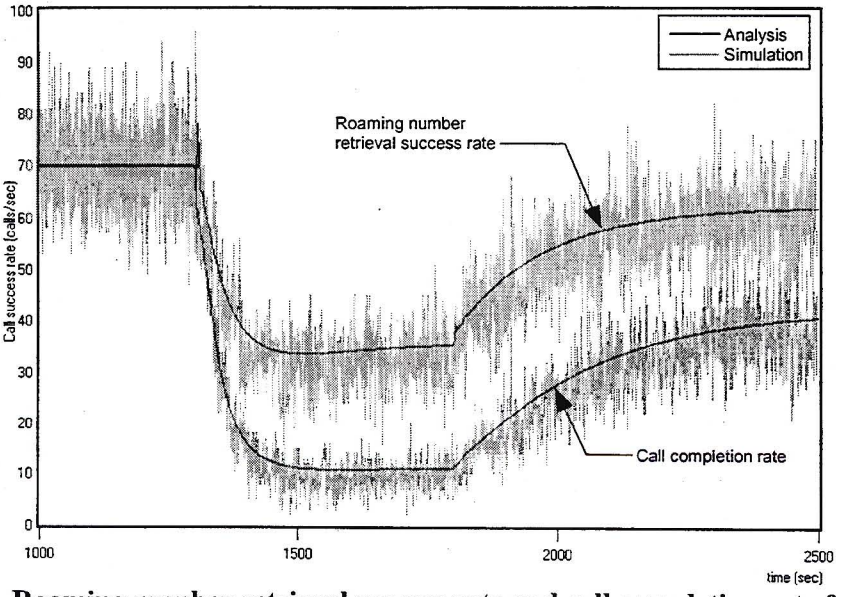


Figure 6-4. Roaming number retrieval success rate and call completion rate for each SP region ($K_1 = K_2 = K_3 = 100$)

Figure 6-5 shows the expected number of messages in a STP's processing buffer for the call overload scenario. Prior to the overload the message arrival rate at each STP is approximately 950 MSUs/sec, hence the buffer occupancy level fluctuates and occasionally approaches its maximum capacity. When the overload is introduced the queue is rapidly filled to its maximum capacity with SCCP messages and their resulting ISUP messages. The ISUP load peaks within the first 50 seconds of the overload and then begins to decrease as soon as the SCCP load increases due call reattempts. Only some of the attempts to obtain a roaming number are

successful and of these only a few calls are successfully established. The call failures trigger reattempts, which gradually increase the volume of SCCP messages (particularly SRI messages). The result is a decrease in the number of ISUP messages admitted into the STPs, until steady state is achieved or the overload is removed.

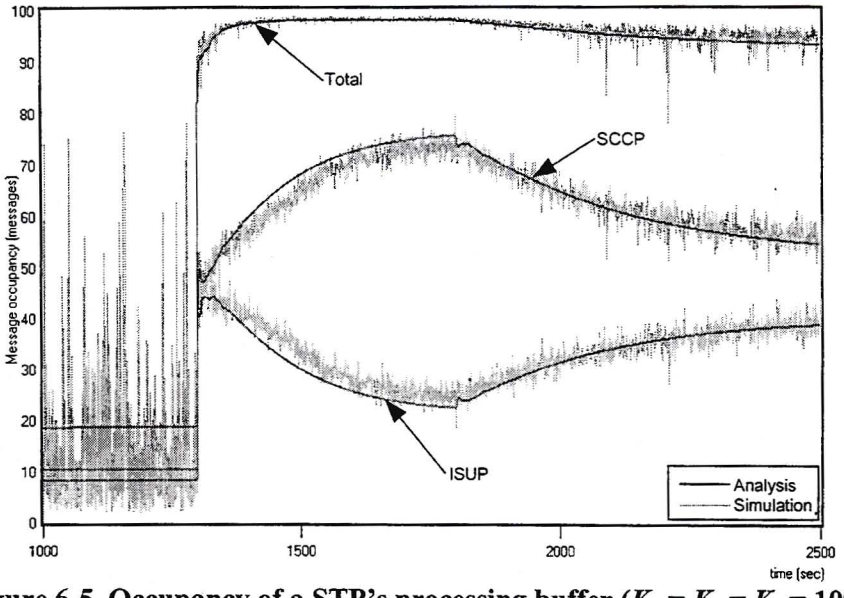


Figure 6-5. Occupancy of a STP's processing buffer ($K_1 = K_2 = K_3 = 100$).

6.3.1.3 The influence of call holding times on network performance

Call holding times in mobile networks were generally much shorter than those observed in PSTN networks. However, the call holding time in mobile networks appears to be increasing due to the decreasing cost of calls and mobile terminals in some countries. In measurements by Chlebus [1997], call holding times in mobile networks were approximately half of the call holding times observed in PSTN networks. In the author's own recent observations of mobile network call holding times in two countries, the mean call holding time in one country was found to be equal to ~90 seconds, while in the second country it was equal to ~150 seconds.

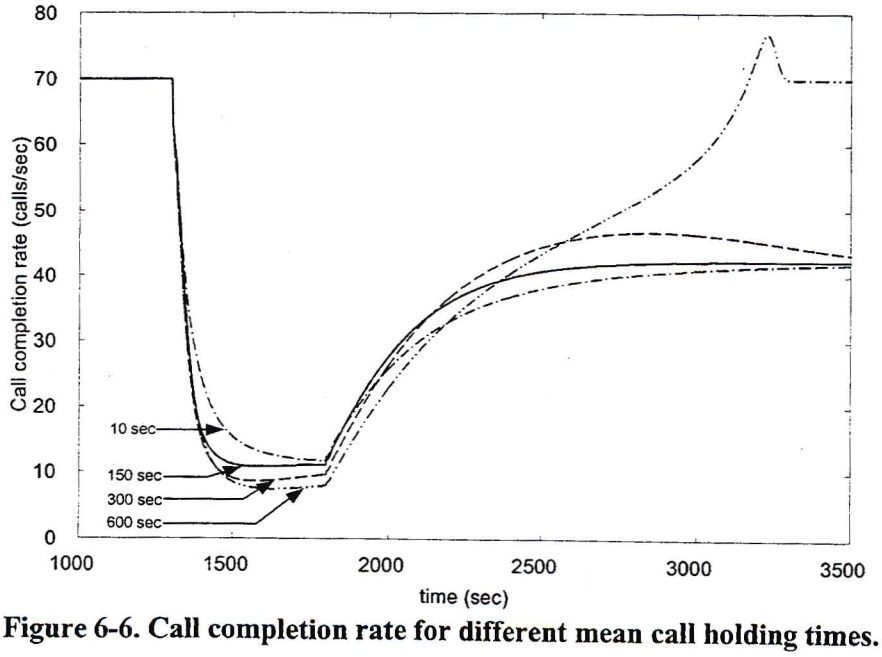


Figure 6-6. Call completion rate for different mean call holding times.

Figure 6-6 shows the influence of call holding times on the call completion rate of a single region. One of the most noticeable differences is a lower call completion rate soon after the overload is introduced for the higher call holding times. The call completion rate is influenced by the REL load that is present during the overload (Figure 6-7). These REL messages are generated by the normal termination of calls that were successful prior to the overload and the REL load is therefore sensitive to the mean call holding time. The REL load is also increased by reattempts from unsuccessful release procedures. The REL and RLC messages add to the network load and suppress the throughput of SCCP messages and hence the call completion rate is also reduced. Eventually the number of active calls in the network decreases and consequently the REL load also decreases and then the network tends towards the steady-state equilibrium solution.

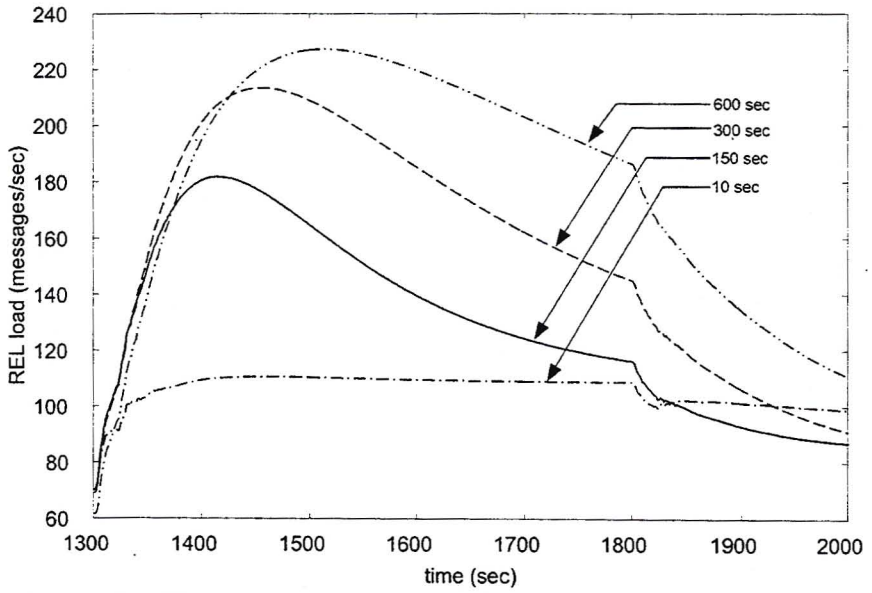


Figure 6-7. REL message load for different mean call holding times.

In the scenario where the mean call holding time is in the order of 500 seconds or greater, when the overload is removed the network drifts towards the uncongested state, as it has not approached equilibrium in the congested state. If the overload is maintained for a longer duration the call completion rate will gradually increase to the same steady state solution as that obtained with the other mean call holding times, and the network will remain congested when the overload is removed.

6.3.2 Multiple Discard Thresholds and Multiple User Part Interactions

This section examines the transient performance of a mobile network with message priorities and selective discarding during a regional overload when only the regional STPs are congested. The results obtained here are used as benchmarks to evaluate the effectiveness of the congestion control mechanisms examined in the next chapter. This is the typical type of overload experienced in a scenario where

- a VLR may have failed and a large number of location updates are initiated when it recovers, or
- a Base Station Controller that is responsible for coverage at a boundary of a location area has failed and mobile terminals that are able to receive spill-over radio coverage from an adjacent VLR's location area initiate location updates.

A background load of 30 location update attempts/sec and 30 calls/sec is initiated at $t = 0$ seconds in each region. At $t = 1300$ seconds the location update load in region-1 is increased by

90 attempts/sec for a period of 500 seconds. The mobility model used to generate the background traffic load assumes that 50% of the location updates from a region are due to subscribers who have moved between VLRs within the same region. The other 50% of the location updates are due to subscribers who have entered the region from VLRs in the other regions. Location updates that are part of the overload traffic stream are due to subscriber movements between VLRs in region-1 (i.e. the PVLR is always in region-1). Furthermore, calls are equally distributed between the MSCs.

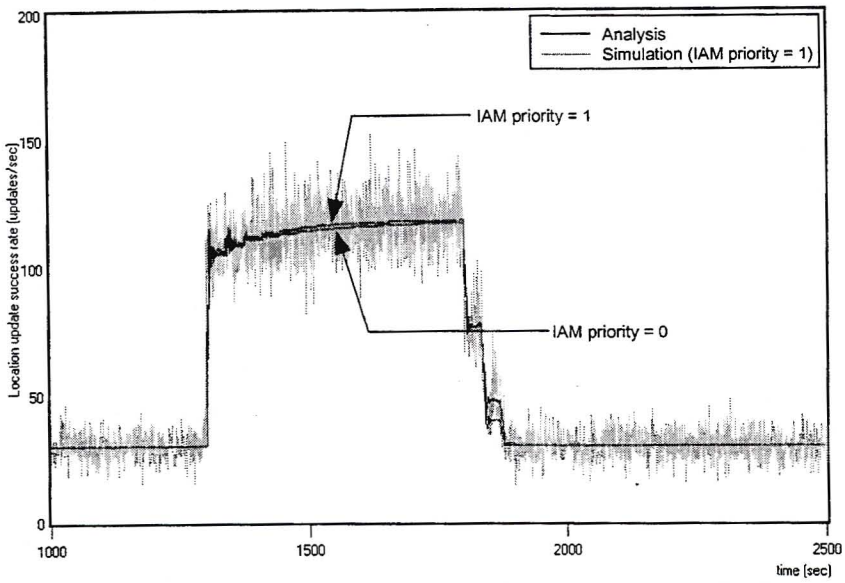


Figure 6-8. Location update success rate ($K_1 = 100$, $K_2 = 120$, $K_3 = 130$).

The MAP messages are assigned the following priorities: UL = 0, ISD = 1, ISDA = 2, ULA = 2, CL = 0, CLA = 1, SRI = 0, PRN = 1, PRNA = 1 and SRIA =2. The load generating messages (UL and SRI) are assigned the lowest priorities. The CL message is also assigned a priority of zero since the cancel location procedure is performed as an independent transaction on the HLR. The ISD and PRN messages are assigned a priority of one since they are triggered by linked transactions on the HLR. The analysis below considers the scenarios where IAM messages are assigned a priority of either 0 or 1 while the other ISUP messages are assigned the following priorities: ACM = 1, ANM = 2, REL = 1 and RLC = 2.

Figure 6-8 shows the location update success rate in region-1. The influence of the IAM message priority on location updates is negligible with more than 97.5% of the location updates being successfully completed before the overload is removed. The impact of the overload on the roaming number retrieval success rate and on the call completion rate for region-1 is shown in Figure 6-9. The traffic streams, roaming number retrieval success rate and the call completion rate oscillate with a period of ~36 seconds when the overload is introduced. This period is approximately equal to the sum of T_{SRI} and the mean call reattempt period. In the scenario where IAM messages are assigned a priority of zero the call completion rate decreases while the roaming number retrieval success rate increases as more failed calls are reattempted. In the scenario where IAM messages are assigned a priority of one almost none of the ISUP messages are discarded and therefore the call completion rate closely follows the roaming number retrieval success rate. An examination of the STP queue lengths also shows that congestion is limited to the STPs serving region-1 (where the mean queue lengths are close to K_1) while the other STPs are unaffected by the overload (as their mean queue length is less than 10 messages).

As in the previous section, these results also indicate that SCCP messages dominate in the use of the congested resources. This corresponds with the conclusions obtained by Mayer [1997] for a PSTN environment with multiple user parts. In his results one user part always dominates in the use of the congested facilities, while the throughput of the other user part drops to zero.

However, the findings here indicate that through the appropriate selection of message priorities and flow controls it is possible to ensure that minimal discarding is experienced by the ISUP messages to achieve an optimal throughput of calls. In the case of a GSM network environment it is recommended that only the UL, CL and SRI messages should be assigned the lowest priority, while all the other messages should be assigned non-zero priorities.

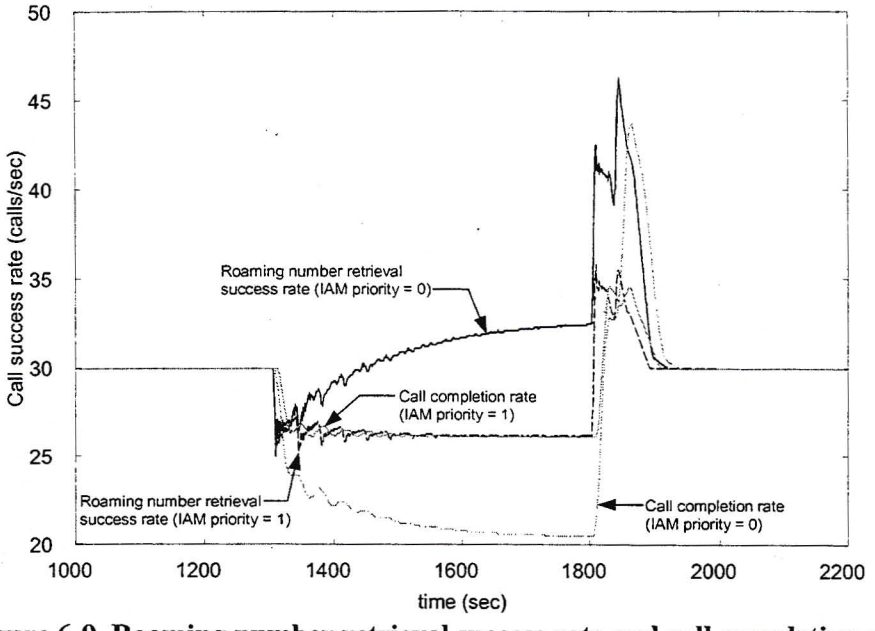


Figure 6-9. Roaming number retrieval success rate and call completion rate ($K_1 = 100, K_2 = 120, K_3 = 130$).

6.3.3 *Influence of Subscriber Profile Transfers on Super-Charged Networks*

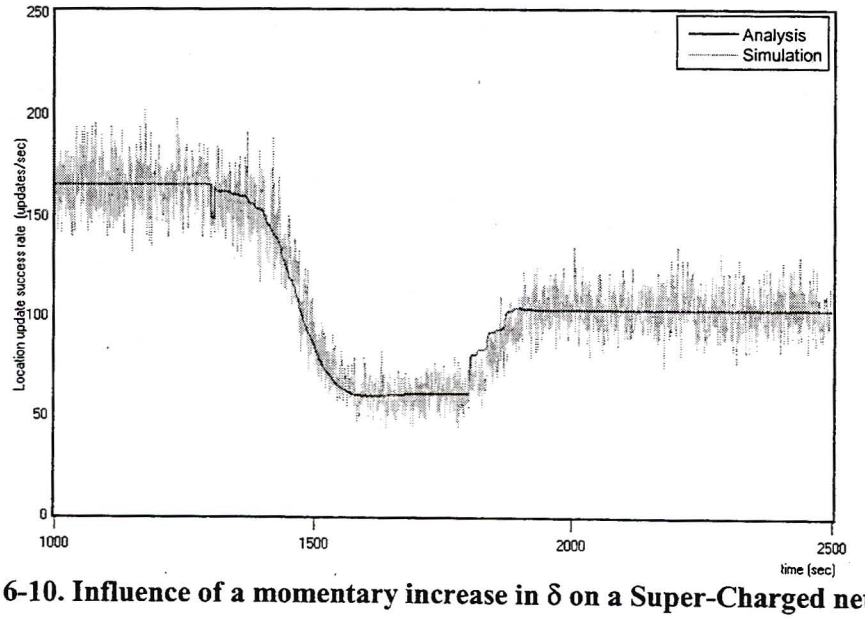


Figure 6-10. Influence of a momentary increase in δ on a Super-Charged network.

Signalling network congestion is possible in a Super-Charged network where sufficient reserve signalling capacity is not available. Various events including operator controlled subscriber profile changes, VLR database management procedures and HLR/VLR recoveries can provoke an increase in the number of location updates that require the transfer of subscribers' profiles.

Figure 6-10 shows the location update success rate for a scenario where the location update arrival rate is 165 attempts/sec in each region, two ISD messages are required to transfer a profile and $\delta = 0.3$ at the start of the analysis and simulation. At $t = 1300$ seconds δ is increased to 0.5 for a period of 500 seconds. The increase in the number of location updates that require the transfer of ISD messages leads to STP congestion, which persists even after δ returns to its normal level. In this case the network will only return to the uncongested state when either the offered load or δ is reduced to a lower value for a sufficiently long period.

6.4 Summary

Previous studies on mobility management have concentrated on the steady state behaviour of mobile networks, while those that do consider the time varying behaviour of mobility management protocols have not explicitly modelled the individual signalling message flows, protocols timers and procedures, or the impact of failures and congestion on protocol performance. The chapter extends the mathematical models derived in the earlier chapters to allow for the transient analysis of mobility management procedures, including the SLU protocol, the Super-Charged Location Update protocol and the LiLLP protocol.

Network traffic during congestion is dominated by SCCP messages (particularly UL and SRI messages). In the analysis of the location update success rate, the rate at which the network progresses into the congested state is mainly dependent of the number of ISD messages that are required to transfer subscriber profiles. Again, these results highlight the advantage of using high-speed ATM signalling links and SIGTRAN signalling links to transfer large subscriber profiles in a single ISD message.

In the analysis of the MAP call delivery and ISUP call set-up procedures the gradual build-up in the SCCP load is primarily responsible for suppressing the throughput of ISUP messages and the degradation in the call completion rate. The MAP timer values are found to have a small impact on network performance, instead network performance is more sensitive to the mean call holding time. REL messages from calls that were successful prior to the overload are responsible for the high REL load that is observed during the overload period. The REL load impacts the throughput of SCCP messages and therefore reduces the call completion rate obtained during the initial period of the overload. Networks with very large call holding times are also likely to return to the uncongested state after brief periods of overload as they take longer to reach the equilibrium state during congestion.

In the analysis of a regional overload where the STPs have multiple discard thresholds, the results indicate that with the appropriate selection of message priorities it is possible to ensure that very few ISUP messages are discarded and furthermore congestion does not propagate to the STPs that do not serve the affected region during low to moderate overloads.

7. Congestion and Flow Control in SS7

7.1 Introduction

Simple message discard schemes are obviously not the most effective flow control mechanisms, but they do provide a foundation over which more complex algorithms can be developed. The SS7 congestion and flow control mechanisms rely on choke packets and probe packets to convey congestion status information between the network entities. The process of relaying this information is a rather simple task, but the process of correctly detecting the congestion and responding appropriately to the congestion, in order to achieve optimum performance in all eventualities without further aggravating the situation, is a rather complex task.

Research on congestion and flow control techniques in packet switched networks has persisted since the `70s when computer networks first became popular. Research in this area still continues, as communication networks grow larger and are required to support a vast array of new services and applications. A number of studies have examined the applicability of link level flow controls as suitable flow control mechanisms in the event of STP processor congestion. Rumsewicz [1994] examined the influence of the TFC transmission frequency on the call completion rate in the NOCP. He concluded that the transmission of a TFC message for one out of every eight low priority messages received is appropriate. Northcote & Rumsewicz [1995] compared the IO and NOCP controls to a combined control (CC) scheme, which incorporates both the IO and NOCP. They found that the CC performed better than the weaker of the IO and NOCP and occasionally performed as well as the better of the two. Results obtained from studies that examine the effectiveness of the IO and NOCP during link level congestion could also apply equally well to processor congestion. Smith [1994], for example, found that the TFC procedure overcontrols traffic during congestion, whereas network latency leads to synchronisation of traffic from the source nodes. Zepf & Rufa's [1994] analysis of a sustained overload scenario showed that the IO and NOCP do not work correctly in networks with intelligent network services or a large number of users.

Previous work on SS7 congestion and flow controls has also primarily focused on ISUP traffic in a PSTN environment. Zepf & Rufa [1994] and Mayer [1997] considered SCCP traffic in a PSTN environment and Mayer points out that it is necessary for the SCCP and ISUP congestion controls to maintain a balanced relationship to ensure that one user part does not impede the throughput of the other. Studies on SS7 congestion and flow control in mobile networks are scarce. Nagarajan [1999], for example, examined signalling link congestion between a Base Station Controller and a MSC, where only SCCP traffic is present.

Congestion and flow control mechanisms that are initially designed and optimised for a particular network environment or overload scenario may not work as efficiently, or may completely fail, under alternate circumstances. This chapter evaluates the performance of the IO, the NOCP and NOWP during STP processor congestion in a multiple STP network. The overload scenarios considered include a network wide and focused overload in a PSTN environment, and a regional overload in a mobile network. Furthermore, various different implementations of the ISUP and SCCP congestion control mechanisms are also considered, as non-linear throttling schemes are found to perform better than the commonly used linear step-wise reduction schemes.

Even though the analysis in this chapter focuses on congestion and flow control mechanisms in a traditional signalling network, the results presented here are also applicable to signalling networks where IP and ATM are used as lower layer transport protocols. In a SIGTRAN based

network where the MTP Level 3 User Adaptation protocol [IETF RFC 3332] and/or the MTP Level 2 Peer/User Adaptation protocols [IETF RFC 3331] are utilised the MTP Level 3 congestion and flow control mechanisms would still be utilised for STP or signalling gateway congestion.

7.2 Implementation of SS7 Congestion and Flow Control Mechanisms

Section 1.9 described the operation of the MTP flow control mechanisms and the user part congestion control actions, together with a brief discussion of their limitations. Various aspects of the SS7 protocol, as defined in the ITU-T Recommendations, are stated as being implementation dependent. For example; ITU-T Recommendation Q.764, which addresses the ISUP congestion control procedures, states that the number of steps of traffic reduction and amount of increase/decrease of traffic at the various steps are considered to be an implementation matter. However, most studies that examined the performance of the IO have utilised a linear step-wise reduction/incremental procedure. Likewise ITU-T Recommendation Q.715 states that the SCCP and ISUP congestion control parameters should be selected to allow for co-ordination between them. This chapter therefore compares the performance of non-linear congestion control procedures to the commonly applied linear control schemes. In addition the NOWP is also examined, as there exists no published literature, to date, on the performance of this control scheme. The following list describes the operation of the ISUP congestion and flow control mechanisms that are examined in the subsequent sections. Section 7.4 describes how the ISUP congestion controls are adapted for implementation and synchronisation in SCCP.

IO_1: The congestion control method used for IO_1 is described in detail in Section 1.9.2. Upon receipt of a congestion indication (CI), ISUP reduces traffic to the affected destination by one step equal to '$1/X$', where X is the maximum number of reduction steps allowed. At step X all the traffic to the affected destination is blocked. When T30 expires, traffic to the affected destination is increased by one step (of size '$1/X$') and T30 is restarted. Here, ISUP does not distinguish between message types, and each message has an equal probability of being blocked. The IO flow control procedure described in Section 1.9.1.1 is implemented in controls IO_1 through to IO_6.

IO_2: Here, congestion control is identical to IO_1, except that only IAM messages are throttled. This congestion control mechanism is commonly used in studies that address the performance of the International Option.

IO_3: In this congestion control scheme, all the IAM messages to the affected destination are blocked when the first CI is received, and timers T29 and T30 are started. Congestion indications received while T29 is active are ignored. When T30 expires, the IAM message load is increased by '$1/X$' and T30 is restarted. This process continues until transmission of the full IAM message load has resumed. If a CI is received while only T30 is active, the entire IAM message load is again blocked and T29 and T30 are restarted.

IO_4: The congestion control mechanism implemented here is identical to IO_2 except that a T29 timer is not present.

IO_5: This congestion control is identical to IO_3, but the entire IAM message load is only blocked if ISUP is transmitting at its full capacity to the affected destination. If a CI is received while only T30 is active, the outgoing IAM load is only reduced by one step (i.e. '$1/X$'), if possible.

IO_6: In this scheme, congestion control is similar to IO_2, except that only '$1/(x + 1)$' of the total IAM message load to the affected destination is allowed into the network, for each reduction step x ($0 \leq x \leq X$). For example, when the first CI is received 50% of the IAM traffic is throttled. If another CI is received after T29 has expired and while T30 is active, the load is reduced by one more step, such that only 33.3% of the IAM traffic load is allowed into the network. At $x = X$ the entire IAM message load is blocked.

NOCP: In this scheme a TFC message is generated for every n messages received, at the congested STP, with a priority that is less than or equal to the current congestion status. At each SP, MTP level 3 maintains a congestion status counter for the affected destinations and informs the user parts of changes in the current congestion status, through CI messages. Upon receipt of a CI, ISUP throttles all traffic with a priority that is less than the current congestion status, to a specified destination. The SRSCT procedure, described in Section 1.9.1.2, is used to determine if congestion on the affected route set has abated.

NOWP: In this scheme, when the queue length of the STP's processing buffer reaches the congestion onset threshold, the congestion state is set to a predefined state s (where $1 \leq s \leq 3$) and the signalling points are informed of the congestion situation through TFC messages. The procedure used for subsequent updates of the congestion status is described in Section 1.9.1.3 and the flow control procedure implemented here is illustrated in Figure 1-17. The ISDN User Part has '$X = 3r$' load reduction steps, where r is a predetermined integer. Upon receipt of the first CI, the outgoing IAM message load is reduced by 'cr/X' of the total IAM message load, where c is the congestion level specified in the CI message. A congestion level of three will therefore block all the IAM messages to a given destination. Timers T29 and T30 are also started when a CI message arrives. Subsequent CI messages that arrive while T29 is active are ignored. If a CI arrives after T29 has expired, the outgoing IAM message load is reduced by 'cr/X' of the total IAM message load, only if the current outgoing IAM load is greater than or equal to the new load. When T30 expires, the outgoing IAM message load is increased by '$1/X$' and T30 is restarted.

Bernoulli trials are used in all of the congestion controls to select the throttled messages from the outgoing traffic stream. If an IAM message is blocked in the user part or the call fails due to message discarding, the call is later reattempted with a probability of 0.7. A blocked IAM message does not invoke the release procedure, however if a call fails because an IAM or ANM message was discarded at a congested STP, then the release procedure is initiated when the T_{IAM} timer expires. Blocked ANM messages also result in a failed call attempt and the release procedure is initiated when T_{IAM} expires. If a REL message is blocked or discarded and timer T5 hasn't expired, then the release procedure is reattempted when T1 expires.

Performance in the following investigation is measured in terms of the call completion rate, the percentage of successfully completed calls and the number of TFC messages generated. The call completion rate determines the number of calls successfully established, on receipt of an ANM message by the calling party. The percentage of successfully completed calls is calculated as follows:

$$Percentage\ success = \frac{\text{Call completion rate}}{\text{Outgoing IAM message load}} \times 100 \qquad \textbf{(7-1)}$$

This performance measure determines the probability with which an IAM message, that is not throttled, will result in a successful call. The number of TFC messages generated determines the

overhead imposed by the flow control mechanisms on the transmission resources, memory resources and network processing elements.

The simulation parameters used to model the congestion and flow control mechanisms described above are identical to those defined in Sections 3.4 and 4.3, together with the following additions (unless otherwise stated):

- The buffer thresholds for the NOCP flow controls are as follows (in terms of the number of messages in the queue): $A_1 = 70$, $O_1 = 75$, $D_1 = 100$, $A_2 = 100$, $O_2 = 105$, $D_2 = 120$, $A_3 = 120$, $O_3 = 122$ and $D_3 = 130$.
- The buffer thresholds for the IO and NOWP flow controls are as follows (in terms of the number of messages in the queue): A = 70, O = 75 and D = 100.
- Signalling messages are assigned the following lengths in octets: IAM = 48, ACM = 20, ANM = 18, REL = 22, RLC = 17, TFC = 15 and RCT = 15.
- Signalling messages are assigned the following discard priorities in the NOCP: IAM = 0, ACM = 1, ANM = 2, REL = 1, RLC = 2 and TFC = 2. The RCT messages are assigned a discard priority, at creation time, by the signalling route set congestion test procedure. Its priority is equal to one less than the currently known congestion status.
- The ISUP congestion control timers T29 and T30 are set to 0.3 seconds and 5.0 seconds, respectively. Ten reduction steps ($X = 10$) are used to throttle traffic in ISUP.
- The NOCP flow control timers are as follows: T15 = 3.0 seconds and T16 = 2.0 seconds. The maximum timer values, as defined in the ITU-T Recommendations, are used since the Recommendations state that the minimum values are for use on routes with long propagation delays (e.g. routes including satellites).
- The NOWP flow control timers are, Tx = 0.3 seconds and Ty = 0.3 seconds. When the onset threshold is exceeded the initial congestion state is $s = 3$. Three reduction steps ($r = 1$) are used to throttle traffic in ISUP, where each step represents one level of congestion. The timer values selected for Tx and Ty were chosen to be larger than the maximum expected end-to-end delay, since no suggested values are given in the ITU-T Recommendations and no previous studies have considered the NOWP. Section 7.5.2 examines the NOWP flow control timers in more detail.
- TFC messages are created for one out of every eight messages received ($n = 8$) at the congested buffer. New TFC messages are inserted at the end of the processing queue. When the congested buffer reaches its maximum capacity, TFC messages are not created until buffer resources are again available.

7.3 Comparison between the different Congestion and Flow control mechanisms

In order to compare the congestion and flow control mechanisms listed above in different overload scenarios the following scenarios were considered for the fixed network illustrated in Figure 3-4:

- Network Wide Overload: In this scenario a background first offered load of 107 calls/sec, is introduced into the network from each region at time $t = 0$ seconds. The first offered load is increased to 150 calls/sec at $t = 1300$ seconds and is maintained at this level for a period of 500 seconds and then returned to 107 calls/sec.
- Focused Overload with a single STP failure: In this scenario, STP-1 (refer to Figure 3-4) is disabled at the start of the simulation and a background traffic load of 50 calls/sec is introduced into the network from each region. At $t = 1300$ seconds the first offered load to region-1, from the other regions, is increased by 20 calls/sec for a period of 500 seconds. The effective new call arrival rate from each of the other regions to region-1 is therefore 28.571 calls/sec.

The results presented below are based on measurements taken during the last 200 seconds of the overload, i.e. from time $t = 1600$ seconds to $t = 1800$ seconds. Figures illustrating the call completion rate and message loads were obtained by taking a 20 second moving average of the measured data, in order to remove the high frequency components and to clearly illustrate the relative differences in performance between the various congestion and flow control mechanisms. Figures of the STP queue length and message occupancy were obtained by recording their average values over a one-second period, and the figures illustrating the fluctuation in a STP's congestion level show the actual changes in the congestion level over a small period of time.

7.3.1 Network Wide Overload

Table 7-1 lists the IAM and REL message loads arriving from region-1 for the various congestion and flow control mechanisms, together with region-1's call completion rate and the probability that an arbitrary call that is not throttled will be successful. Even though the data presented in the table is for SP region-1, similar results are obtained at the other regions. Table 7-2 lists the average buffer occupancy of STP-1 and the average number of TFC and RCT messages present in the buffer.

	IAM message load from region-1 (messages/sec)	*Call completion rate for region-1 (calls/sec)*	*Percentage success*	*REL message load from region-1 (messages/sec)*
IO_1	138.241	74.488	53.88	259.887
IO_2	129.256	122.374	94.68	137.483
IO_3	126.389	126.064	99.74	126.158
IO_4	130.567	123.507	94.59	134.616
IO_5	125.576	125.113	99.63	127.759
IO_6	128.424	125.118	97.43	133.862
NOCP	124.640	112.473	90.24	129.291
NOWP	119.315	118.941	99.69	123.182

Table 7-1. Effectiveness of the congestion and flow control mechanisms for a network wide overload.

IO_1 has the lowest call completion rate and only about half of the IAM messages admitted into the network result in successful call completions. The congestion control mechanism used here does not distinguish between message types and therefore all the messages have an equal probability of being throttled in ISUP. ANM messages that are blocked by the ISUP congestion controls result in failed call attempts, which consequently invoke the release procedure after the respective T_{IAM} timer expires and possibly a call reattempt. The throttling of REL and RLC messages leads to reattempts T1 seconds later and this action further increases in the REL message load. Soon after the overload is introduced, the STP processing buffers are filled to their maximum capacity within a short period of time. Messages are then discarded at the STPs and TFC messages are generated. The STPs continue to generate TFC messages, until the effect of the earlier TFC messages becomes apparent. This leads to overcontrol of the signalling traffic streams and the queue length of the STP buffers eventually drops to below 20 messages. Once the ISUP congestion control timers expire, the incoming message load again exceeds the network's maximum message handling capability. The STP buffers are again filled to their maximum capacity and the process repeats itself until the first offered load is reduced to an acceptable level. This oscillatory behaviour results in an average queue length that is approximately equal to half of the total buffer size.

	Average no. of TFC messages	*Average Queue Length*	*Average no. of RCT messages*
IO_1	3.444	45.776	-
IO_2	3.444	46.197	-
IO_3	0.355	16.148	-
IO_4	2.886	41.493	-
IO_5	0.718	21.839	-
IO_6	2.551	39.869	-
NOCP	2.395	29.157	0.434
NOWP	0.984	19.598	-

Table 7-2. Processing buffer utilisation in STP-1 during a network wide overload.

The IO_2 congestion controls are similar to IO_1, except that only IAM messages are throttled. A high call completion rate is obtained because none of the ANM messages are throttled in the user parts. Furthermore, fewer REL messages consume buffer resources when IO_2 is used. The explanation for the manner in which the STP buffer is occupied is similar to the explanation given for IO_1, above. As a result the average number of TFC messages generated and the average queue lengths, for both congestion control schemes, are almost identical. Figure 7-1 shows the buffer occupancy for one of the STPs. The buffer occupancy remains above the onset threshold until sufficient TFC messages have been generated to reduce the message load arriving at each STP to a reasonable level. During this period the queue length occasionally reaches the discard threshold and messages are discarded. Even though the congestion control procedure described in the ITU-T Recommendation Q.764 does not explicitly state that ISUP should only throttle IAM messages, IO_2 is widely accepted as being a better congestion control scheme than IO_1. An additional benefit gained here, is that by only throttling IAM messages a voice circuit does not have to be reserved for the call and the caller can be immediately informed of the network overload.

IO_4 performs slightly better than IO_2, by allowing slightly more IAM messages into the network (Table 7-1) and therefore a higher call completion rate is achieved. Fewer REL and TFC messages are also generated. IO_4 performs better, since its congestion controls are able to reduce the IAM message load sooner than the congestion controls in IO_2 and therefore fewer messages are discarded at the congested STPs. As a result the average queue length is also slightly lower. Here, IAM throttling at a SP is proportional to the number of TFC messages received. This is desirable since more TFC messages will be transmitted to the nodes that are generating the larger traffic loads, and hence they will reduce their IAM loads more severely. Fewer TFC messages are therefore required to reduce the IAM message load to the required level.

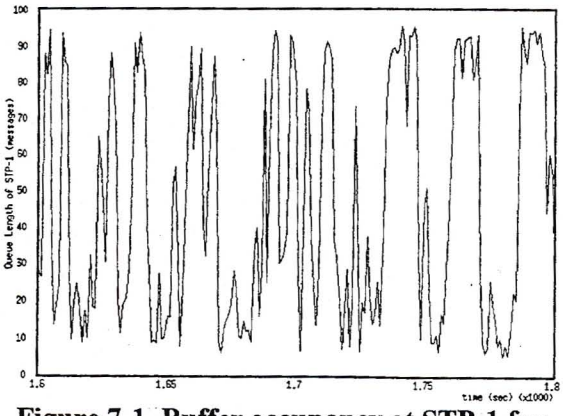


Figure 7-1. Buffer occupancy at STP-1 for IO_2.

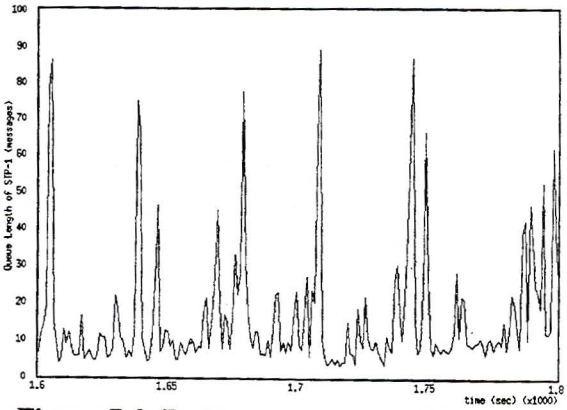


Figure 7-2. Buffer occupancy at STP-1 for IO_3.

IO_6 aims to reduce the number of TFC messages generated at the onset of congestion by throttling traffic severely when the first TFC message arrives and then by successively smaller steps for subsequent TFC messages. The effect is that slightly fewer TFC messages are required to obtain the desired effect and the STPs are in a congested state for a smaller period of time. Furthermore, the reintroduction of IAM traffic after the abatement of congestion is faster than when IO_2 or IO_4 are used, as fewer source-destinations pairs are throttled and fewer load incremental steps are necessary. When the congestion control timers expire, ISUP increases the IAM load more intensely with each incremental step.

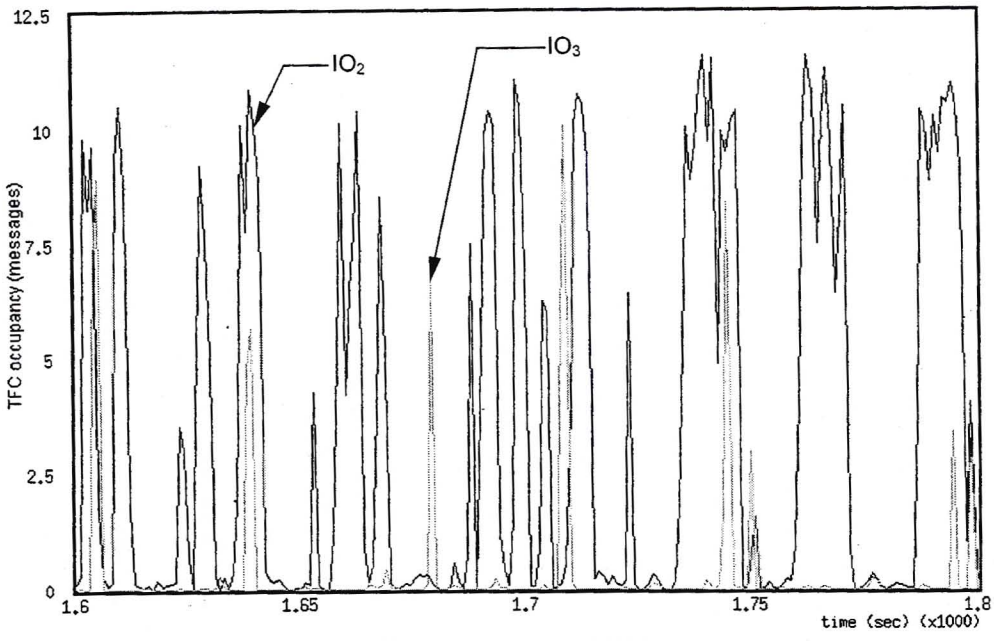


Figure 7-3. Number of TFC messages in a STP buffer (for IO_2 and IO_3).

The above results clearly indicate that a large number of TFC messages are generated when the queue length of a STP's processing buffer exceeds the congestion onset threshold. The STPs continue to generate TFC messages until the source nodes have reduced their loads to a tolerable level. However, in most cases the slow response time of the source nodes results in excessive discarding of signalling messages and consequently a large number of TFC messages are generated. Most of these TFC messages are also ignored if they arrive while T29 is active. In some instances, the sources are also slow to respond to the abatement of congestion. IO_3 and IO_5 attempt to reduce message discarding at the congested nodes, by throttling all the IAM messages to the affected destinations when the first TFC message arrives. Fewer TFC messages are therefore generated at the congested STPs, since the reduction in the IAM traffic load is apparent much sooner. Note; each TFC only refers to a single source-destination pair, therefore it is possible for a source to throttle all the IAM traffic to a particular SP and yet continue transmitting to the other SPs in the same region. The ISDN User Part will gradually reintroduce the IAM traffic load if no further CI messages are received from MTP level 3. In both congestion control schemes, an IAM message that is not throttled at the source has greater than a 99.6% probability of yielding a successful call. IO_3 and IO_5 also generate the smallest number of REL messages when compared to the other IO schemes. In both congestion controls fewer messages are discarded at the STPs and the IAM message load is more accurately controlled. Figures 7-1 and 7-2 show the time varying queue lengths of the IO_2 and IO_3 STP buffers, while Figure 7-3 shows a comparison between the number of TFC messages present in a STP processing buffer for both these control schemes. When the IO_3 congestion control is used, a large number of TFC messages are generated at approximately every 40 seconds, while with the IO_2 scheme TFC messages are usually present in the STPs' processing buffer. With the IO_3 and IO_5 congestion controls it is possible that only IAM traffic between some of the source-destination pairs will be throttled, while the other source-destination pairs continue to transmit

their entire load. However, when the traffic load from the throttled source-destination pairs increases to a level that again induces congestion in the STPs, then the resulting TFC messages will mostly likely be generated for the source-destination pairs that are still transmitting at their full loads (since their traffic loads are much higher).

The NOWP performs as well as the IO_3 and IO_5 control schemes, in terms of the call success rate. The congestion control implemented in this scheme is in principle similar to that used in IO_5 because all the IAM messages are blocked between some of the source-destinations pairs when the first few TFC messages arrive at their respective destinations, thereafter traffic is throttled in incremental steps. However, fewer IAM messages are produced in this case since the congested STPs continue to create TFC messages once the queue length drops well below the abatement threshold. The TFC messages continue to be created until the congestion status returns to zero. The NOWP therefore overcontrols the IAM message load. Furthermore, the source-destination pairs that are completely throttled are also slow to respond to congestion abatement. Hence, the call completion rate is also low.

The NOCP has the lowest call completion rate and probability of originating a successful call, when compared to the other control schemes (except for IO_1). This scheme performs poorly since ISUP either allows or blocks all the low priority traffic to the affected destination. There are no intermediate load reduction steps as defined in the IO and NOWP user part congestion controls. When the queue length of the STPs' processing buffer exceeds the first onset threshold TFC messages with a congestion status of one are generated. Occasionally the queue length exceeds the level two onset threshold and TFC messages with a congestion status of two are generated. The STP queue length rarely exceeds the level three onset threshold. If a TFC message arrives at a SP, all the messages to the affected destination with a priority lower than the specified congestion status are blocked. When an adequate number of messages have been blocked, the queue lengths of the affected buffers decrease rapidly to well below the level one abatement threshold. However, throttling continues until the SRSCT procedure detects that a route to the affected destination is no longer congested. A major drawback of the NOCP is synchronisation between the various source-destination traffic streams. When the onset threshold is exceeded TFC messages are generated for a large number of source-destination pairs. The traffic between the respective source-destination pairs is then throttled. Once the SRSCT timers for these source-destination pairs expire, all the SPs receive the same information as to whether congestion has subsided or not. If the STPs are no longer congested all the SPs resume transmission of the affected message steams at approximately the same time. This again induces congestion in the STP processing buffers, and TFC messages are again generated to most of the source-destination pairs, and the process then repeats itself.

Unlike Smith's [1994] investigation, where he attributes the traffic oscillations to network latency, the above results indicate that the coarse nature and operation of the NOCP congestion controls is responsible for the oscillations. The operation of the IO congestion control mechanism suggests that these oscillations could be suppressed if the throttled traffic is gradually reintroduced into the network.

The above results indicate that the NOCP is an inferior control scheme when compared to the IO_2 control scheme. This conclusion is also consistent with the conclusions obtained by other researchers for single STP networks ([Rumsewicz & Smith, 1995] and [Northcote & Rumsewicz, 1995]). Nevertheless, sections 7.3.2 and 7.5.1.2 show that this is not always the case.

7.3.2 Focused Overload with a single STP failure

A focused overload during a STP failure is the most likely scenario when a congestion situation is possible. This is also the type of scenario addressed by most researchers who examine

overloads in single STP networks (they basically ignore the actual network structure and assume that all the SPs are directly connected to the STP, which is the focus of the overload traffic). This section examines the validity of the conclusions acquired in those studies and also investigates the performance of the proposed control schemes. The results presented in this section and the subsequent sections do not include IO_1, since it invariably performs very poorly and is therefore not worth scrutinising any further.

	IAM message load from region-1 (messages/sec)	*Call completion rate for region-1 (calls/sec)*	*Percentage success*	*REL message load from region-1 (messages/sec)*
IO_2	37.384	32.685	87.43	45.980
IO_3	39.847	39.172	98.31	42.813
IO_4	36.399	31.901	87.64	44.680
IO_5	41.911	41.232	98.38	44.182
IO_6	38.261	34.842	91.06	45.305
NOCP	43.798	37.512	85.65	47.143
NOWP	39.507	38.345	97.06	42.103

Table 7-3. Effectiveness of the congestion and flow control mechanisms at region-1 during a focused overload.

Figure 7-5 shows the call completion rates for region-1 during a focused overload to this region. IO_2 has the lowest call completion rate relative to the other control schemes illustrated in the figure. Table 7-3 confirms the poor performance of the IO_2 control scheme, however slightly more calls are completed with IO_2 than with the IO_4 control scheme. Both control schemes perform poorly because the flow control mechanisms are unable to throttle the traffic from a large number of source-destination pairs. In this scenario there are 180 source-destination pairs of overload traffic streams. The IO_2 and IO_4 control schemes are unable to reduce the overload traffic fast enough, in order to minimise the number of messages discarded. However, rather than reducing traffic from the sources of the overload, they overcontrol the new call traffic originating from region-1. Figure 7-4 shows the time varying queue length of STP-2's processing buffer. Its average queue length usually lies above the abatement threshold during the overload period (Table 7-5), as the flow control mechanisms are unable to reduce the traffic load to within manageable limits for a long enough period. These results demonstrate the ineffectiveness of the IO_2 (and the IO_4) congestion controls to help alleviate congestion in the affected STP during a focused overload. These results correspond with the results obtained by Zepf & Rufa [1994] which show that the current congestion and flow control implementations do not work in an environment where the resources of the congested entity are shared between a large number of user parts.

The highest call completion rates and call success rates are obtained with the IO_3 and IO_5 control schemes (Tables 7-3 and 7-4). These congestion controls block the entire IAM message load to the affected destination in response to the initial congestion indications. The traffic loads are therefore reduced to within manageable limits soon after the first few TFC messages arrive at their respective destinations. As most of the messages arriving at STP-2 are from the overload traffic streams, most of the TFC messages are sent to the sources that contribute to the overload. Hence, traffic originating from region-1 is not severely throttled. IO_5 performs slightly better than IO_3, since IO_3 tends to overcontrol the IAM traffic. Both schemes have a number of desirable characteristics that demonstrate their efficiency; these include:

- a high call success probability and call completion rate in region-1, which implies that the most of the suppressed calls are from traffic streams that contribute to the overload.
- a small difference between the REL message load and the IAM message load, which indicates that very few REL messages are discarded at STP-2.

- a small average queue length, which indicates that the buffer occupancy in STP-2 does not frequently exceed the onset threshold nor does it remain congested for a long duration.
- a small number of TFC messages are generated, which signifies a low flow control overhead.

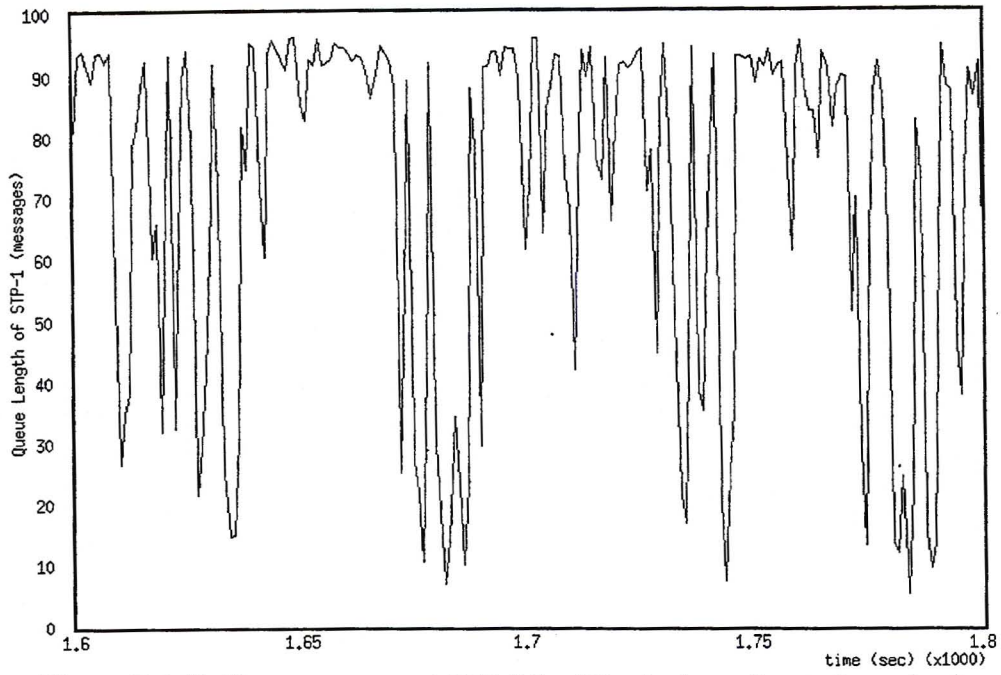


Figure 7-4. Buffer occupancy at STP-2 for IO_2, during a focused overload.

The results obtained for IO_6 are intermediate to the results obtained with IO_2 and IO_3, since IO_6 throttles the IAM traffic to region-1 severely in response to the initial TFC messages and then gradually reduces the level of throttling. IO_6 takes longer than IO_3 to reduce the overload traffic and STP-2 is therefore congested for a slightly longer period. However, the severe throttling in response to the initial TFC messages helps to maintain the average queue length below the abatement threshold. Nevertheless, IO_6 is not as effective as IO_3 or IO_5.

As in the previous section, the performance of the NOWP is comparable to the performance of the IO_3 and IO_5 controls. The NOWP allows slightly fewer IAM messages in to the network as it overcontrols traffic, by continuing to generate TFC messages for at least a period of '2×Ty' seconds after the queue length has dropped below the abatement threshold. With this scheme a large number of TFC messages are sent, to the sources of the overload traffic, when the onset threshold is crossed. But, once the queue length drops below the abatement threshold, TFC messages continue to be generated for messages arriving from region-1. These messages are usually ACM and ANM responses to the successful IAM messages of the overload traffic. These TFC messages therefore cause undesirable localised throttling.

The NOCP congestion and flow controls allow for a higher call completion rate in region-1 than the IO_2 controls (at least 75% of the background calls are completed compared to 65% with IO_2). However, more calls are successful in the entire network when the IO_2 controls are used. Nevertheless, it would be more desirable to allow calls from the background traffic stream to have a higher call completion rate rather than the calls that contribute to the overload. The priority assignment scheme employed in the NOCP also throttles the high priority messages when more severe congestion occurs in STP-2; hence the NOCP admits fewer REL messages into the network than the IO_2 scheme. As with the IO_3 and IO_5 control schemes, the IAM traffic load is reduced soon after the initial TFC messages arrive at their destinations, but the IAM load surges soon after the source nodes detect the abatement of congestion and STP-2 again becomes

congested. The NOCP is therefore more responsive than IO_2, but its performance can be improved if the blocked traffic is gradually reintroduced into the network.

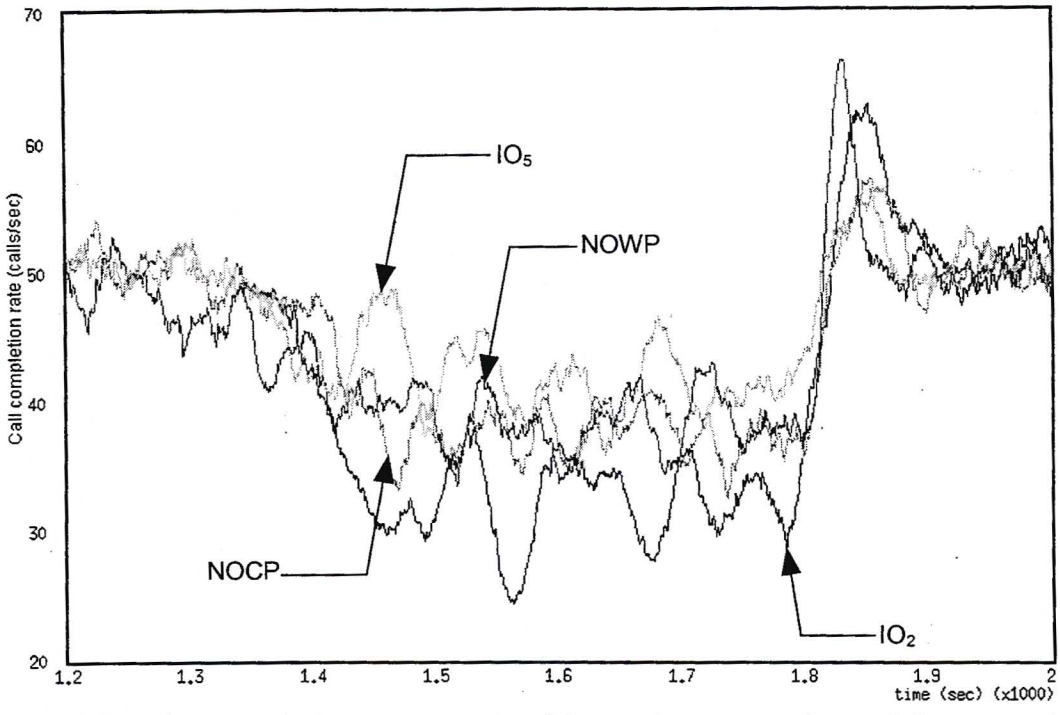


Figure 7-5. Call completion rate at region-1 for various congestion and flow control schemes, during a focused overload to this region.

The above results contradict the results obtained Northcote & Rumsewicz [1995] for a single STP network, where they show that the IO_2 scheme performs better than the NOCP. In their model an equal volume of messages, due to background calls, arrive from each SP. In addition there are only 50 source-destination pairs of overload traffic streams. In the model used here, most of the background calls carried by STP-2 originate from either region-1 or region-2 and there are 180 source-destination pairs of overload traffic streams. Hence, most of the messages from the background traffic stream that are discarded at STP-2 are either from region-1 or region-2, and therefore a large number of calls to and from these regions are failures. Furthermore, IO_2 is unable to reduce the traffic effectively for such a large number of source-destination pairs.

	IAM message load arriving into the network (messages/sec)	*Call completion rate for the entire network (calls/sec)*	*Percentage success*	*REL message load arriving into the network (messages/sec)*
IO_2	373.192	348.631	93.42	402.296
IO_3	362.103	359.517	99.29	363.483
IO_4	374.015	350.837	93.80	398.079
IO_5	364.793	362.320	99.32	366.488
IO_6	372.576	355.808	95.50	390.113
NOCP	375.266	346.586	92.36	383.601
NOWP	360.133	355.778	98.79	362.966

Table 7-4. Effectiveness of the congestion and flow control mechanisms in the entire network during a focused overload.

Manfield et al. [1993] remark that a RCT message may not necessarily travel on the route where the congestion was detected. In the overload scenario examined here, it is possible that TFC messages will be generated for traffic between region-2 and its non-adjacent regions. For example, a message from a SPx in region-5 to a SPy in region-2, via STP-2, might invoke a TFC message. SPx would then block all the IAM traffic to SPy. After T15 expires SPx will send a RCT message to SPy. If the RCT message is sent through STP-2 and it is still congested then SPx will receive another TFC message and it will restart T15. However, if STP-2 is congested and the RCT message is sent via STP-3 no subsequent TFC message will arrive before T16 expires and then SPx would resume transmission of the throttled traffic stream. Measurements taken during the last 200 seconds of the overload show that there are on average 0.011 RCT messages in the processing buffer of STP-3 (i.e. approximately 1% of the total number of RCT messages in STP-2). Some of these RCT messages are also due to source-destination pairs that exist between region-3 and region-1. The presence of very few RCT messages in STP-3 therefore implies that the impact of incorrect congestion information on calls to and from region-2 is very small. In fact, most of the RCT messages present in the network relate to traffic streams to and from region-1.

	Average no. of TFC messages	*Average Queue Length*	*Average no. of RCT messages*
IO_2	5.963	70.719	-
IO_3	0.737	21.433	-
IO_4	5.805	70.898	-
IO_5	1.334	27.892	-
IO_6	4.367	58.511	-
NOCP	3.481	42.830	1.017
NOWP	2.236	29.431	-

Table 7-5. Processing buffer utilisation in STP-2 during a focused overload.

Figures 7-6 and 7-7 show the transitions in the congestion status level, of various flow control mechanisms at STP-2, over a small period of time. Note, the changes in the congestion status levels are discrete integer values and there are no intermediate levels as shown in the figure. The interconnecting lines between the transition points are drawn merely for the ease of viewing. In the NOCP there is often a burst of IAM messages followed by a burst of ACM messages soon thereafter. The queue length of STP-2 therefore oscillates between the level two onset threshold and the level one discard threshold during congestion. The ACM messages arriving at the STP after a burst of IAM messages cause the queue length to exceed the level two onset threshold. Consequently, no further IAM messages are admitted into the queue and the arrival rate of the ACM messages decreases. The queue length then drops below the level one discard threshold and more IAM messages are admitted into the queue. These IAM messages are again followed by ACM messages and the queue length increases to above the level two onset threshold. This oscillatory process continues until sufficient source-destinations pairs have been throttled. Occasionally a burst of REL messages (due to the earlier discarding of IAM messages) will also cause the STP's queue length to increase to the level 2 discard threshold. The subsequent RLC traffic will cause the queue length to exceed the level 3 onset threshold.

Figure 7-6 shows that the congestion status of the NOWP flow control resides in states two and one for a period of Ty seconds. This indicates that once the queue length drops below the abatement threshold for Ty seconds the onset threshold is not exceeded until after the congestion status returns to state zero. Furthermore, unlike the other flow control methods, where the congestion status is updated when the buffer threshold settings are crossed, the NOWP's congestion status is not susceptible to oscillations due to random traffic fluctuations.

However, the MTP flow control timers have to be suitably large to prevent them from tracking the random traffic fluctuations.

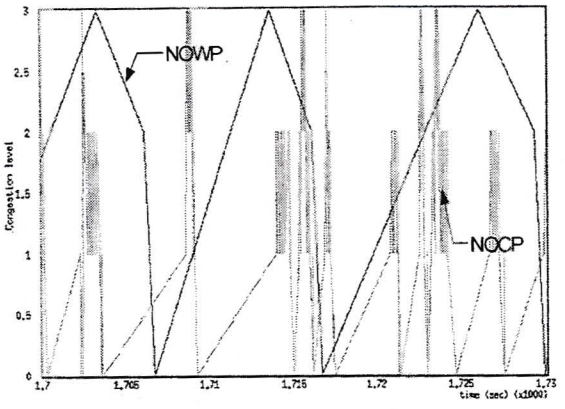


Figure 7-6. Congestion level in STP-2 for the NOCP and NOWP control schemes.

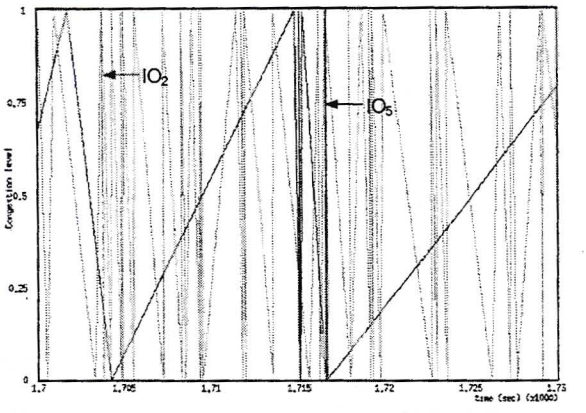


Figure 7-7. Congestion level in STP-2 for the IO_2 and IO_5 control schemes.

When the IO_2 control scheme is used the buffer occupancy regularly fluctuates from above the onset threshold to below abatement threshold. The IAM traffic streams that are only throttled by one step return to their full capacity within a few seconds and they again induce congestion. IO_5 is able to throttle the IAM loads for a longer period, before the traffic load to STP-2 again exceeds its message handling capability (Figure 7-7). STP-2 therefore becomes congested less often when either IO_3 or IO_5 are used.

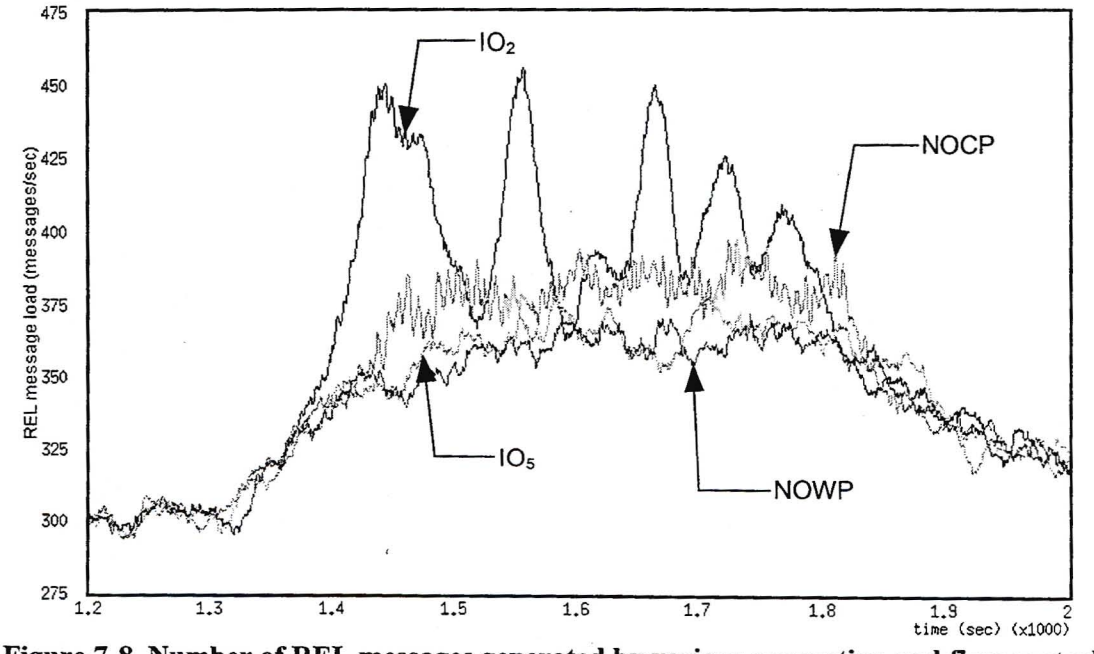


Figure 7-8. Number of REL messages generated by various congestion and flow control schemes, during a focused overload.

Figure 7-8 shows the total number of REL messages present in the network for various congestion and flow control mechanisms, during the focused overload scenario. The high discarding of IAM, ANM, REL and RLC messages at the congested STP aggravates congestion when the IO_2 control scheme is used. Discarding of these messages generates additional REL messages and eventually induces a REL message avalanche. These REL messages are also responsible for the high buffer occupancy observed in STP-2, as the TFC messages are unable to alleviate the REL message load. While the other STPs did not become congested, it is

possible that under alternate circumstances the REL avalanche could cause the congestion to propagate to other STPs.

The NOCP control scheme is also prone to REL message avalanches, but the REL load can be controlled. The traffic loads observed with the NOCP are therefore not excessively high as those observed with the IO_2 control scheme. However, when the NOCP is used, the traffic streams oscillate with a period of approximately 6 seconds. When STP-2 becomes congested, the queue length remains above the level one onset threshold for a period of approximately one second, before sufficient source-destination pairs have been throttled. Some of the REL messages are also discarded during this period and they are reattempted 5 seconds later. The queue length then drops below the level one abatement threshold and remains at this level for 5 seconds before the T16 timers and release procedure timers expire. The traffic load then increases and the process is repeated.

When the IO_3, IO_5 and NOWP control schemes are used, the REL message load follows the trend of the IAM message load. With these schemes, very few messages are discarded at STP-2 and therefore most of the REL messages can be attributed to normal call terminations.

7.4 Congestion Control in Multiple MTP User Parts

Call establishment in GSM relies on procedures within MAP to locate a called subscriber (i.e. the *send routing information* and *provide roaming number* procedures described in Section 1.8.2) and procedures within ISUP to eventually set-up the circuit switched path (Section 1.6). Since MAP uses SCCP to route its signalling messages and often utilises the same network resources as ISUP signalling messages, a balanced relationship is necessary between the SCCP and ISUP congestion control procedures to ensure an optimal throughput of calls and other services.

7.4.1 Implementation of SCCP Congestion Control

The SCCP congestion control mechanism assigns an importance value to each SCCP message type. These importance values are used to control the blocking of SCCP messages at the different Restriction Levels (RLs), whereby the messages with the lowest importance are blocked at RL = 1 and the messages with the highest importance are blocked at the highest RL. However, MAP only uses the SCCP Class 0 and Class 1 connectionless services to transfer messages. Furthermore, since MAP also performs the necessary procedures to segment large signalling messages into smaller messages all the signalling messages are transferred as SCCP Unitdata (UDT) messages.

The SCCP congestion control parameters (as described in Section 1.9.3) are implemented as follows to allow for synchronous operation with the ISUP congestion controls described earlier.

- The IO control schemes are implemented with 2 Restriction levels and 10 Restriction sublevels (N = 1 and M = 10).
- The NOCP is implemented with 4 Restriction levels and 1 Restriction sublevel (N = 3 and M = 1).
- The NOWP control scheme is implemented with 2 Restriction levels and 3 Restriction sublevels (N = 1 and M = 3).
- In the IO_2 to IO_6 and NOWP control schemes receipt of a TFC triggers throttling of UL and SRI messages in a MSC and CL messages in a HLR. These messages are blocked, as they are the initial messages of *independent* MAP transactions.

The simulation parameters specified in Section 7.2 are complemented by the following for the mobile network environment:

- MAP signalling messages are assigned the following lengths in octets: UL = 70, ULA = 57, CL = 55, CLA = 40, ISD = 279, ISDA = 80, SRI = 57, SRIA = 69, PRN = 70 and PRNA = 59.
- MAP signalling messages are assigned the following discard priorities in the NOCP: UL = 0, ULA = 2, CL = 0, CLA = 1, ISD = 1, ISDA = 2, SRI = 0, SRIA = 2, PRN = 1 and PRNA = 1.
- The SCCP timers Ta and Td are analogous to ISUP's T29 and T30 timers and are therefore implemented identically, with Ta = 0.3 seconds and Td = 5.0 seconds.
- Two ISD messages are required to transfer a subscriber profile. The remainder of the mobile network simulation parameters are identical to those listed in Sections 5.3 and 6.3.

7.4.2 Effectiveness of Congestion Controls in GSM Networks

The overload scenario examined in this section is identical to that examined in Section 6.3.2. In mobile networks there are always significantly more MSCs/VLRs compared to the number of HLRs. This analysis differs from the earlier sections, in that during a location update overload or call overload the overload traffic streams are always from the MSCs/VLRs to the HLRs. The HLRs also generate additional messages in response to these messages. In this analysis there are only 36 source-destination pairs of overload traffic streams through each of the affected STPs.

Table 7-6 shows the location management performance measures obtained for each of the congestion control mechanisms. Results for the IO_1 control scheme, which discards UDT messages irrespective of the type of MAP message contained therein, are not shown as it performs far worse than any of the other control schemes. The IO_5 control scheme was able to achieve the highest location update success rate and HLR update rate relative to the other control schemes. The main advantage of this scheme is again its ability to rapidly block traffic at the initial onset of congestion and then to apply a fine-grained congestion control to the subsequent traffic.

	UL Load (MSUs/sec)	*Location Update Success Rate (updates/sec)*	*HLR Updates (updates/sec)*	*Location Cancellation Success Rate (cancellations/sec)*
IO_2	95.670	89.660	38.990	41.305
IO_3	89.399	88.246	39.099	26.241
IO_4	95.591	90.064	42.325	39.557
IO_5	109.478	104.621	43.596	38.365
IO_6	97.084	89.823	40.493	13.877
NOCP	85.463	55.872	34.099	19.217
NOWP	78.404	77.739	35.064	18.970

Table 7-6. Performance of the MAP location management procedures in region-1.

The IO_3 control scheme does not perform as well since it blocks all the traffic soon after the initial onset of congestion and then gradually admits the UL traffic, which again triggers congestion and causes the source nodes to block all their traffic. The IO_3 control essentially overcontrols traffic from the small number of sources present. Messages that are generated while the congestion controls are active therefore experience a low probability of discard. This can also be observed by examining the total number of messages admitted into the network in Table 7-7 and the STP buffer utilisation in Table 7-8. Similarly, the NOWP and IO_6 control schemes also overcontrol traffic at the source nodes.

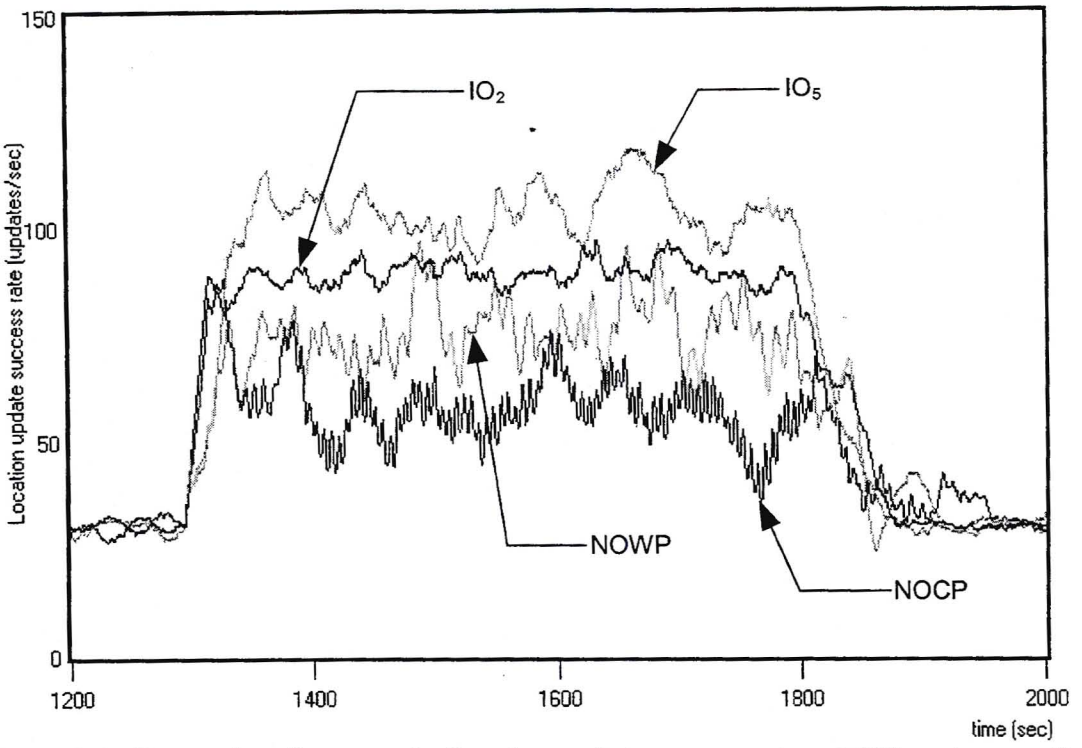


Figure 7-9. Comparison between the location update success rates of different congestion control schemes in region-1.

The lowest performance results are obtained with the NOCP. This is primarily due to overcontrol of signalling traffic at the source nodes, since the NOCP admits the smallest number of UL and SCCP messages (Table 7-6 and Table 7-7). Furthermore, the blocking of non-load generating messages such as ISD, ISDA and ULA messages also reduces the location update success rate. The mobility management traffic, the STP queue lengths and mobility management performance measures are all found to oscillate with a period of approximately 6 seconds (Figure 7-9). These oscillations are similar to those observed in Section 7.3.2. The source nodes take approximately one second to reduce the SCCP traffic to an acceptable level and five seconds later when the congestion control timers expire the STP again becomes congested and the cycle is repeated. Results for the scenario where IAM messages are assigned a priority of one are not shown as only a marginal improvement in network performance was achieved.

	SCCP Load (MSUs/sec)	*ISUP Load (MSUs/sec)*
IO_2	581.350	121.227
IO_3	525.640	110.704
IO_4	592.744	124.246
IO_5	625.478	116.424
IO_6	548.443	119.000
NOCP	414.862	83.443
NOWP	459.670	99.315

Table 7-7. SCCP and ISUP message loads from region-1.

Table 7-9 lists the performance measurements obtained for the MAP call delivery procedures and ISUP call set-up procedures. The performance of each of the congestion control mechanisms with respect to the roaming number retrieval success rate is similar to that obtained for the location update success rate. However, in ISUP the IO_3, IO_5 and NOWP congestion control mechanisms are responsible for blocking the largest proportion of IAM messages, while

the other control schemes block only a small fraction of IAM load. Various investigations with the IO_2 control scheme in different overload scenarios, including a call overload, indicate that with this scheme IAM messages always experience a negligible amount of blocking. An examination of the call completion rates and total number of ISUP messages present in the network suggests that ISUP messages are exposed to a negligible amount of blocking compared to SCCP traffic. Furthermore, an examination of the TFC traffic received by the MSCs indicates that most of the TFC messages include an HLR as the affected destination. This results in a greater throttling of traffic within the MSCs' SCCP, as most of the TFCs are ignored by ISUP, since it does not generate signalling traffic towards the HLRs.

The above observations led to the analysis of a network where the IO_5 control scheme was implemented in SCCP and the IO_2 control scheme was implemented in ISUP. The performance measures obtained for the MAP procedures where unchanged with the IO_5 control scheme while the percentage of IAM messages that were blocked was significantly reduced. As a result a 10% improvement in the call completion rate was achieved.

	Average no. of TFC messages	*Average Queue Length*	*Average no. of RCT messages*
IO_2	1.442	28.478	-
IO_3	0.318	12.043	-
IO_4	1.429	28.745	-
IO_5	0.920	25.100	-
IO_6	1.343	24.318	-
NOCP	1.642	19.045	0.044
NOWP	0.456	10.373	-

Table 7-8. Processing buffer utilisation in STP-1.

A comparison between the results obtained in this section with those obtained in Section 6.3.2, shows that a much higher location update success rate and call completion rate can be obtained with only simple message discard schemes in some overload scenarios. The congestion and flow controls are weak in low to moderate overload scenarios since the blocking of load generating traffic at the source nodes reduces the number of calls and location update attempts admitted into the network to below the network's maximum throughput capacity. In very high overload scenarios congestion and flow controls are still beneficial since they can prevent congestion from propagating throughout the network and they reduce the end-to-end message transfer delay.

	SRI Load (MSUs /sec)	*SRI Success Rate (calls/sec)*	*IAM Load (MSUs/sec)*	*Call Completion Rate (calls/sec)*	*% of blocked IAM messages*
IO_2	19.015	18.320	18.232	17.970	0.48
IO_3	18.640	18.522	16.980	17.049	8.33
IO_4	19.571	18.990	18.936	18.685	0.28
IO_5	24.502	23.911	20.695	20.335	13.45
IO_6	19.118	18.463	17.828	17.350	3.44
NOCP	16.527	13.015	12.429	11.103	4.50
NOWP	15.887	15.793	13.887	13.709	12.07

Table 7-9. Performance of the MAP call delivery and ISUP call set-up procedures in region-1.

These results indicate that the effectiveness of the SCCP control scheme is important to achieve a high SRI success rate and a correspondingly high call completion rate. Unlike the results obtained by Mayer [1997] where one user part dominates in the use of the congested resources while the other user part suffers; in GSM where SCCP messages are the dominant traffic type, congestion controls and message discarding have a relatively small impact on the throughput of ISUP messages. Instead, the ISUP call completion rate is dependent on the success rate of the MAP call delivery procedures. Similar conclusions are also obtained if the alternate case is also considered, where the location management traffic is low and a burst of calls originate from one region. Finally, these results indicate that the IO_5 control scheme or a combination of the IO_5 control scheme in SCCP and the IO_2 control scheme in ISUP should be employed in mobile networks to sustain a high throughput of location updates and calls during congestion scenarios.

7.5 Optimisation of the Congestion and Flow Controls

The following subsections focus on optimisation of the congestion and flow control parameters. Even though the analysis focuses on the PSTN environment similar conclusions are also obtained for some of these scenarios in a mobile network environment.

7.5.1 Optimising the performance of the flow control procedure

TFC messages consume network resources during congestion, when the available resources are scarce. To avoid reducing the traffic load too rapidly the ISUP congestion controls (as specified in ITU-T Recommendation Q.764) ignore congestion indications to an affected destination if T29 is active. Performance is therefore sacrificed if too many TFC messages are sent to a particular destination. These redundant TFC messages consume buffer resources at the expense of call set-up messages. It would therefore be desirable to reduce the number of redundant TFC messages present in the network, in order to make optimum use of the available resources. Sections 7.5.1.1 and 7.5.1.2 investigate the impact of the TFC transmission rate on the call completion rate. The network wide and focused overload scenarios examined here are identical to those examined in Sections 7.3.1 and 7.3.2.

7.5.1.1 The Impact of TFC Messages in a Network Wide Overload Scenario

Sections 7.5.1.1 and 7.5.1.2 only discuss the IO_2, IO_3, NOCP and NOWP control schemes, since the results obtained for the IO_4 and IO_5 control schemes are similar to those obtained for the IO_2 and IO_3 control schemes, respectively. The results obtained for the IO_6 control scheme are intermediate to those obtained for the IO_2 and IO_3 control schemes.

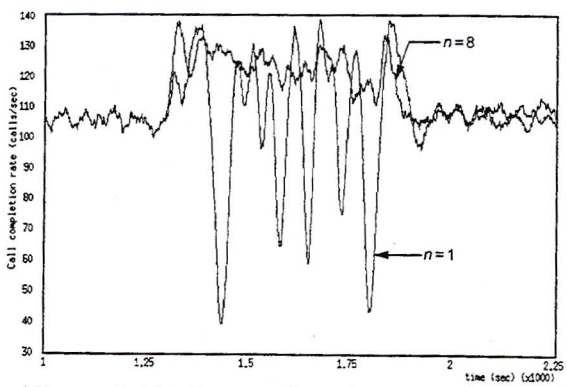


Figure 7-10. Comparison between the call completion rates obtained for TFC transmission intervals of $n = 1$ and $n = 8$ in the IO_2 control scheme.

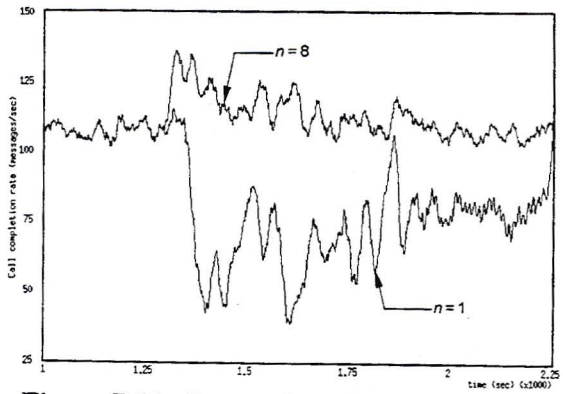


Figure 7-11. Comparison between the call completion rates obtained for TFC transmission intervals of $n = 1$ and $n = 8$ in the NOCP control scheme.

When a TFC message is transmitted for each signalling message received (i.e. $n = 1$), the TFC messages consume about half of the available buffer resources. Message discarding at the STPs is therefore increased since the queue length grows much faster than in the scenario where $n = 8$. Furthermore, the call completion rate oscillates and occasionally drops below 50 calls/sec. This oscillation is more prominent in the IO control schemes (see Figure 7-10) and has a period of between 50 and 70 seconds. The period of oscillation is usually an integer multiple of T30, and depends on how severely the IAM traffic loads are throttled. In the NOCP, network performance is severely degraded and congestion persists for about 450 seconds after the overload traffic is removed (Figure 7-11).

Table 7-10 lists the call completion rates for the different congestion and flow control schemes when different TFC transmission intervals are used. For a TFC transmission interval of $n = 20$, the call completion rate of the IO and NOCP control schemes decreases slightly, as the STPs take longer to reduce the IAM message loads to moderate levels. However, the performance of the NOWP control scheme improves when the TFC transmission rate is reduced, since the effect of overcontrolling the IAM message load with too many TFC messages is reduced.

ITU-T Recommendation Q.704 states that a TFC message should be sent for each message received with a priority that is less than the current congestion status, in the NOCP. However, the above results suggest that a TFC transmission interval of $n = 8$ is suitable and recommended for optimum performance of the NOCP and IO control schemes, but a higher TFC transmission interval is necessary for optimum performance of the NOWP.

	Call Completion rate for $n = 1$ (messages/sec)	*Call Completion rate for $n = 8$ (messages/sec)*	*Call Completion rate for $n = 20$ (messages/sec)*
IO_2	105.921	122.374	120.847
IO_3	102.783	126.064	125.232
NOCP	62.729	112.473	112.222
NOWP	103.695	118.941	126.059

Table 7-10. Call completion rates for different TFC transmission intervals at a single SP region.

7.5.1.2 The Impact of TFC Messages in a Focused Overload Scenario

The conclusions obtained in this section for the IO_3, NOCP and NOWP control schemes are comparable to those obtained for the network wide overload scenario. The IO_2, IO_3 and NOCP control schemes perform best when $n = 8$ and the NOWP's performance improves if higher TFC transmission intervals are used. However, when $n = 20$ the IO_2 control scheme fails and the entire network becomes congested at approximately 300 seconds after the overload is introduced. Figure 7-12 illustrates the call completion rates for some of these congestion and flow control schemes. For the period from $t = 1600$ seconds to $t = 1800$ seconds an average of ~3.5 calls/sec are successfully completed in region-1 and ~124.2 calls/sec are completed in the rest of the network, when IO_2 is used. This implies that less than half of the background traffic calls from the other regions are successful. When the simulation ended at $t = 2300$ seconds the network using the IO_2 control scheme was still in the congested state.

The IO_2 control scheme fails because not enough TFC messages are generated at the onset of congestion to reduce the IAM traffic load to a tolerable level. Furthermore, if a STP's processing buffer is filled to its maximum capacity, TFC messages from the other STPs are also discarded. One should also consider the following situation; if only a single buffer is available and a signalling message arrives, this message is allocated the buffer space. But, if this message is also the 20th message since the last TFC was created a TFC message will not be created in

this situation since no additional resources are available to retain the TFC. In addition, the REL avalanche also aggravates the congestion situation and prolongs the period of congestion.

The IO_3 and NOWP are able to avoid a similar fate because the few TFC messages that are generated at the onset of congestion are sufficient to reduce the IAM load to a reasonable level within a short period of time. In the NOCP the highest priority is assigned to the TFC messages, these messages therefore experience minimal discarding in the congested STPs. The sources are then able to throttle traffic, based on the current congestion status. Nevertheless, the control schemes that perform well at $n = 20$ could also fail if much lower TFC transmission rates are used.

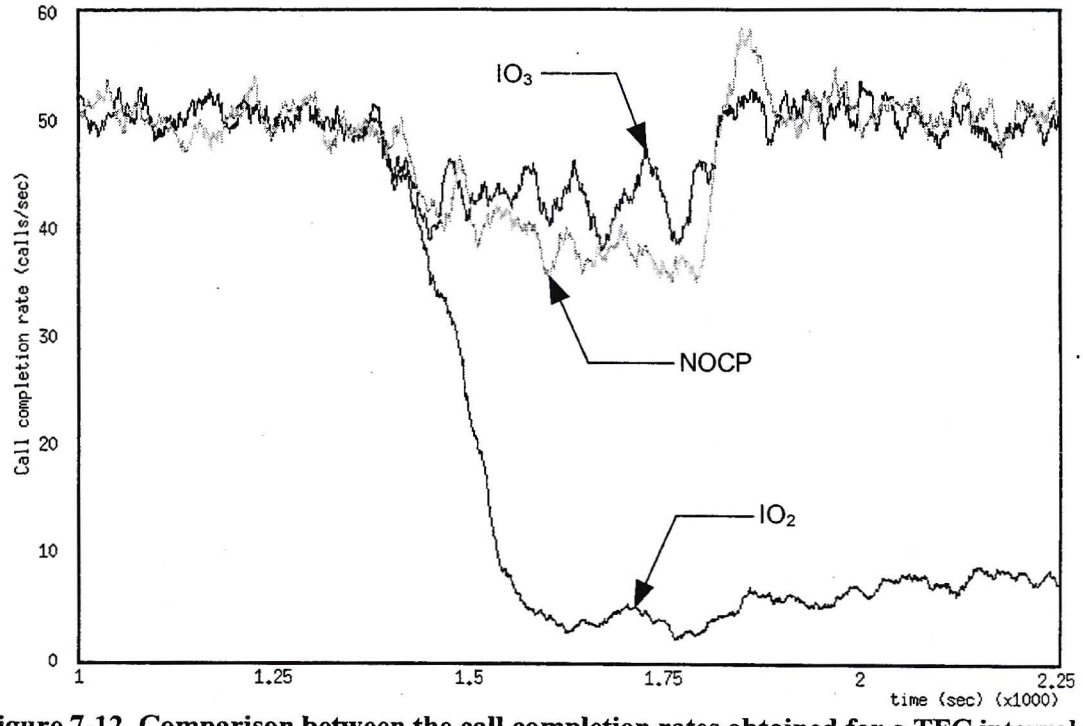


Figure 7-12. Comparison between the call completion rates obtained for a TFC interval of $n = 20$ in the IO_2, IO_3 and NOCP congestion and flow control schemes.

The above results indicate that the IO_2 control scheme is not robust over a wide range of overload and failure scenarios, especially in environments where the TFC messages are not sent very frequently or where they are susceptible to discarding at the transit nodes during severe congestion situations. The congestion control schemes that block the entire IAM load to the affected destination in response to the initial TFC message are found to be more robust and are able to maintain the integrity of the network during adverse conditions.

7.5.2 The Influence of Flow control timers on network performance in the National Option without Congestion Priorities

The ITU-T Recommendations do not specify the range of the NOWP flow control timers nor do they specify its initial congestion status. These parameters are considered as being implementation dependent issues. To date, there also exists no published literature on the performance of the NOWP during overload scenarios. This section therefore investigates the performance of the NOWP flow controls for various timer settings and initial congestion status values.

Table 7-11 lists the IAM message load, the call completion rate and the percentage of successful calls for various settings of the NOWP flow control parameters. These results pertain

to the focused overload scenario, but the same conclusions are acquired for the network wide overload scenario.

From the results below it is evident that the highest call completion rates are usually obtained with the smallest timer values. When Ty is very large the network is slow to respond to the abatement of congestion and thus continues to generate TFC messages, which overcontrol the IAM message load. If Tx is very large than it is possible that the congestion level will take too long, during severe overloads, to progress from status level one to status level three. A large number of TFC messages will therefore have to be generated during severe congestion scenarios, if congestion level three is not easily attained.

Initial congestion status 's'	*Tx (seconds)*	*Ty (seconds)*	*IAM message load from region-1 (messages/sec)*	*Call completion rate for region-1 (calls/sec)*	*Percentage success*
1	0.1	0.1	40.704	39.360	96.70
1	0.3	0.3	40.916	39.517	96.58
1	0.1	1.0	38.759	37.635	97.10
1	1.0	0.1	42.621	37.448	87.86
1	0.5	6.0	33.094	32.143	97.13
1	6.0	0.5	44.724	30.650	68.53
3	0.1	0.1	40.828	39.823	97.54
3	0.3	0.3	39.507	38.345	97.06
3	0.1	1.0	38.901	37.754	97.05
3	1.0	0.1	41.621	40.823	98.08
3	0.5	6.0	33.778	33.182	98.24
3	6.0	0.5	39.980	38.675	96.74

Table 7-11. Effectiveness of the NOWP flow control, in region-1, for various parameter settings.

For the control scheme where $s = 1$, a small Tx and Ty that are greater than the average end-to-end transfer delay are considered as the recommended timer settings. With these timer settings, the congestion status is updated to reflect the impact of the initial TFC messages on the incoming traffic streams. If the TFC messages were effective at reducing the load the congestion status is decremented, otherwise it is increased.

For the control scheme where $s = 3$, a small value of Ty is recommended while a Tx of ~1 second is recommended. As these flow controls are likely to overcontrol the source nodes with the initial TFC messages, a large Tx reduces the probability of the congestion status returning to level three due to random traffic fluctuations, unless the incoming load has not been adequately throttled by the previous TFC messages. Performance in the NOWP could be improved if the initial congestion state is a function of the traffic arrival rate, rather than the queue length.

7.5.3 Optimisation of the ISUP congestion controls

Other than employing alternate load reduction/incremental schemes, the ISUP congestion controls also allow for the selection of alternate timer values and load reduction/incremental step-sizes. Table 7-12 lists the call completion rates obtained for various load reduction/incremental step-sizes in a focused overload scenario.

When the number of reduction/incremental steps is reduced from ten to five (or from nine to three for the NOWP) an improvement in the call completion rate is apparent for all the congestion control schemes. The IO_2 and IO_4 congestion control schemes have a higher call completion rate because the IAM load is throttled much faster and fewer load reductions are necessary to obtain the desired effect. The call completion rates of the IO_3, IO_5 and NOWP congestion controls improves since traffic is reintroduced much faster in response to congestion abatement. In the IO_6 control scheme, changing the number of steps does not change the load reduction size at each step but instead it determines the step at which the IAM traffic load is completely blocked. Its call completion rate increases since fewer reduction steps are necessary to completely block the IAM load.

	Call completion rate for X = 5 (calls/sec)	*Call completion rate for X = 10 (calls/sec)*	*Call completion rate for X = 15 (calls/sec)*
IO_2	33.532	32.685	29.946
IO_3	41.300	39.134	41.478
IO_4	35.631	31.901	32.128
IO_5	42.453	41.232	42.153
IO_6	35.571	34.842	35.537
	Call completion rate for r = 1 (calls/sec)	***Call completion rate for r = 3 (calls/sec)***	***Call completion rate for r = 5 (calls/sec)***
NOWP	38.345	38.182	37.754

Table 7-12. Call completion rates, at region-1, obtained for different reduction/incremental step sizes.

When the number of reduction/incremental steps is increased from ten to fifteen, the call completion rate of IO_2 decreases, since more reduction steps are necessary reduce the IAM load to the required level. However, even though IO_4 has to reduce the IAM load by the same number of steps as IO_2, IO_4 is more responsive to the TFC arrival rate and therefore reduces its load much faster. In fact, all the IO control schemes that throttle their IAM loads severely in response to TFC messages perform better when the reduction/incremental step-size is increased. Their average IAM load is also larger, since the reintroduction of IAM traffic is less intense and STP-2 becomes congested less often and less severely. In the NOWP, STP-2 remains in a congested state for at least '3×Ty' seconds, even though the message arrival rate is less intense and fewer TFC messages are necessary to throttle the IAM load. This steadfast overcontrol of IAM traffic effectively reduces the IAM load allowed into the network and consequently also decreases the call completion rate.

In the IO and NOWP, the T29 and T30 congestion control timers determine how soon after throttling full transmission of a throttled traffic stream will resume; while in the NOCP the flow control timers, T15 and T16, determine for how long traffic from ISUP is blocked. The ITU-T Recommendations specify the following ranges for the ISUP congestion control timers and the NOCP flow control timers:

- T15 = 2 to 3 seconds,
- T16 = 1.4 to 2 seconds,
- T29 = 0.3 to 0.6 seconds and
- T30 = 5 to 10 seconds.

The following table lists the call completion rates obtained for different timer values, during the focused overload scenario:

	T29 = 0.3sec and T30 = 5.0sec.	*T29 = 0.6sec and T30 = 10.0sec.*
IO_2	32.685	33.724
IO_3	39.172	40.158
IO_4	31.901	34.813
IO_5	41.232	41.601
IO_6	34.842	36.773
NOWP	38.345	38.828
	T15 = 2.0sec and T16 = 1.4sec	***T15 = 3.0sec and T16 = 2.0sec***
NOCP	34.616	37.512

Table 7-13. The effect of timer settings on the call completion rates in region-1.

The above results show that for all the congestion control schemes considered, a higher call completion rate is accomplished if the larger timer values are used. With larger timer values the source nodes are slow to reintroduce the throttled IAM traffic and therefore fewer IAM messages are discarded at the congested STP.

7.6 Summary

The study of congestion and flow control techniques is an area of ongoing research. Signalling network outages [Bolotin et al., 1994] prompted a number of researchers to examine the applicability of link level flow controls as suitable flow control mechanisms in the event of STP processor congestion. However, most of these studies only consider the IO and NOCP in a small (usually single STP) network architecture. In addition, these studies only implement a linear step-wise throttling mechanism in ISUP, even though other research on overload controls has indicated that it is best to reduce the source traffic sharply at the detected onset of congestion and then to gradually restore the load as performance improves [Manfield et al., 1993].

This chapter evaluated the performance of the IO, the NOCP and the NOWP during STP processor congestion in a PSTN and mobile network. In addition, various different congestion control implementations, that incorporate non-linear throttling schemes, were also considered. These controls were examined in various overload scenarios for different parameter values, in order to determine which control schemes are robust and effective in different environments.

The results strongly indicate that, during congestion, the IAM traffic load to the affected destination must be sharply reduced at the initial onset of congestion, and then gradually increased if no further TFC messages are received. The control schemes that utilised this method produced high call completion rates and location management success rates in all of the overload scenarios examined. Networks using the superior congestion control schemes have the following characteristics, which demonstrates the effectiveness of these controls:

- The REL message load is slightly larger than the IAM message load, which indicates that very few REL or RLC messages are discarded at the congested STPs.
- The congestion is localised during focused overloads and the overload does not significantly impact the performance of the other STPs, nor does the congestion propagate to the other STPs.
- The average queue lengths in the affected STPs are small and the STPs are only congested for a brief interval before the load is reduced by the congestion controls.

- A small number of TFC messages are generated, which signifies a low flow control overhead.

The IO_2 control scheme, which is used by most researchers, performs poorly relative to the other control schemes examined in a PSTN environment. It is found to be susceptible to complete failure during some focused overload scenarios. In an environment where traffic streams from a large number of signalling points traverse the affected STP, the IO_2 flow controls are unable to throttle a sufficient number of source-destination pairs. The STP is therefore usually in the congested state.

The NOCP performs better than the IO_2 control scheme in some of the scenarios examined, as it is able to alleviate congestion at the affected STP by blocking traffic from most of the overload traffic streams soon after congestion is detected. However, when the source nodes detect congestion has abated the blocked traffic streams are restored to their full capacity, almost simultaneously. This consequently results in a large number of messages being periodically discarded at the affected STP. Within a mobile network the NOCP control performed poorly as excessive blocking of the mobility management traffic resulted in very few messages being admitted into the network. A much lower location update success rate and roaming number retrieval success rate was therefore obtained relative to the other congestion control schemes. PSTN networks using the IO_2 and NOCP control schemes are also susceptible to REL avalanches. These REL messages are primarily responsible for the degraded performance observed with the IO_2 control scheme in some scenarios.

The NOWP performs as well as the proposed IO control schemes in some cases. However if the flow controls are initiated with an initial congestion status of three, then the IAM message load is overcontrolled as TFC messages continue to be generated even after the congestion has abated. Nevertheless, performance in the NOWP (when $s = 3$) can be improved if TFC messages are generated less frequently, in which case more IAM messages are admitted into the network and consequently more calls are successful. An examination of various TFC flow control timer settings shows that a small Tx and Ty, that are greater than the average end-to-end message transfer delay, are recommended for optimum performance. If the initial congestion status is three then a larger Tx can improve the call completion rate, slightly.

Finally, an examination of ISUP and SCCP congestion controls in a GSM network indicates that the ISUP messages are subject to minimal discarding at the congested STPs. The traffic streams with the highest loads are between the MSCs/VLRs and the HLRs and these streams are more rigorously throttled by the congestion control mechanisms. In mobile networks the call completion rate is found to be sensitive to the success rate of the send routing information procedure.

8. Conclusion

The Signalling System No. 7 (SS7) protocol is used in modern telecommunications networks to carry signalling messages related to call control, database transactions and network management on a packet-switched network that is logically independent of the underlying circuit-switched trunk network. Signalling System No. 7 is based on a layered protocol architecture, which includes three Message Transfer Part (MTP) levels and an Application Part level, which includes various User Part protocols. Congestion situations may arise within each individual SS7 level, i.e., in level 2 (link receiver congestion or route set congestion), in level 3 (nodal congestion), and in level 4 (user part congestion) [Zepf et al., 1991].

Route set congestion occurs when predefined thresholds in the transmission and retransmission buffers of signalling links are exceeded. The ITU-T Recommendations define three types of flow control mechanisms for route set congestion, namely the International Option (IO), the National Option with congestion priorities (NOCP) and the National Option without congestion priorities (NOWP). The user part congestion controls respond to route set congestion indications by reducing traffic to the affected destination. MTP level 3 is responsible for network management and has to ensure the reliable transfer of signalling messages from source to destination. Level 3 processor congestion occurs when traffic arriving from the signalling links overwhelms the routing processor. However, flow controls for routing processor congestion are referred to as implementation dependent in ITU-T Recommendation Q.704. In most implementations the routing processor's flow controls simply discard the excess signalling traffic.

To date, numerous studies have examined the performance of signalling networks. Some of these studies use methods that have already been developed for the analysis of packet switched networks, namely the well-known Jackson and BCMP queuing network models. While these methods are usually used for the analysis of a single protocol level, a number of modelling methodologies are also available for the analysis of multilayered protocol architectures. However, some of the techniques used to analyse multilayered protocol architectures are known to be very complex and detailed and are occasionally difficult to solve analytically.

Studies of SS7 network performance during congestion scenarios show that the current congestion and flow control mechanisms perform poorly in some overload scenarios and are completely ineffective in a multiple user part environment ([Mayer, 1997] and [Zepf & Rufa, 1994]). Various researchers have also observed that signalling networks are prone to experience release (REL) message avalanches during congestion (Smith [1994] and Rumsewicz [1994]). Customer reattempts and message reattempts, due to application level recovery procedures, have also been found to seriously degrade network performance and they also impede recovery from congestion [Rumsewicz, 1993]. As a result, some researchers have suggested protocol changes to address the shortcomings of the current implementations. A combined control scheme, for example, that incorporates both the IO and NOCP congestion controls, was found to perform better than the weaker of the two congestion controls and at times it performed as well as the better of the two [Northcote & Rumsewicz, 1995].

There exists limited literature on the performance evaluation of large SS7 networks. Most studies have either analysed the network elements in isolation or concentrated on congestion of a single link or node, typically in a scenario where one node of a mated STP pair has failed [Skoog, 1988]. These studies do not consider the overall network structure and model all the sources as being directly connected to a single STP, which is the intermediate node during a focused overload [Rumsewicz, 1994]. In addition, previous studies [Smith, 1994] and protocol standards [ITU-T Recommendation Q.704] have also concentrated on transmission bottlenecks

that result in link level congestion. However, the trend towards the integration of broadband ATM signalling links and high-speed IP based signalling links into current and future networks aims to address the bandwidth bottleneck and shortcomings of narrowband signalling links. This may consequently lead to processor capacity becoming the primary bottleneck during periods of unexpected traffic peaks. Furthermore, previous studies have focused on SS7 congestion scenarios in a PSTN environment and no work exists on the impact of STP congestion on mobile networks, where signalling traffic is dominated by mobility management messages.

This study analysed the performance of large Signalling System No. 7 networks during signalling transfer point congestion. Analytical models were developed to analyse the impact of STP congestion on network performance in both the PSTN and PLMN environments when only simple messages discard schemes are employed. Simulations were also developed to validate the analytical models and to investigate the suitability and effectiveness of various congestion and flow control mechanisms.

The steady state equilibrium model, derived here, principally provides a means of quickly estimating the safe operating regions of a signalling network, while the transient model provides a more intuitive perspective of the traffic processes that eventually lead to network congestion. Comparisons between the results obtained by analysis and those obtained via simulations show that the analytical models provide a good representation of network performance. The analytical models, developed here, have the following improvements over previous models:

- any number of STPs and SPs may be included in the network structure,
- different message priority schemes may be analysed,
- different routing algorithms may be evaluated,
- a call matrix and mobility matrix allow for different traffic scenarios to be examined,
- and the analytical models are not confined to a single network configuration.

An analytical model for a priority based queuing discipline with three discard thresholds was also developed. Here, multiple discard thresholds are used to control the admission of the various message types into the queue. The discard process modelled here is analogous to message discarding in the National Option with congestion priorities, when the flow control mechanisms are ineffective. With the analytical models, one is able to calculate various performance measures, such as the call completion rate, location update success rate, message throughput, and the delay for various first offered loads. The network models can also be used to determine the maximum number of calls and location updates sustainable by a signalling network, to select appropriate buffer thresholds and priority schemes, and to analyse various network performance measures during different overload scenarios.

The results obtained show that it is possible for large signalling networks to become congested after a few hundred seconds of overload and thereafter remain in a congested state. Customer reattempts and application level recovery actions are mainly responsible for the substantial increase in signalling traffic, observed during congestion situations. In networks with unbalanced traffic loads, node failures or focused overloads, congestion commences in the STP that is carrying the largest traffic load. Once a PSTN's network element is congested, a large number of release attempts (due to call failures) are initiated. These REL messages aggravate congestion at the already congested STP, but they also increase the traffic load carried by the intermediate STPs. Congestion therefore propagates when the REL load to the congested region overwhelms the uncongested STPs. These REL messages are responsible for maintaining the high signalling load and sustained congestion situation when the overload traffic is removed. Furthermore, the results show that in large signalling networks, transit messages (i.e. messages that traverse two or more STPs) have a much lower throughput than signalling messages that only traverse a single STP, especially during severe congestion situations. The call completion rate and location update success rate between SPs in the non-adjacent regions therefore drops to zero when the network is congested. Assigning a higher priority to transit messages could therefore help to better utilise the available resources and improve message throughput.

In the analysis of message priority schemes in a PSTN environment, the 01212_{ISUP} priority scheme experiences a periodic surge in the REL message load every T_{IAM} seconds, which results in a step-wise decrease in the mean call completion rate and a corresponding step-wise increase in the mean STP queue length. This burst in the REL message load is much larger and more prominent when very small call holding times are used. However when very large call holding times are present, the network is able to sustain more calls before the call completion rate drops to zero.

An investigation on the influence of buffer threshold settings on network performance indicates that very large buffers do not necessarily improve message throughput or the call completion rate once the network is congested; but they do increase the queuing delay endured by signalling messages in the MTP level 3 processing buffer. The probability of discarding high priority messages, such as ANM, ULA and network management messages, in appropriately selected message priority schemes is found to be negligible. These messages therefore experience a high throughput as the processing buffer is rarely filled to its maximum capacity.

Mobile networks differ from the PSTN scenario, in that congestion is aggravated by the flood of UL and SRI messages triggered by application level procedures and customer reattempts. A decrease in the location update success rate is accompanied by a corresponding deterioration in the accuracy of the HLR's database and in the location cancellation success rate. While previous studies have generally ignored the impact of larger subscriber profiles on signalling network performance, results here show that when two or more ISD messages are required, to transfer a profile, the peak message handling capability of the network is reduced and the location update success rate drops close to zero more rapidly. These results highlight the importance of using the large message size capability of high speed ATM signalling links and SIGTRAN signalling links to transfer the entire subscriber profile in a single ISD message. Furthermore, unlike the PSTN environment where the call completion rate drops to zero when some selective discard schemes are used, in the mobile environment if suitable priorities are assigned to the MAP and ISUP messages it is possible to maintain the network's performance at an optimum level for both user parts with a minimal degradation in the effective throughput.

In addition to the GSM mobility management procedures the Simple Location Update (SLU) Protocol, the Super-Charged Location Update Protocol and the Lightweight Location Lookup Protocol (LiLLP) were also analysed analytically. The SLU protocol (which is commonly used by other researchers) is found to be inadequate at modelling the performance and behaviour of GSM. Not only does the SLU protocol support a higher location update success rate, since no ISD and ISDA messages are used, but unlike GSM the network performance degrades gradually for higher offered loads in a congested network. Furthermore, with SLU the network does not remain congested once the overload traffic is removed. The Super-Charger mechanism, which is specified by 3GPP for use in GSM and 3G networks, is found to significantly reduce signalling load and throughput can be increased by up to 300% when compared to the GSM scenario that requires the transfer of two ISD messages. The LiLLP protocol also offers some relief to the signalling network by reducing the number of messages generated by the MAP call delivery procedures, but the ISUP call completion rate is only marginally improved at very high loads.

Most studies that consider the efficacy of user part congestion controls, only implement a linear step-wise load throttling scheme, even though ITU-T Recommendations state that the amount of increase or decrease in the load at each step is implementation dependent. The performance of various proposed ISUP and SCCP congestion control schemes were compared to the control schemes used by other researchers. In addition, the performance of the NOWP was also investigated, as there are no previous studies on the performance of this control scheme. The control schemes were examined in various overload scenarios and with different parameter values, in order to determine which control schemes are robust and effective in different environments.

Results strongly indicate that, during congestion, the load generating traffic to the affected destination must be sharply reduced at the initial onset of congestion, and then gradually increased if no further TFC messages are received. The congestion control schemes that utilised this method produced the highest call completion rates and location update success rates in all of the overload scenarios examined. Additionally, in a mobile network environment ISUP messages are found to be subject to minimal discarding as the SCCP congestion controls are more active due to the higher mobility management load that is present. The call completion rate in mobile networks is essentially sensitive to the success rate of the send routing information procedure. Networks using the superior congestion control schemes have the following characteristics, which demonstrates their effectiveness:

- The REL message load is slightly larger than the IAM message load, which indicates that very few REL or RLC messages are discarded at the congested STPs.
- The congestion is localised and overloads do not significantly impact the performance of the other STPs.
- The average queue lengths in the affected STPs are small and the STPs are only congested for a brief interval before the load is reduced.
- A small number of TFC messages are generated, which signifies a low flow control overhead.

The step-wise load reduction congestion control scheme, which is used by most researchers, performs poorly relative to the other control schemes examined and is found to be susceptible to complete failure during focused overload scenarios. This scheme is also unable to throttle a sufficient number of source-destination pairs in an environment where traffic streams from a large number signalling points traverse the congested STP. The NOCP performs better than the step-wise load reduction control scheme in some of the scenarios examined, but especially in the focused overload scenarios. The NOCP is able to alleviate congestion at the affected STP by blocking traffic from most of the overload traffic streams soon after congestion is detected. However, the blocked traffic steams are restored to their full capacity, almost simultaneously, soon after the source nodes detect congestion has abated. This consequently results in a large number of messages being periodically discarded at the affected STP. PSTN networks using the step-wise load reduction control scheme and NOCP control scheme are also susceptible to REL avalanches. These REL messages are primarily responsible for the degraded performance.

The NOWP performs as well as the proposed IO control schemes. However, if the flow controls are initiated with an initial congestion status of three then the IAM message load is overcontrolled as TFC messages continue to be generated even after the congestion has abated. Nevertheless, performance in the NOWP can be improved if TFC messages are generated less frequently, in which case more IAM messages are admitted into the network and consequently more calls are successful. An examination of various TFC flow control timer settings shows that small timer values, that are greater than the average end-to-end message transfer delay, generally provide good performance.

The next step, in the advancement of signalling protocols is to meet the evolving requirements of telecommunications networks. Even though Internet protocols have been integrated into future telecommunications standards to some extend, SS7 still continues to serve a pivotal role in 3G networks and the evolution of the SS7 transport and application level protocols still continues within the ITU-T, ETSI, 3GPP and recently within the IETF. The following lists some future research in this area:

- ATM based signalling links have only been used in small isolated deployments over the past few years and are expected to be more widely used in 3G networks, while the specifications for IP based SIGTRAN signalling links have only very recently been approved as draft protocol standards by the IETF. Yet very little research and analysis exists on ATM and IP based signalling link performance or the appropriate selection of transport protocol parameters to meet SS7's reliability requirements.

- SIGTRAN allows for logical signalling points to be created (where a point code can be distributed across more than one host). No work exists as yet on the synchronisation and failure-over requirements of SS7 applications in a distributed environment or the impact of congestion in a single host on network performance.
- The operation of congestion control mechanisms and application level procedures in traditional circuit switched telecommunications networks and voice over IP networks are very dissimilar. For example, the congestion control actions defined for the Session Initiation Protocol are not as robust as those defined for ISUP and SS7. An investigation is necessary to examine how congestion in one network impacts the other network and to determine whether the congestion control mechanisms of both networks can be synchronised.
- Nagarajan's [1999] analysis of signalling traffic between a BSC and a MSC does not explicitly model individual messages nor does it consider SCCP connection orientated procedures and higher layer application procedures. A more through analysis of signalling between the MSC and multiple radio network elements in GSM and 3G networks is an area that needs to be investigated in greater depth to examine the impact of congestion on the strict timing requirements of radio network procedures.
- The ITU-T has recently specified the Bearer Independent Call Control protocol to address the shortcomings of ISUP and work is currently in progress to define a Transport Independent-SCCP. While the procedures supported by these protocols are compatible with existing ISUP and SCCP procedures they also support enhanced functionality. Future research needs to consider how the congestion control mechanisms of these user parts should respond to congestion detected by different signalling transport protocols.

"The challenge now remains for the effective design of a transport independent congestion control mechanism, that is not only robust over a wide range of system parameters and implementation scenarios but is also able to cope with the requirements of new and evolving services."

A. Appendix

A.1 Algorithm for the Determination of the Steady-State Equilibrium Solution

The following iterative algorithms were used to solve the system of equations given in Chapters 3 and 5, for first offered loads of α and β:

Calculation of the system parameters for the congested and uncongested regions:

STEP 1. Select starting values for each p_{ij}, the blocking probability at threshold '$i+1$' ($0 \le i \le 2$) of STP-j ($1 \le j \le N$, where N is the number of STPs in the network). Experimentation shows that by selecting an initial value of zero for all the probabilities of discard, the system of equations converges towards the uncongested state, where multiple solutions exist. An initial value of one for all the probabilities of discard results in convergence towards the congested state. In regions where only one solution exists, selecting any arbitrary initial value from zero to one will cause the system of equations to converge towards the correct solution.

STEP 2. Select ε, the precision to be used for determining the stopping criteria.

STEP 3. Calculate the transmission rates of the message streams from each SP region (i.e. $\lambda_{IAM\,xyz}$, $\lambda_{ACM\,xyz}$, etc.) with equations (3-14) to (3-17) and the call completion rate for each SP region with equation (3-18) in the PSTN analysis. The equations for messages arrival rates in the PLMN analysis are given in Chapter 5.

STEP 4. Calculate the message arrival rates at each STP with equation (3-21).

STEP 5. Let $p'_{ij} = p_{ij}$.

STEP 6. Calculate the equilibrium distribution and p_{ij}, with equations (3-2) and (3-3), by using the message arrival rates calculated in STEP 4.

STEP 7. If $|p_{ij} - p'_{ij}| < \varepsilon$ for all p_{ij}'s,

- then go to STEP 8;
- else go to STEP 3;

where $|x|$ is the absolute value of x.

STEP 8. Finally, calculate the throughput (with equation 3-19), link loads (with equations 3-26 to 3-28), queue lengths (with equation 3-4) and the end-to-end message transfer delay (with equation 3-30).

Calculation of system parameters for the meta-stable region:

STEP 1. First calculate the system parameters for the congested and uncongested regions, as described above. If the system parameters converge to two different solutions then it implies the existence of three solutions (unless the two solutions exist at the end-points where the meta-stable region originates from the congested region or terminates at the uncongested regions).

STEP 2. Let pu_{ij} equal the probability of discards determined for the congested region and let pl_{ij} equal the probability of discards determined for the uncongested region.

STEP 3. Also let $pu'_{ij} = pu_{ij}$ and $pl'_{ij} = pl_{ij}$.

STEP 4. Select ε, the precision to be used for determining the stopping criteria.

STEP 5. Let $p_{ij} = (pu'_{ij} + pl'_{ij})/2$.

STEP 6. Calculate the transmission rates of the message streams from each SP region (i.e. $\lambda_{IAM\,xyz}$, $\lambda_{ACM\,xyz}$, etc.) with equations (3-14) to (3-17) and the call completion rate for each SP region with equation (3-18) in the PSTN analysis. The equations for messages arrival rates in the PLMN analysis are given in Chapter 5.

STEP 7. Calculate the message arrival rates at each STP with equation (3-21).

STEP 8. Let $p'_{ij} = p_{ij}$.

STEP 9. Calculate the equilibrium distribution and p_{ij}, with equations (3-2) and (3-3), by using the message arrival rates calculated in STEP 7.

STEP 10. If $|p_{ij} - p'_{ij}| < \varepsilon$ for all p_{ij}'s,
- then go to STEP 11;
- else go to STEP 6.

STEP 11. If $|p_{ij} - pl_{ij}| < \varepsilon$ for all p_{ij}'s,
- then $pl'_{ij} = (pu'_{ij} + pl'_{ij})/2$ and go to STEP 5.

STEP 12. If $|p_{ij} - pu_{ij}| < \varepsilon$ for all p_{ij}'s,
- then $pu'_{ij} = (pu'_{ij} + pl'_{ij})/2$ and go to STEP 5.

STEP 13. Finally, calculate the throughput (with equation 3-19), link loads (with equations 3-26 to 3-28), queue lengths (with equation 3-4) and the end-to-end message transfer delay (with equation 3-30).

A.2 Routing Tables

For the purpose of this study, the shortest path routing algorithm was used. Figure A-1 illustrates the traffic allocation figures for uniformly distributed traffic loads, when no link or node failures are present. The traffic allocations can be obtained by examining the routing of messages between each source-destination pair. In the figure, λ_i can be interpreted as the total number new call arrivals at SPs in Region-*i*, or alternatively the total number of IAM messages generated by SPs in Region-*i*. Each region consists of n signalling points ($n = 6$ was used in the PSTN analysis). The access links, or SP-STP links, shown in the figure therefore consist of n signalling links (one from each signalling point).

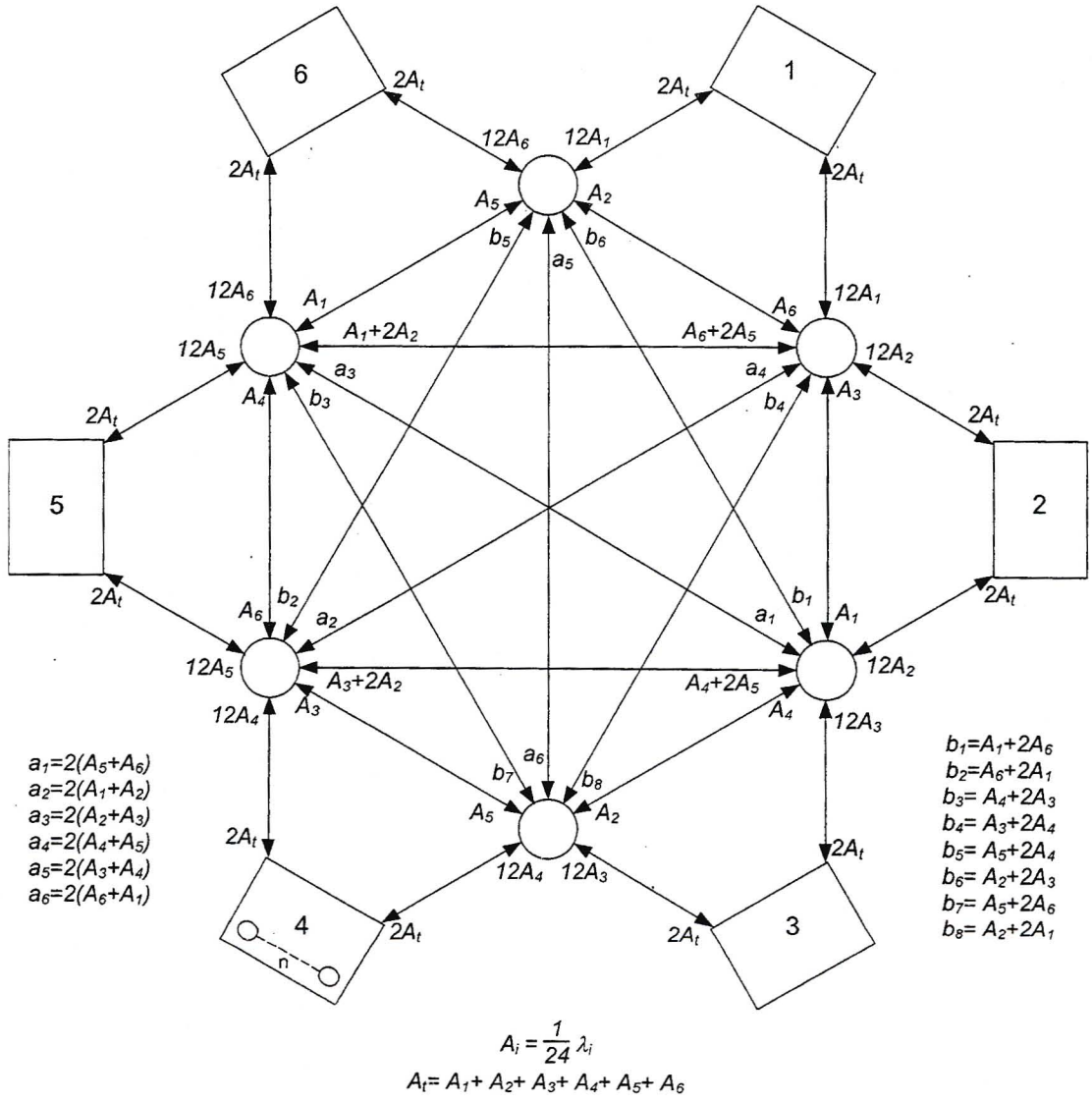


Figure A-1. Traffic distribution for the network examined.

Table A-1 lists the traffic routes followed by calls originating from Region-1 and destined to SPs in a non-adjacent region (4), adjacent region (6) or in the same region (1). However, routing between the other regions can be easily resolved from these examples. Where more then one route exists between two SP regions, the traffic load is balanced equally across each route. Messages travelling in the reverse direction follow the same route as that used by the messages that initiated them (e.g. a RLC message will traverse the same STPs as the REL message that triggered it).

Traffic Routing from SP Region-1 to SP Region-4 (routing between SPs in non-adjacent region)	
Forward	**Reverse**
R-1→STP-1→STP-5→R-4	R-4→STP-5→STP-1→R-1
R-1→STP-1→STP-4→R-4	R-4→STP-4→STP-1→R-1
R-1→STP-2→STP-5→R-4	R-4→STP-5→STP-2→R-1
R-1→STP-2→STP-4→R-4	R-4→STP-4→STP-2→R-1
Traffic Routing from SP Region-1 to SP Region-6 (routing between SPs in adjacent region)	
Forward	**Reverse**
R-1→STP-1→R-6	R-6→STP-1→R-1
Traffic Routing from SP Region-1 to SPs within Region-1 (routing between SPs in the same region)	
Forward	**Reverse**
R-1→STP-1→R-1	R-1→ STP-1→R-1
R-1→STP-2→R-1	R-1→ STP-2→R-1

Table A-1. Routing table for calls from a SP in Region-1 to destinations in the adjacent, non-adjacent or in the same region (under normal conditions).

Table A-2 lists the routes followed when STP-1 has failed. Traffic from Region-1 to Region-6 is now routed via STPs 2 and 6, while traffic to Regions 1 and 4 continues to use the available routes.

Traffic Routing from SP Region-1 to SP Region-4 (SPs in non-adjacent region)	
Forward	**Reverse**
R-1→STP-2→STP-5→R-4	R-4→STP-5→STP-2→R-1
R-1→STP-2→STP-4→R-4	R-4→STP-4→STP-2→R-1
Traffic Routing from SP Region-1 to SP Region-2 (SPs in adjacent region)	
Forward	**Reverse**
R-1→STP-2→STP-6→R-6	R-6→STP-6→STP-2→R-1
Traffic Routing from SP Region-1 to SPs within Region-1 (SPs in the same region)	
Forward	**Reverse**
R-1→STP-2→R-1	R-1→ STP-2→R-1

Table A-2. Routing table for calls from a SP in Region-1 to destinations in the adjacent, non-adjacent or in the same region (when STP-1 has failed).

A.3 Simulation Modelling Tools

It is often difficult to express complex real-world systems mathematically and experimental evaluation is not always possible. Simulation models provide an attractive alternative, as they are designed to duplicate the behaviour or real systems. They also provide a method of reproducing the salient features and random fluctuations that characterise the events in a real system, e.g. the arrival rate of customers at a queue. Unlike a mathematical analysis, in a simulation the arrival of packets at a node can be described by individual events. This is not possible in analytical models since the arrival rate is modelled as a probability distribution. Section A.3.1 provides an overview of OPNET, which was used to develop the SS7 simulator described in Section A.3.2.

A.3.1 Optimised Network Engineering Tools

OPtimised Network Engineering Tools (OPNET) is a simulation development environment, designed specifically for the modelling of communication networks. OPNET is based on a hierarchical modelling methodology, comparable to the structure of real world networks. Various graphical editors are available to describe different aspects of a communications system. A brief description of the different graphical editors is given below.

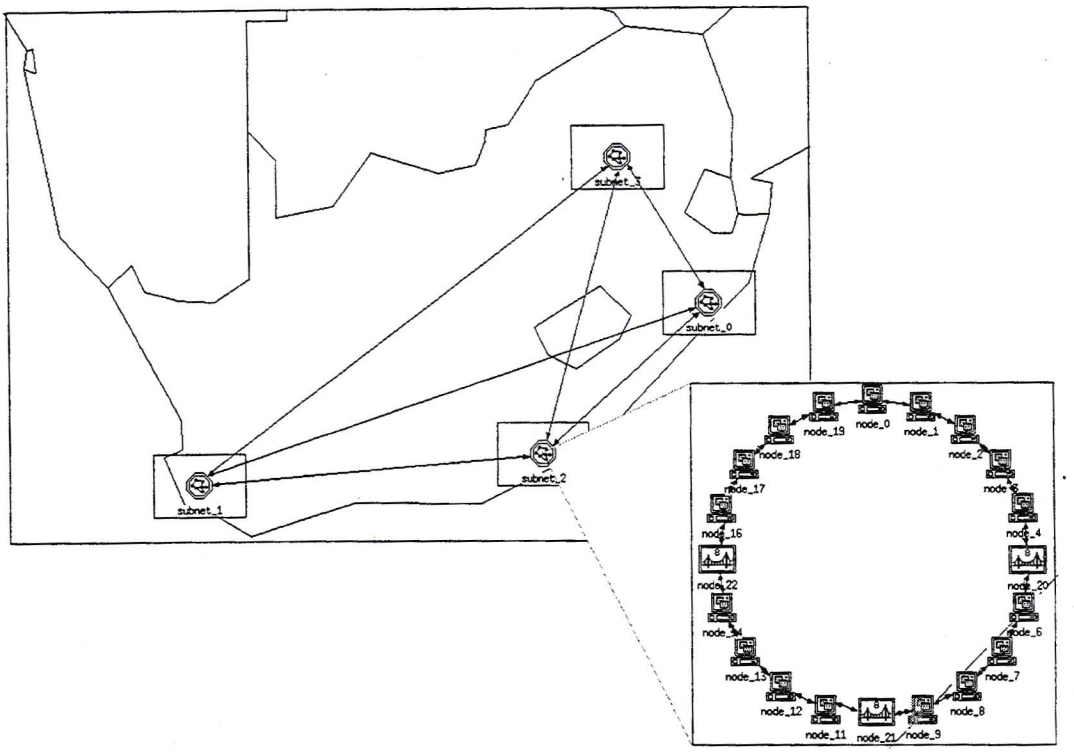


Figure A-2. A simulation model in the network domain.

a) Network Editor: The network editor is used to realise the highest, most easily perceptible, level of the network topology. Communication entities within the network domain represent nodes, links, routers, workstations, etc. Cartographic maps can also be added to the background to define the relationship between communication entities in the physical world. A network is constructed by simply connecting different communication entities to each other. The relative position of nodes in the network domain provides a perception of distance, and allows for the automatic calculation of propagation delays during a simulation run. The radio modeller also allows for three-dimensional trajectories to be defined for the mobile nodes. In addition, a complex network can be simplified by creating abstract subnetworks. Each subnetwork, for example, can represent a local area network (Figure A-2).

b) Node Editor: The node editor is used to construct the internal architecture of communications equipment. Node models are developed with smaller components called modules. Modules represent packet generators, processors, queues, transmitters and receivers. The behaviour of processors and queues can be fully programmed and defined within the Process Editor. Modules interact and communicate with each other via packet streams and statistic wires. Packet streams allow data packets to be transferred between the interconnected modules, while statistic wires are used to convey status information between the different modules. This type of modelling approach can be used to construct nodes based on layered communication protocol architectures, as shown in Figure A-3.

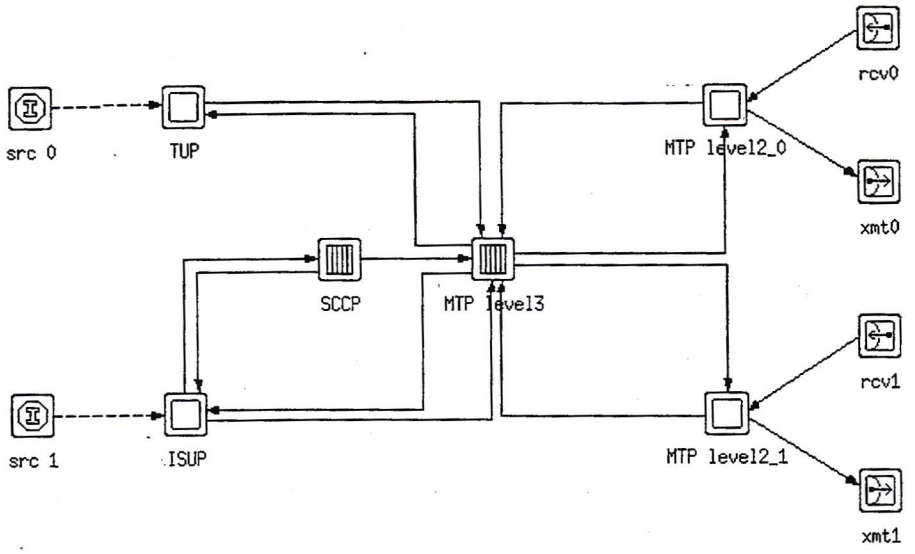


Figure A-3. An example of the layered structure of process modules in a signalling point.

c) Process Editor: Process models specify the internal operation of the processor and queue modules used in the node domain. A process is analogous to a set on instructions executed by a software program in response to an event or an interrupt. OPNET is a discrete event driven simulator. A simulation time and a process model are therefore associated with each event. Events are arranged in a time-ordered list and the simulation progresses by executing the process associated with the event at the front of the list. The functionality of each process is coded using Proto-C, a combination of state transition diagrams, the standard C language and a library of simulation kernel procedures (Figure A-4 shows an example of a state transition diagram). The simulation kernel library contains a set of C callable functions for procedures that are commonly utilised in simulation development; such as random number generators, event generators, routing algorithms, etc. Processes are also capable of generating child processes dynamically during a simulation; this is necessary for modelling multithreading and multitasking environments, e.g. a server that spawns a thread for each new connection established or a M/M/K queue.

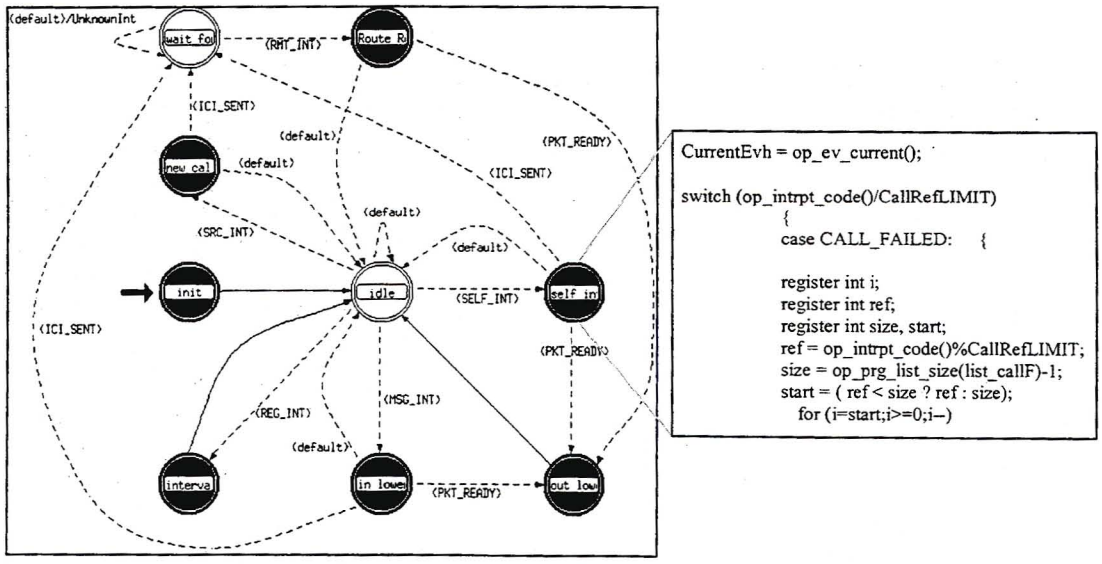


Figure A-4. Example of a state transition diagram and the code associated with one of the states.

d) Probe Editor: The probe editor is used to select the data collection points in a network. Common measurements include queue size, statistic transitions and link utilisation. User

defined data that needs to be measured is defined in the process editor and then selected for measurement with a probe. Animation probes are also available in the probe editor. With animations, the dynamic behaviour of the simulation can be viewed at the network, node and process levels. Animations are useful for debugging, e.g. for monitoring routing, packet sequences and process execution.

e) Parameter Editors: The various parameter editors are used for the specification of packet structures, antenna gain patterns, and probability density functions. A packet structure, including its field names, the associated data types and their default values, are defined in an environment that is graphically representative of the actual packet. Probability density functions are defined with a two-dimensional graphical editor or via an input file, while a three dimensional graphical editor is used to visualise and create the antenna gain patterns. Link models, their associated code and parameters are also defined a parameter editor.

OPNET has a built-in analysis tool. The tool is used to view data as time series plots, probability density functions, cumulative density functions, histograms and scatter plots. Data can also be imported or exported for external processing.

Further details on modelling with OPNET can be obtained from the OPNET Modeler reference manuals [MIL 3, 1997] or from MIL 3's Internet website at *http://www.mil3.com/*.

A.3.2 The SS7 Simulator

The SS7 simulator software code used in this study (except for the link model and OPNET simulation kernel functions) was developed entirely by the author. The link level model is based on the "point-to-point duplex link" model provided with OPNET. When this study commenced no SS7 models or GSM/UMTS models were available with OPNET. Recent versions of OPNET include a SS7 module and UMTS module. However an examination of the SS7 and UMTS modules indicates that they are not comprehensive models and are unsuitable for the type of work performed in this study. The SS7 model, for example, only supports link congestion and flow control mechanisms based on the NOCP, while the UMTS model is designed for analysis of the radio access network rather than the core network.

A.3.2.1 Operational Description

The SS7 simulation developed for this project is a detailed model of the MTP level 3 flow control procedures, the ISUP/SCCP congestion control functions and ISUP/MAP procedures. The MTP flow control procedures are used to monitor the occupancy of the STP processing buffer, to discard messages during congestion and to send TFC messages to the source nodes. When MTP level 3 in a signalling point receives a TFC message it notifies the user parts of the congestion situation with an interface control information (ICI) message. The user parts respond by invoking the appropriate congestion control mechanism. This simulator is able to model the following congestion and flow control mechanisms for STP congestion scenarios:

- Simple priority based message discarding.
- The International option (IO), including the various congestion control schemes analysed in Chapter 7.
- The National option with congestion priorities (NOCP).
- The National option without congestion priorities (NOWP).

User-defined parameters are set via model attributes and simulation attributes. The model attributes allow different values to be set for each instance of a process model (e.g. the message service rate of each STP), while the simulation attributes are used to define global values (e.g. message priorities).

Seventeen distinct packet formats are supported for the various types of messages used in the simulation (e.g. IAM, ACM, ANM, etc.). Each ISUP or MAP message has a *call reference field* or *transaction reference field* that is used to reference the data structures containing the event handles of timers currently pending. The *reference field* can also be extracted from the *interrupt code* of the ISUP and MAP timers. In addition, this field is also necessary to identify the data structure containing the destination node address, routing information and the event handles that are associated with a particular call or transaction. RCT message packets only have a *priority field* that is inspected at a congested node, to determine whether or not the packet should be discarded. TFC messages have a *congested destination* field and a *congestion level* field. Note; a *source address field* is not necessary since a default field specifies the address of the node that created the packet. Routing is accomplished by using the OPNET routing functions to find the shortest routes between the source and destination nodes. If multiple routes exist between the source and destination nodes, a route is randomly selected from the route set.

A.3.2.2 Model Scope and Limitations

The simulation model is based on a layered architecture. This allows for the process models to be replaced by new models, without the need to completely rewrite the entire simulation code (as long as the necessary interfaces are maintained).

The simulation also allows for the modelling of node and link failure conditions. However, the TFP, TFR and TFA procedures are not modelled, therefore failures/recoveries during the course of the simulation are not supported. Instead, a *status* model attribute is provided. This parameter is used to extract the network topology, necessary for creating the routing tables, during the initialisation phase of the simulation. The disabled nodes and links are therefore not included in the network topology. This method of enabling/disabling nodes and links allows one to use the same network model to analyse different failure scenarios, without having to create a separate model in each instance.

The effect of link level procedures (including the impact of interframe fill-in sequences and link level flow controls) on the overall performance of the system is assumed to be negligible. The "point-to-point duplex link" model provided with OPNET is therefore used. This model does not implement all the MTP level 2 functions.

The delay incurred by a message at a node, including the message processing time, depends upon the hardware architecture and the routine software procedures executed by the core processor. Selecting an appropriate stochastic distribution for the processing delays incurred in a node is therefore difficult, without actual measurements of a real implementation under various loading conditions. The service process in the message processing queues is therefore assumed to be an exponential distribution. The processing time incurred to create a TFC message and to discard a message in a congested STP is assumed to be negligible, while message routing is assumed to consume the bulk of the processing overhead in the core processor. New TFC messages are inserted at the end of the processing buffer, if buffer capacity is available. The TFC messages, like the other messages, therefore also consume processing time before being routed to the correct outgoing link. If the queue has reached its maximum size (determined by the highest discard threshold) all new messages that are received are discarded, and no further TFC messages are created until buffer storage capacity is again available.

A.3.2.3 Signalling Point Architecture

Figure A-5 illustrates the node level model structure of a signalling point. Six link transmitters and receivers are used in the following node model, however these are completely implementation dependent and the simulation supports any number of transmitters and receivers. Packets sent from the Message Routing process model to a link transmitter are queued in a buffer and serviced at a user-defined rate.

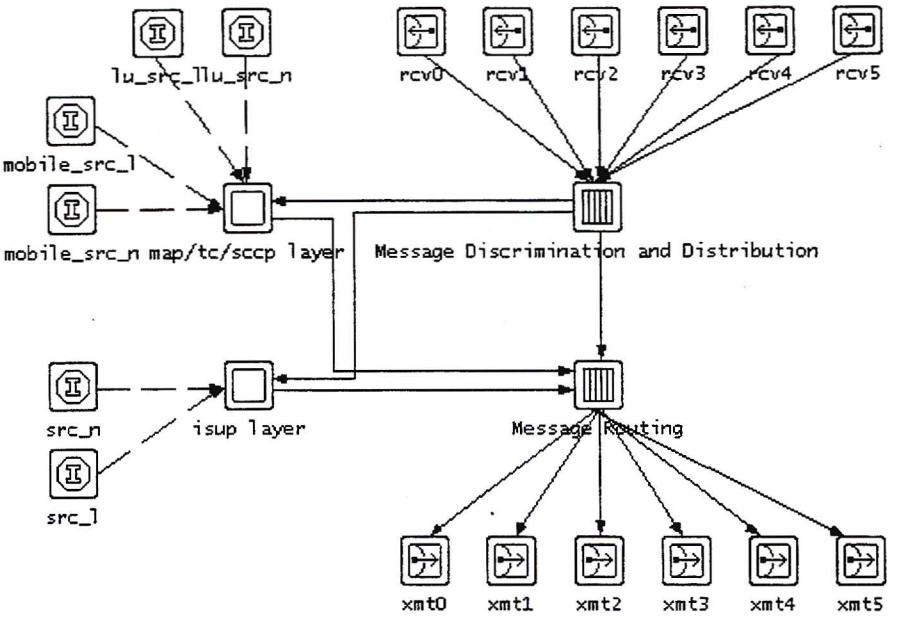


Figure A-5. Node model for the Signalling Points.

The message discrimination and distribution functions are implemented in a single process model that examines the messages that arrive at the signalling point. Figure A-6 shows the state transition diagram for these functions.

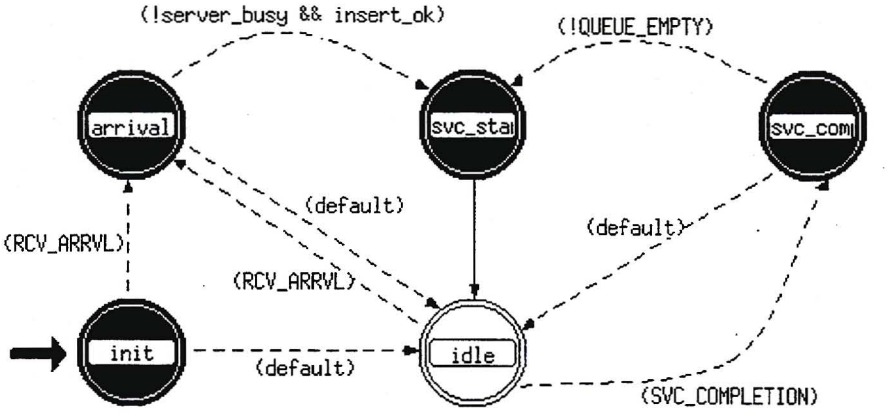


Figure A-6. Process model for the MTP level 3 message discrimination and distribution functions in a Signalling Point.

The primary function of the message routing process model (Figure A-7) is to send messages via the correct outgoing links to their respective destinations. This process model is also responsible for flow control in the NOCP. Two timers, T15 and T16, are started when a TFC message is received and an ICI package with a congestion indication is sent to the user parts. ISUP, SCCP and RCT messages are queued in a FIFO buffer with an exponential service time to account for the processing delay required to analyse and forward the packet to the appropriate outgoing signalling link.

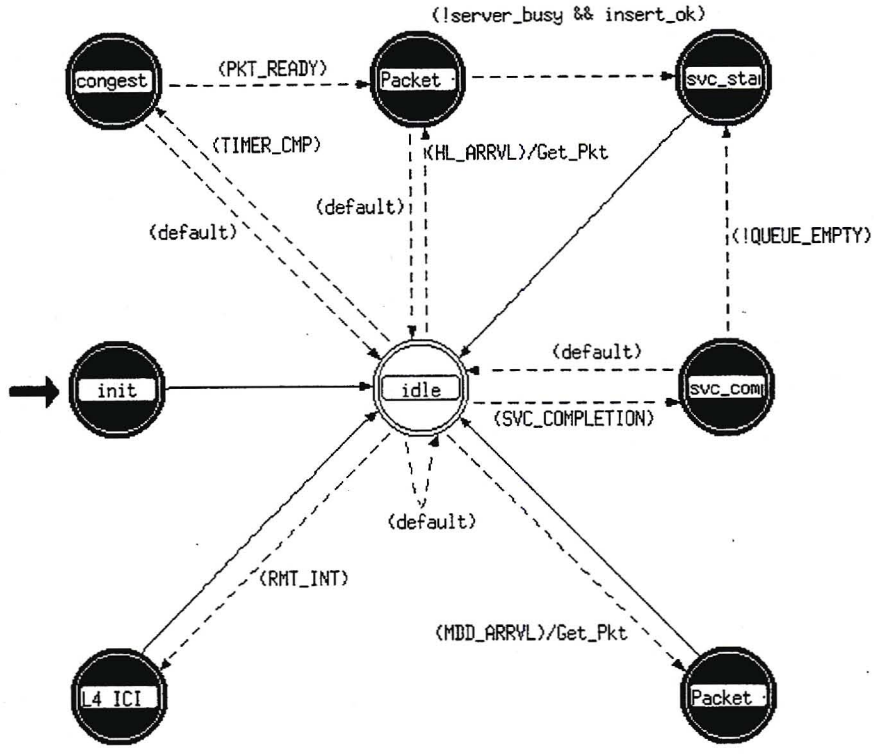
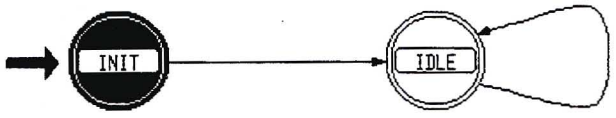


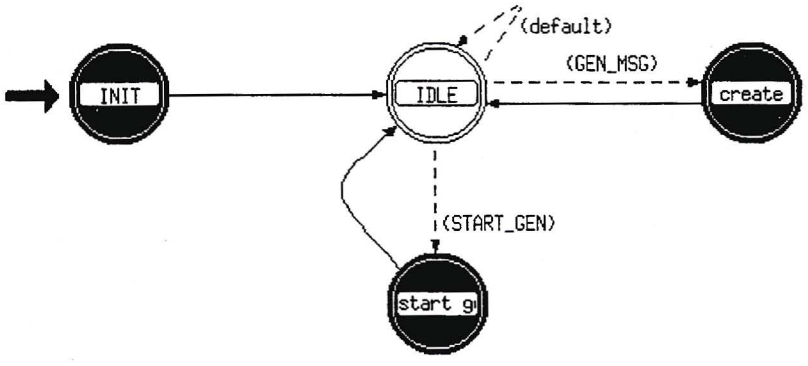
Figure A-7. Process model for the MTP level 3 message routing function in a Signalling Point.

Two types of Poisson traffic generators (Figure A-8) are used to generate new call arrivals and new location update attempts. The traffic generators are connected to the user parts through statistic wires. A new call arrival or location update toggles the level on the associated statistic wire, which consequently interrupts the process model. The process, then responds by initiating the necessary call set-up or location update processes. The background traffic source model toggles the level of its statistic wire in response to an interrupt and then schedules a new interrupt for the next call arrival. The process then returns to the idle state. The overload traffic source model requires additional model attributes to specify the start and end times of the overload traffic. This model is also able to reduce the overload traffic in a step-wise manner, in order to obtain results for multiple overload levels from a single simulation run.

In Figure A-5 the pair of traffic generators connected to the ISUP process model are used to generate the PSTN traffic load. The MAP process model has 4 traffic generators connected to it, two are used to generate mobile originating call traffic and two are used to generate the location update attempts.



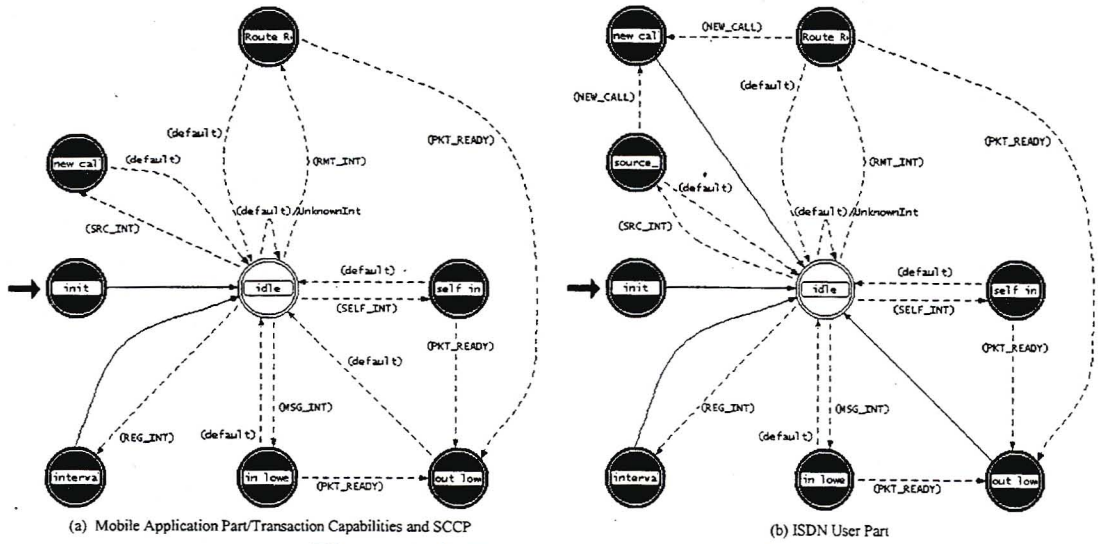
(a) Background Traffic Source



(b) Overload Traffic Source

Figure A-8. Process models for the new call traffic generators.

The ISUP and MAP process models (Figure A-9) generate all the messages related to calls and location updates. A single process model supports all the functions applicable to the MAP layer, transaction capabilities layer and SCCP layer. These models explicitly model the location management, call delivery and call set-up procedures. The ISUP and SCCP procedures respond to congestion indications, by throttling traffic to the affected destination and by activating the necessary congestion control timers. Each process model also maintains a list of all the calls and transactions in progress; which includes their call reference number, their destination address, the route followed and the event handles of their associated timers. The *interval interrupt* state is executed at regular intervals (e.g. 1-second periods), to record the system's status information (such as the call completion rate) in the output file. Both process models also communicate with each other through the ICI package to trigger ISUP call set-up from MAP or to indicate ISUP call failure to MAP.



(a) Mobile Application Part/Transaction Capabilities and SCCP

(b) ISDN User Part

Figure A-9. User part process models.

A.3.2.4 Signalling Transfer Point Architecture

Figure A-10 illustrates the node level model of a signalling transfer point. A single queuing process model is used to implement the MTP level 3 functions in the STP. These functions include message discrimination, message routing and signalling network management. The message distribution functions are not modelled since no user parts are present. As in the signalling point model, the number of link transmitters and link receivers is also user definable.

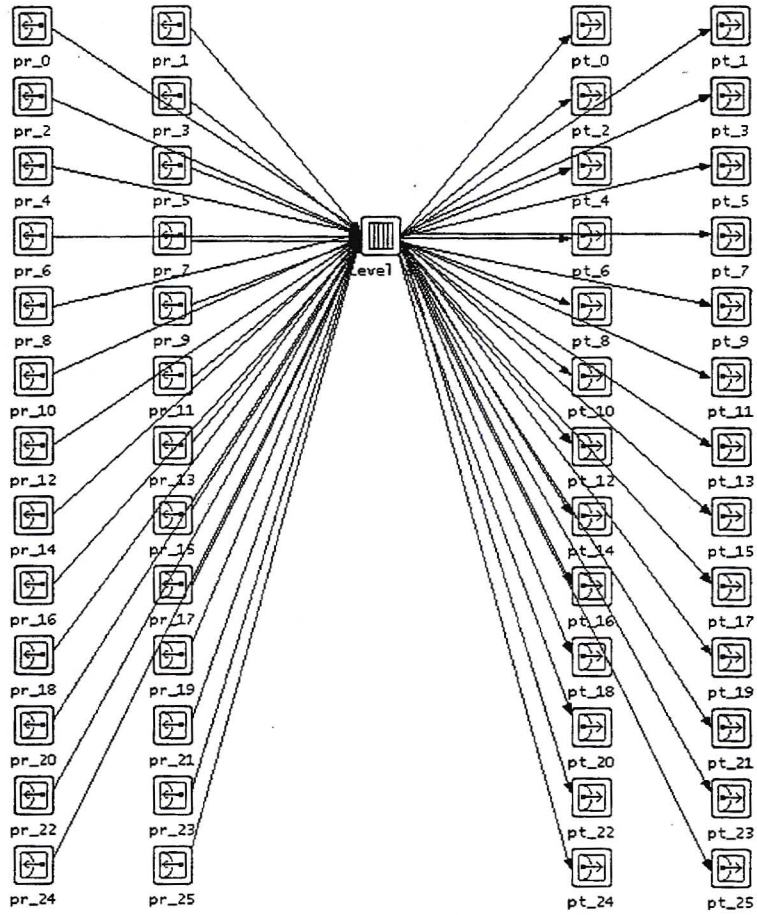


Figure A-10. Node model for the Signalling Transfer Points.

The MTP level 3 process model (Figure A-11) is modelled as a FCFS buffer with an exponential service time. Separate arrival states are used to model the different flow control procedures. These include; message discarding with no feedback control, IO flow control, NOCP flow control and NOWP flow control. On receipt of a signalling message, the process enters the appropriate flow control state to determine if the message can be admitted into the queue and if a TFC message should be generated. The system then starts the service process or returns to the idle state if a message is already being serviced. When the process model receives a self-interrupt, indicating the end of the service process, the routing label of the first message in the queue is examined and the message is forwarded to the appropriate outgoing signalling link transmitter.

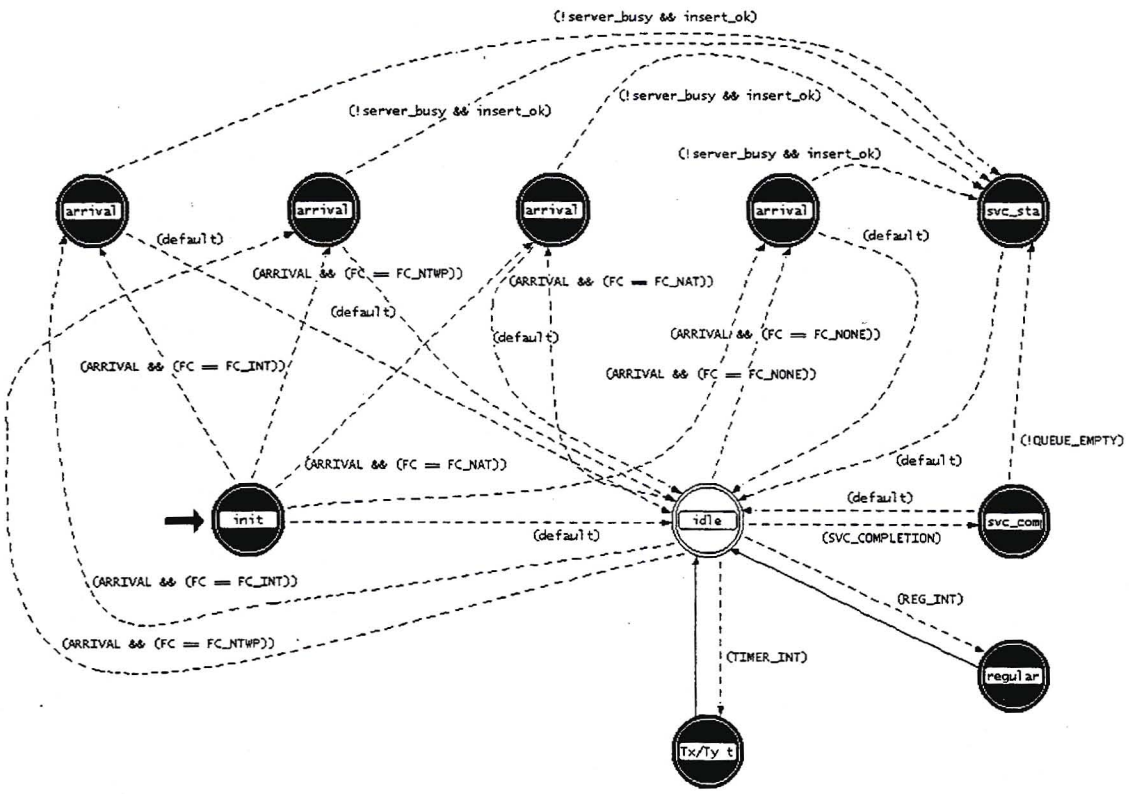


Figure A-11. Process model for the MTP level 3 functions in a Signalling Transfer Point.

References

3GPP, (2000), "3GPP Working Procedures," 17 July 2000, http://www.3gpp.org.

3GPP, (2002), "Technical Specification Group Core Network, Bearer-independent circuit-switched core network, (Release 5)," TS 22.205, v5.3.0.

3GPP, (2002), "Technical Specification Group Services and System Aspects, General Packet Radio Service (GPRS), Service description, Stage 2, (Release 5)," TS 23.060, v5.2.0.

3GPP, (2002), "Technical Specification Group Core Network, Super-Charger technical realization, Stage 2, (Release 5)," TS 23.116, v5.0.0.

3GPP, (2001), "Technical Specification Group Core Network, Technical Report on the Gateway Location Register, (Release 4)," TS 23.909, v4.0.0.

3GPP, (2001), "Technical Specification Group Core Network, Technical Report on Super-Charger, (Release 4)," TS 23.912, v4.1.0.

3GPP, (2002), "Technical Specification Group Core Network, Mobile Radio Interface Layer 3 Specification, Core Network Protocols, Stage 3, (Release 5)," TS 24.008, v5.4.0.

3GPP, (2002), "Technical Specification Group Core Network, Mobile Application Part (MAP) specification, (Release 5)," TS 29.002, v5.3.0.

American National Standards Institute, (2001), "Signaling System No. 7 (SS7) - Message Transfer Part (MTP)," T1.111.

Akyildiz, I F, McNair, J, Ho, J S M, Uzunalioglu, H & Wang, W, (1999), "Mobility Management in Next-Generation Wireless Systems," Proceedings of the IEEE, vol. 87, no. 8, August 1999, p. 1347-1384.

Bafutto, M, Kühn, P J & Willmann, G, (1994), "Capacity and Performance Analysis of Signaling Networks in Multivendor Environments," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 490-499.

Baskett, F, Chandy, K M, Muntz, R R & Palacios, F G, (1975),"Open, Closed, and Mixed Networks with Different Classes of Customers," Journal of the Association for Computing Machinery, vol. 22, no. 2, April 1975, p. 248-260.

Bolotin, V A, (1994), "Modeling Call Holding Time Distributions for CCS Network Design and Performance Analysis," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 433-438.

Bolotin, V A, Kühn, P J, Pack C D & Skoog, R A, (1994), "Guest Editorial Common Channel Signaling Networks: Performance, Engineering, Protocols, and Capacity Management," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 377-378.

Bonatti, M & Decina, M, (1988), "Traffic Engineering for ISDN Design and Planning," Elsevier Science Publishers B.V., North-Holland, 1988.

Burns, J E & Ghosal, D, (1997), "Automatic Detection and Control of Media Stimulated Focused Overloads," ITC-15, V. Ramaswami and P. E. Wirth (Editors), Elsevier Science Publishers B. V., p. 889-901.

Chana, A & Takawira, F, (1996), "Modelling Congestion Control in SS7 networks," Teletraffic Systems Engineering Seminar, Electronic Engineering, University of Natal, South Africa, September 1996.

Chana, A & Takawira, F, (1997), "Performance of the Transfer-Controlled Procedure in Controlling SS7 Congestion," Teletraffic `97, Rhodes University, South Africa, September 1997.

Chana, A & Takawira, F, (2000), "The Impact of Location Management Protocols on the Signalling Traffic Load in Mobile Networks," SATCAM 2000, September 2000.

Chana, A & Takawira, F, (2002), "An Analysis of Mobile Network Performance during Signalling Network Congestion," submitted to IEEE Transactions on Vehicular Technology.

Chana, A & Takawira, F, (2002), "SS7 Congestion Control Performance in Mobile Networks," submitted to IEEE Transactions on Vehicular Technology.

Chana, A & Takawira, F, (2002), "The impact of STP Congestion on the performance of Mobility Management Procedures," submitted to Wireless Networks.

Chana, A & Takawira, F, (2003), "An Examination of Signaling Traffic in a Super-Charged Mobile Network," accepted for presentation at WCNC 2003.

Chandramouli, Y & Krishnan, K R, (1999), "Signaling Network Impact of Local Number Portability," ITC-16, P. Key and D. Smith (Editors), Elsevier Science Publishers B. V., p. 1-10.

Chlebus, E, (1997), "Empirical validation of call holding time distribution in cellular communications systems," ITC-15, V. Ramaswami and P. E. Wirth (Editors), Elsevier Science Publishers B. V., p. 1179-1188.

Cohen, J W, (1957), "Basic Problems of Telephone Traffic Theory and the Influence of Repeated Calls," Philips Telecommunications Review, vol. 18, no. 2, p. 49-100.

Conway, A E, (1989), "Performance Modeling of Multi-Layered OSI Communication Architectures," Proceedings of the International Conference in Communications, June 1989, p. 651-657.

Conway, A E, (1990), "Queueing Network Modeling of Signaling System No. 7," Proceedings of Globecom `90, December 1990, p. 552-558.

Conway, A E, (1991), "A Perspective on the Analytical Performance Evaluation of Multilayered Communication Protocol Architectures," IEEE Journal on Selected Areas in Communications, vol. 9, no. 1, January 1991, p. 4-14.

Cui, Y, Lam, D, Widom, J & Cox, D C, (1998), "Efficient PCS Call Setup Protocols," Proceedings of IEEE INFOCOM `98, March/April 1998, p. 728-736.

Davies, D W, (1972), "The control of congestion in packet-switching networks," IEEE Transactions on Communications, vol. COM-20, June 1972.

ETSI, (1996) "Digital cellular telecommunications system (Phase 2); Network Architecture," GSM 03.02, Third Edition, November 1996.

ETSI, (1998) "Digital cellular telecommunications system (Phase 2); Mobile radio interface layer 3 specification," GSM 04.08, Eleventh Edition, October 1998.

ETSI, (1995) "European digital cellular telecommunications system (Phase 2); Signalling transport mechanism specification for the Base Station System - Mobile-services Switching Centre (BSS - MSC) interface," GSM 08.06, Second Edition, July 1995.

ETSI, (1998) "Digital cellular telecommunications system (Phase 2); Mobile-services Switching Centre - Base Station System (MSC - BSS) interface; Layer 3 specification," GSM 08.08, Sixth Edition, October 1998.

ETSI, (2000) "Digital cellular telecommunications system (Phase 2+); Mobile Application Part (MAP) specification," GSM 09.02, Tenth Edition, Release 1996, April 2000.

Eynard, C, Lenti, M, Lombardo, A, Marengo, O & Palazzo, S, (1995), "A Methodology for the Performance Evaluation of Data Query Strategies in Universal Mobile Telecommunication Systems (UMTS)," IEEE Journal on Selected Areas in Communications, vol. 13, no. 5, June 1995, p. 893-907.

Fasbender, A, Hoff, S & Pietschmann, M, (1995), "Mobility Management in Third Generation Mobile Networks," Proceedings of the IFIP TC6 International Workshop on Personal Wireless Communications (Wireless Local Access), April 1995, p. 33-46.

Gerla, M & Kleinrock, L, (1980), "Flow Control: A Comparative Survey," IEEE Transactions on Communications, vol. COM-28, no. 4, April 1980, p. 553-574.

Gianini, J & Pettitt, B, (1997), "Link Delays in Signalling System No. 7 Networks," ITC-15, V. Ramaswami and P. E. Wirth (Editors), Elsevier Science Publishers B. V., p. 1199-1208.

Goldberg, R R & Shrader, D C, (1990), "Common Channel Signaling Interface for Local Exchange Carrier to Interexchange Carrier Interconnection," IEEE Communications Magazine, July 1990, p. 65-71.

Grangé, J L & Gien, M, (1979), "Proceedings of the Symposium on Flow Control in Computer Networks," North-Holland Publishing Co., Amsterdam, February 1979.

Gross, D & Harris, C M, (1985), "Fundamentals of Queueing Theory," John Wiley & Sons, 2nd ed.

Hung, H N, Lin Y B, Peng, N F & Yang, S R, (2001), "Resolving Mobile Database Overflow with Most Idle Replacement," IEEE Journal on Selected Areas in Communications, vol. 19, no. 10, October 2001, p. 1953-1961.

Irland, M I, (1978), "Buffer Management in a Packet Switch," IEEE Transactions on Communications, vol. COM-26, no. 3, March 1978, p. 328-337.

IETF, (1999), "Framework Architecture for Signaling Transport," RFC 2719, October 1999.

IETF, (2000), "Stream Control Transmission Protocol," RFC 2960, October 2000.

IETF, (2002), "Signaling System 7 (SS7) Message Transfer Part 2 (MTP2) - User Adaptation Layer," RFC 3331, September 2002.

IETF, (2002), "Signaling System 7 (SS7) Message Transfer Part 3 (MTP3) - User Adaptation Layer (M3UA)," RFC 3332, September 2002.

ITU-T Recommendations, Q701-Q706, (1993-1996), "Specifications of Signalling System No. 7, Message transfer part (MTP)," International Telecommunication Union, Geneva.

ITU-T Recommendations, Q711-Q716, (1993-1996), "Specifications of Signalling System No. 7, Signalling Connection Control Part (SCCP)," International Telecommunication Union, Geneva.

ITU-T Recommendations, Q761-Q769, (1993-2000), "Specifications of Signalling System No. 7, ISDN user part," International Telecommunication Union, Geneva.

ITU-T Recommendations, Q771-Q775, (1997), "Specifications of Signalling System No. 7, Transaction capabilities application part," International Telecommunication Union, Geneva.

ITU-T Recommendations, Q2210, (1996), "Broadband ISDN - Signalling network protocols, Message transfer part level 3 functions and messages using the services of ITU-T Recommendation Q.2140," International Telecommunication Union, Geneva.

Jain, R, (1990), "Congestion Control in Computer Networks: Issues and Trends," IEEE Network Magazine, May 1990, p. 24-30.

Jain, R, Lin, Y B, Lo, C & Mohan, S, (1994), "A Caching Strategy to Reduce Network Impacts of PCS," IEEE Journal on Selected Areas in Communications, vol. 12, no. 8, October 1994, p. 1434-1444.

Jain, R & Lin, Y B, (1995), "An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS," Wireless Networks, vol. 1, no. 2, July 1995, p. 197-210.

Kant, K, (1997), "Flow-control Issues in ATM Signalling Link Deployment," ITC-15, V. Ramaswami and P. E. Wirth (Editors), Elsevier Science Publishers B. V., p. 1219-1228.

Kant, K & Ong, L (1997), "Signaling in Emerging Telecommunications and Data Networks," Proceedings of the IEEE, vol. 85, no. 10, October 1997, p. 1612-1621.

Kasera, S, Pinheiro, J, Loader, C, Karaul M, Hari, A & LaPorta, T, (2001), "Fast and Robust Signaling Overload Control," Proceedings of the 9th International Conference on Network Protocols, 2001.

Karmarkar, V V, (1994), "Assuring SS7 Dependability: A Robustness Characterization of Signaling Network Elements," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 475-489.

Kleinrock, L & Lam, S S, (1975), "Packet switching in a multiaccess broadcasting channel: Performance evaluation," IEEE Transactions on Communications, vol. COM-23, April 1975, p. 410-423.

Kosal, H & Skoog, R A, (1994) "A Control Mechanism to Prevent Correlated Message Arrivals from Degrading Signaling No. 7 Network Performance," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 439-445.

Krishna, P, Vaidya, N H & Pradhan, D K, (1996), "Efficient Location Management in Mobile Wireless Networks," Technical Report # 96030, Department of Computer Science, Texas A&M University, July 1996.

Kritzinger, P S, (1986), "A Performance Model of the OSI Communication Architecture," IEEE Transactions on Communications, vol. COM-34, no. 6, June 1986, p. 554-563.

La Porta, T F, Veeraraghavan, M & Buskens, R W, (1996), "Comparison of Signaling Loads for PCS Systems," IEEE/ACM Transactions on Networking, vol. 4, no. 6, December 1996, p. 840-855.

Lam, D, Cox, D C & Widom, J, (1997), "Teletraffic Modeling for Personal Communications Services," IEEE Communications Magazine, February 1997, p. 79-87.

Lam, S S & Luke Lien, Y C, (1981), "Congestion Control of Packet Communication Networks by Input Buffer Limits - A Simulation Study," IEEE Transactions on Computers, vol. C-30, no. 10, October 1981, p. 733-742.

Lazar, A A, Tseng, K H & Lim, K S, (1992), "Delay Analysis of the Singapore National CCSS#7 Network Under Fault and Unbalanced Loading Conditions," Proceedings of ICCS/ISITA '92, Singapore, p.994-998.

Lazar, A A, Tseng, K H, Lim, K S & Choe, W, (1994), "A Scalable and Reusable Emulator for Evaluating the Performance of SS7 Networks," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 395-404.

Lin, Y B, (1997), "Reducing Location Update Cost in a PCS Network," IEEE/ACM Transactions on Networking, vol. 5, no. 1, February 1997, p. 25-33.

Lin, Y B, (1998), "Deregistration Strategies for PCS Networks," IEEE Transactions on Vehicular Technology, vol. 47, no. 1, February 1998, p. 49-57.

Lin, Y B, (2000), "Overflow Control for Cellular Mobility Database," IEEE Transactions on Vehicular Technology, vol. 49, no. 2, March 2000, p. 520-530.

Lin, Y B & DeVries, S K, (1995), "PCS Network Signaling using SS7," IEEE Personal Communications Magazine, June 1995, p. 44-54.

Lin, Y B & Tsai, W N, (1998), "Location Tracking with Distributed HLR's and Pointer Forwarding," IEEE Transactions on Vehicular Technology, vol. 47, no. 1, February 1998, p. 58-64.

Lingnau, A & Drobnik, O, (1998), "Mobile Agents in a Mobile Communications System Database," 3rd IFIP TC6 Workshop on Personal Wireless Communications (Wireless Local Access), Tokyo, Japan, April 1998.

Majithia, J C, Irland, M, Grange, J L, Cohen, N & O'Donnell, C, (1979), "Experiments in Congestion Control Techniques," from Grangé & Gien [1979], p. 211-234.

Manfield, D R, Millsteed, G & Zukerman, M, (1993), "Congestion Controls in SS7 Signaling Networks," IEEE Communications Magazine, June 1993, p. 50-57.

Manfield, D R, Millsteed, G K & Zukerman, M, (1994) "Performance Analysis of SS7 Congestion Controls Under Sustained Overload," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 405-414.

Mayer, W J, (1997), "Congestion Control Interactions with Multiple User Parts in SS7," ITC-15, V. Ramaswami and P. E. Wirth (Editors), Elsevier Science Publishers B. V., p. 1189-1198.

Mayer, S, (2000), "Impact of GPRS on the Signalling of a GSM-based Network," Proceedings of the 1st Polish-German Teletraffic Symposium (PGTS 2000), Dresden.

Meier-Hellstern, K S & Alonso, E, (1991), "Signaling System No. 7 Messaging in GSM," Rutgers University, Technical Report WINLAB-TR-25, December 1991.

MIL 3, (1997), OPNET Modeler manuals, MIL 3 Inc., Washington DC, version 3.0A.

Modarressi, A R & Skoog, R A, (1990), "Signaling system No. 7: A Tutorial," IEEE Communications Magazine, July 1990, p. 19-35.

Nagarajan, R, (1999), "Threshold-based congestion control for the SS7 signaling network in the GSM digital cellular network," IEEE Transactions on Vehicular Technology, vol. 48, no. 2, March 1999, p. 385-396.

Northcote, B & Rumsewicz, M, (1995), "Congestion Control Dynamics in CCS Networks During STP Processor Overload," Proceedings of ICC `95, Seattle, July 1995.

Pollini, G P, Meier-Hellstern, K S & Goodman, D J, (1995), "Signaling Traffic Volume Generated by Mobile and Personal Communications," IEEE Communications Magazine, June 1995, p. 60-65.

Pollini, G P & Goodman, D J, (1996), "Signaling System Performance Evaluation for Personal Communications," IEEE Transactions on Vehicular Technology, vol. 45, no. 1, February 1996, p. 131-138.

Price, W L, (1977), "Data network simulation experiments at the National Physical Laboratory," Computer Networks, vol. 1, p. 199-208.

Rumsewicz, M P, (1993), "Analysis of the Effects of SS7 Message Discard Schemes on Call Completion Rates During Overload," IEEE/ACM Transactions on Networking, vol. 1, no. 4, August 1993, p. 491-502.

Rumsewicz, M, (1994), "On the Efficacy of using the Transfer-Controlled Procedure during periods of STP processor overload in SS7 networks," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 415-423.

Rumsewicz, M P & Smith, D E, (1995), "A Comparison of SS7 Congestion Control Options during Mass Call-in Situations," IEEE/ACM Transactions on Networking, vol. 3, no. 1, February 1995, p. 1-9.

Schopp, M, (1997), "User Modelling and Performance Evaluation of Distributed Location Management for Personal Communications Services," ITC-15, V. Ramaswami and P. E. Wirth (Editors), Elsevier Science Publishers B. V., 1997, p. 23-34.

Schopp, M, (2000), "Location management in a multi-network-operator environment," Telecommunication Systems, vol. 15, 2000, p. 63-78.

Schwartz, M, (1987), "Telecommunication Networks: Protocols, Modeling and Analysis," Addison-Wesley Publishing Company.

Schwartz, M & Saad, S, (1979), "Analysis of Congestion Control Techniques in Computer Communication Networks," from Grangé & Gien [1979], p. 113-130.

Skoog, R A, (1988), "Performance and Engineering of Common Channel Signaling Networks Supporting ISDN," from Bonatti & Decina [1988], p. 415-424.

Skoog, R A, (1991), "Study of Clustered Arrival Processes and Signaling Link Delays," Teletraffic and Datatraffic in a Period of Change ITC-13, A. Jensen and V. B. Iversen (Editors), Elsevier Science Publishers B. V., North-Holland, p. 61-66.

Smith, D E, (1994), "Effects of Feedback Delay on the Performance of the Transfer-Controlled Procedure in Controlling CCS Network Overloads," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 424-432.

Spragins, J D, Hammond, J L & Pawlikowski, K, (1991), "Telecommunications: Protocols and Design," Addison-Wesley Publishing Company.

Syski, R, (1986), "Introduction to Congestion Theory in Telephone Systems: North-Holland Studies in Telecommunications," Elsevier Science Publishers B.V., Holland.

Tabbane, S, (1995), "An Alternative Strategy for Location Tracking," IEEE Journal on Selected Areas in Communications, vol. 13, no. 5, June 1995, p. 880-892.

Tabbane, S, (1997), "Location Management Methods for Third-Generation Mobile Systems," IEEE Communications Magazine, August 1997, p. 72-84.

Tabbane, S, (1998), "Modelling the MSC/VLR processing load due to mobility management," Proceedings of ICUPC'98, Florence, Italy, October 1998.

Wang, J Z, (1993), "A Fully Distributed Location Registration Strategy for Universal Personal Communication Systems," IEEE Journal on Selected Areas in Communications, vol. 11, no. 6, August 1993, p. 850-860.

Willmann, G, (1989), "Modelling and Performance Evaluation of Multi-Layered Signalling Networks Based on the CCITT No. 7 Specification," Teletraffic Science for New Cost-Effective, Networks and Services ITC-12, p. 930-940.

Willmann, G & Kühn, P J, (1990), "Performance Modeling of Signaling System No. 7," IEEE Communications Magazine, July 1990, p. 44-56.

Wong, V W S & Leung, V C M, (2000), "Location Management for Next-Generation Personal Communication Networks," IEEE Network, September/October 2000, p. 18-24.

Zepf, J, Willmann, G & Rufa, G, (1991), "Transient Analysis of Congestion and Flow Control Mechanisms in Common Channel Signalling Networks," Teletraffic and Datatraffic in a Period of Change, ITC-13, A. Jensen and V. B. Iversen (Editors), Elsevier Science Publishers B. V., North-Holland, p. 413-419.

Zepf, J & Rufa, G, (1994), "Congestion and flow control in Signaling System No. 7 – Impacts of Intelligent Networks and New Services," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 501-509.

Bibliography

3GPP, (2002), "Technical Specification Group Core Network, Gateway Location Register (GLR) - Stage 2, (Release 5)," TS 23.119, v5.0.0.

3GPP, (2002), "Technical Specification Group Core Network, SS7 Signalling Transport in Core Network, (Release 5)," TS 29.202, v5.1.0.

Ajmone Marsan, M, De Carolis, G, Leonardi, E, Lo Cigno, R & Meo, M, (2000), "An approximate method for the computation of blocking probabilities in cellular networks with repeated calls," Telecommunication Systems, vol. 15, p. 53-62.

Barbour, A D, (1976), "Networks of Queues and the Method of Stages," Advances in Applied Probability, vol. 8, p. 584-591.

Benjamin, D A P, (1995), "Parallel Discreet Event Simulation," Teletraffic `95, 1995, p. 518-526.

Benmohamed, L & Meerkov, S M, (1993), "Feedback control of congestion in packet switching networks: The case of a single congested node," IEEE/ACM Transactions on Networking, vol. 1, no. 6, December 1993, p. 693-707.

Beutler, F J & Melamed B, (1978), "Decomposition and Customer Streams of Feedback Networks of Queues in Equilibrium," Operations Research, vol. 26, no. 6, p. 1059-1072.

Bunday, B D, (1986), "Basic Queueing Theory," Edward Arnold Ltd.

Chandy, K M, Howard J H & Towsley D F, (1977), "Product form and local balance in queueing networks," Journal of the Association for Computer Machinery, vol. 24, no. 2, p. 250-263.

Disney, R L & Kiessler, P C, (1987), "Traffic Processes in Queueing Networks – A Markov Renewal Approach," Johns Hopkins University Press.

Duffy, D E, McIntosh, A A, Rosenstein, M & Willinger, W, (1994), "Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 544-551.

Frost, V S & Melamed, B, (1994), "Traffic Modeling for Telecommunications Networks," IEEE Communications Magazine, March 1994, p. 70-81.

Gelenbe, E & Pujolle, G, (1987), "Introduction to Queueing Networks," John Wiley & Sons.

I, C L, Pollini, G P & Gitlin, R D, (1997), "PCS Mobility Management Using the Reverse Virtual Call Setup Algorithm," IEEE/ACM Transactions on Networking, vol. 5, no. 1, February 1997, p. 13-23.

ITU-T Recommendations, E505, (1992), "Measurements of the performance of common channel signalling network," Geneva.

ITU-T Recommendations, Q700, (1993), "Introduction to CCITT Signalling System No. 7," International Telecommunication Union, Geneva.

Jabbari, B, (1991), "Common Channel Signalling System Number 7 for ISDN and Intelligent Networks," Proceedings of the IEEE, vol. 79, no. 2, February 1991, p. 155-168.

Jabbari, B, (1992), "Routing and Congestion Control in Common Channel Signaling System No. 7," Proceedings of the IEEE, vol. 80, no. 4, April 1992, p. 607-617.

Jackson, J R, (1957), "Networks of waiting lines," Operation Research, vol. 15, p. 254-265.

Jackson, J R, (1963), "Jobshop like queueing systems," Management Science, vol. 10, October 1963, p. 131-142.

Kamoun, F & Kleinrock, L, (1980), "Analysis of Shared Finite Storage in a Computer Network Node Environment under General Traffic Conditions," IEEE Transactions on Communications, vol. COM-28, no. 7, July 1980, p. 992-1003.

Kelly, F P, (1975), "Networks of Queues with Customers of Different Types," Journal of Applied Probability, vol. 12, p. 542-554.

Kelly, F P, (1976), "Networks of Queues," Advances in Applied Probability, vol. 8, p. 416-432.

Kleinrock, L, (1974), "Queueing Systems – Volume 1: Theory," John Wiley & Sons.

Kleinrock, L, (1976), "Queueing Systems – Volume 2: Computer Applications," John Wiley & Sons.

Kobayashi, H, (1978), "Modeling and Analysis – An Introduction to system Performance Evaluation Methodology," Addison-Wesley Publishing Company.

Kühn, P J, Pack, C D, & Skoog, R A, (1994), "Common Channel Signaling Networks: Past, Present, Future," IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, April 1994, p. 383-393.

Lam, S S, (1977), "Queuing Networks with Population Size Constraints," IBM Journal of Research and Development, July 1977, p. 370-378.

Lemieux, C, (1979), "Flow Control in Switched Telephone Networks: Theory and Experience. Extension of Theory to Packet Switched Networks," from Grangé & Gien [1979], p. 1-16.

Lipper, E H & Rumsewicz, M P, (1994), "Teletraffic Considerations for the Widespread Deployment of PCS," IEEE Network, vol. 8, no. 5, September/October 1994, p. 40-49.

Matsumoto, J & Mori, H, (1981), "Flow Control in Packet-Switched Networks by Gradual Restrictions of Virtual Calls," IEEE Transactions on Communications, vol. COM-29, no. 4, April 1981, p. 466-473.

Melamed, B, (1979), "Characterizations of Poisson Traffic Streams in Jackson Queueing Networks," Advances in Applied Probability, vol. 11, p. 422-438.

Mitra, N & Usiskin, S D, (1991), "Relationship of the Signaling System No. 7 Protocol Architecture to the OSI Reference Model," IEEE Network Magazine, January 1991, p. 26-36.

Mouley, M & Pautet M B, (1992), "The GSM System for Mobile Communications," CELL & SYS, France.

Muntz, R R, (1972), "Poisson Departure Processes and Queueing Networks," IBM Research Report, RC-4145, IBM Thomas J. Watson Research Centre, Yorktown Heights, New York.

Paterok, M, Herzog, U & Bleisteiner, C, (1989), "The Influence of Repeated Calls on the Performance Measures of Loss Systems," Teletraffic Science for New Cost-Effective, Networks and Services ITC-12, M. Bonatti (Editor), Elsevier Science Publishers B. V., North-Holland, 1989, p. 1413-1419.

Pearce, C E M, (1987), "On the Problem of Re-attempted Calls in Teletraffic," Communications on Statistics, vol. 3, no. 3, p. 393-407.

Pouzin, L, (1981), "Methods, Tools, and Observations on Flow Control in Packet-Switched Data Networks," IEEE Transactions on Communications, vol. COM-29, no. 4, April 1981, p. 413-426.

Rudin, H, "Foreword – Congestion Control: Preview and Some Comments," IEEE Transactions on Communications, vol. COM-29, no. 4, April 1981, p. 373-375.

Schlanger, G G, (1986), "An Overview of Signaling System No. 7," IEEE Journal on Selected Areas in Communications, Vol. SAC-4, 3, 1986, p. 360-365.

Scholtz, F J, (1997), "Statistical Analysis of Common Channel Signalling system No. 7 Traffic," ITC-15, V. Ramaswami and P. E. Wirth (Editors), Elsevier Science Publishers B. V., p. 1229-1236.

Schweitzer, P J, & Lam, S S, (1976), "Buffer Overflow in a Store-and-Forward Network Node," IBM Journal of Research and Development, November 1976, p. 542-550.

Scourias, J, (1996), "Overview of GSM: The Global System for Mobile Communications," Department of Computer Science, University of Waterloo, Technical Report CS-96-15, March 1996.

Skoog, R A, (1989), "Engineering Common Channel Signaling Networks for ISDN," Teletraffic Science for New Cost-Effective, Networks and Services ITC-12, M. Bonatti (Editor), Elsevier Science Publishers B. V., North-Holland, p. 915-921.

Russell, T, (1995), "Signalling System #7," McGraw-Hill, New York.

Unger, B W, Goetz, D J & Maryka, S W, (1994), "Simulation of SS7 Common Channel Signaling," IEEE Communications Magazine, March 1994, p. 52-62.