MATHEMATICAL FOUNDATIONS OF NEXT GENERATION WIRELESS SYSTEMS

by

Hüseyin Birkan Yılmaz

B.S., Mathematics, Boğaziçi University, 2002

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2006

DEDICATION

This thesis is dedicated to

my parents

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who gave me the possibility to complete this thesis. To those individuals I neglect to mention here by name, I still offer my deepest thanks.

First and foremost, I would like to thank my thesis supervisor Assist. Prof. Tuna Tuğcu for many insightful conversations during the development of the ideas in this thesis, and for helpful comments on the text. Without his support, it would be impossible for me to finish this thesis.

I would like to thank Prof. Feodor Vainstein and Assoc. Prof. Alp Eden for their ideas and support during investigation of the mathematical background of this thesis. I am grateful to Prof. Cem Ersoy and Assoc. Prof. Sema Oktuğ for their participation in my thesis committee.

I am grateful to all my friends from Boğaziçi University, especially Deniz Karlı, Yunus Dönmez, Sinan Işık, Recep Duygu Ant, Bora Ferlengez, Yasemin Kara, Evren Önem, Görkem Özkaya, Murat Sağlam, İhsan Ata Topaloğlu, Serdar Sözübek, and Filiz Tümel for their continued moral support. From the staff, Sevgi Dikmen is especially thanked for her care and attention. I am also grateful to Başak Canbak for her moral support at the critical stages of the thesis.

Last, but not least, I would like to thank my family, Milver Yılmaz, Ahmet Yılmaz, and my sister Birgül Yılmaz, for their love, patience, and understanding. They gave me their love and support throughout all stages of my education.

This research is partially supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under grant number $104E032$ and Bogazici University Research Fund under grant number $BAP04S104$.

ABSTRACT

MATHEMATICAL FOUNDATIONS OF NEXT GENERATION WIRELESS SYSTEMS

The heterogeneous structure of *Next Generation Wireless Systems (NGWS)* makes admission control very complex. Accessibility of the subsystems at the time of connection or handoff request, availability of resources in the subsystems, user preferences, and connection class need to be considered in admission control. In this thesis, first we give a general connection admission control algorithm. We also propose the first analytical model in the literature for admission control in NGWS.

NGWS consists of many subsystems. Our analytical model has no limitation on the number of subsystems in NGWS. Due to the existence of multiple subsystems, states of the system are more complex than the states in ordinary homogeneous wireless systems. After defining NGWS and states, we point out the major challenges in modeling for NGWS and evaluating state probabilities.

For evaluating the state probabilities, we present an analytical solution approach. Since the state space explodes, we propose a neat solution to calculate the state probabilities in a reasonable way even as the state space grows. Our proposed solution is based on reasonable assumptions.

ÖZET

GELECEK NESİL KABLOSUZ SİSTEMLERİN MATEMATİKSEL ALTYAPISI

Gelecek Nesil Kablosuz Sistemlerin (GNKS) ayrışık yapısı sisteme kabul kontrolünü çok karmaşıklaştırır. Sisteme kabul kontrolünde bağlantı kurma veya bağlantı el değiştirme isteği sırasında alt sistemlerin ulaşılabilirliğinin, kullanıcının tercihlerinin, ve bağlantı sınıfının hesaba katılması gereklidir. Bu tezde ilk önce genel bir bağlantı kabul kontrol algoritması veriyoruz. Aynı zamanda GNKS dahilinde sisteme kabul kontrol için literatürdeki ilk analitik modeli öneriyoruz.

GNKS birden çok alt sistemden oluşur. Analitik modelimizde alt sistem sayısı üzerine bir limit yoktur. Birden çok alt sistemin varlığından dolayı sistemin durumları alışılmış türdeş kablosuz sistemlerin durumlarından daha karmaşıktır. GNKS ve durumlarını tanımladıktan sonra, GNKS için modellemenin ve durum olasılıklarının hesaplanmasının başlıca zorluklarına değiniyoruz.

Durum olasılıklarını hesaplamak için analitik bir çözüm yaklaşımı sunuyoruz. Durum uzayı aşırı büyüse de durum olasılıklarını makul bir yolla hesaplamak için akıllı bir çözüm öneriyoruz. Önerdiğimiz analitik çözüm de kabul edilir varsayımlara dayanmaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

$[\mathbf{P}_i(t)]_{g_i \in \mathcal{G}}$	Vector of state probabilities of all states at time t
$\mathcal{A}$	Set of all subsets of $\mathcal{B}$
$\mathcal{A}(b)$	Set of areas constituting the cell of access node b
a	Area, that is a subset of $\mathcal{B}$
$\bar{a}$	Physical area
$\mathcal{B}$	Set of all access nodes in the service area
$\widetilde{b}$	Mapping from Ω to $\mathcal{A}$
b_i^s	i^{th} access node of subsystem s
$bw(k)$	Bandwidth requirement of class k connection
$\mathcal{C}$	Set of connection classes in NGWS
c_i^s	Capacity of access node b_i^s
d	Metric that represents the difference between two probability vectors
$f(a_i, k, t)$	Connection profile of class k connections in area a_i at time t
Fr0	States with no free channels in any cell
Fr1	States with one free channel in one of the cells
Fr2	States with two free channels available in all network
g	State of the system
$h_a^k(t)$	Hangup rate of class k connections in area a
K	Resource capacity
$\mathcal{L}_{ac}(rq)$	Ordered list of access nodes specified in request rq
l_i^s	Current load of b_i^s
$\widehat{l_i^s}(rq)$	New load of b_i^s if request rq is accepted
$l_j'^t$	Recorded value of l_j^t at access node b_i^s
$\widehat{l_j'^t}(rq)$	Load of access node b_j^t if rq is accepted, based on recorded value $l_j'^t$
$\overline{m}$	Mean of the number of calls
N_c	Maximum number of ongoing calls
$n_{a_i}(t)$	Number of users in area a_i at time t

$\mathbf{P_b}$	Grade of the service
$\mathbf{P_c}$	Conditional probability
$\mathbf{P_c}(blocking)$	Conditional probability of blocking new connection attempts
$\mathbf{P_c}(dropping)$	Conditional probability of dropping active connections
$P_{handoff}\,(b,k,\Delta t)$	Probability that a handoff attempt of class k occurs for access node b
$P_{hangup}\,(b,k,\Delta t)$	Probability that MT terminates the connection voluntarily
P_{hd}	Target handoff dropping probability
$\mathbf{P}_i\,(t)$	Probability that the system is in state g_i at time t
$\mathbf{P}_i^{'}\,(t)$	Rate of change in state probability of g_i during Δt
$P_{inter}\,(b,k,\Delta t)$	Direct inter-subsystem handoff probability of class k connection through access node b
$\widetilde{P}_{inter}\,(b,k,\Delta t)$	Indirect inter-subsystem handoff probability of class k connection through access node b
$P_{intra}\,(b,k,\Delta t)$	Intra-cell movement and intra-subsystem handoff probability of class k connection over access node b
P_{nb}	Target new call blocking probability
$P_{new}\,(b,k,\Delta t)$	Probability that a new connection attempt of class k for access node b
$P_{ni}\,(b,k,\Delta t)$	Probability that a direct incoming new connection attempt of class k for access node b
$\widetilde{P}_{ni}\,(b,k,\Delta t)$	Probability that an indirect incoming new connection attempt of class k for access node b
$P_{no}\,(b,k,\Delta t)$	Probability that a direct outgoing new connection attempt of class k for access node b
$p(a_i,k,s,t)$	Probability that a user in area a_i with a class k connection prefers subsystem SS^s at time t
$p(k,s)$	Probability that MT prefers subsystem s for a class k connection
R	Current allocated resources
R_{th}	Threshold for resource allocation
$R^k\,(v)$	Probability that access node v rejects a request of class k
$r(t)$	Number of arrived calls in time interval $[t-1,t)$

$r_{a_i}^k(t)$	Connection generation rate of class k connections in area a_i at time t
rq	Request
$\mathbb{S}$	Set of subsystems
s	Subsystem in NGWS
$s(b)$	Subsystem to which access node b belongs
$s(t)$	Number of blocked calls in time interval $[t-1, t)$
$\mathbb{V}(b_i^s)$	Vicinity of b_i^s
$V_{a_j,a_i}^k(t)$	Migration rate of a class k connection from area a_j to a_i at time t
w	Weight for exponential weighted moving average
$\mathcal{X}$	Set of all states
$x_a^k(b)$	Number of active connections of class k communication with access node b in area a
$\alpha(b)$	Total probability that MT prefers other access nodes serving the same area, but those access nodes cannot accommodate the request due to lack of resources
ΔR	Increase in resource allocation
ε	Threshold for convergence
$\Gamma(e, s)$	Transition graph for elementary event e and subsystem s
λ_{ij}	Rate of flow from state g_i to g_j
$\mathcal{V}$	Set of arcs in transition graphs
Ω	Service area
σ	Variance for the number of calls
$\sim$	Equivalence relation on Ω introduced by $\widetilde{b}$

3G	Third Generation Cellular Systems
4G	Fourth Generation Cellular Systems
ABA	Adaptive Bandwidth Allocation
AP	Access Point
BLR	Boundary Location Register
BSC	Base Station Controller
BTS	Base Transceiver System

CAC	Call Admission Control
CGR	Connection generation rate (iteration parameter)
DECT	Digital Enhanced Cordless Telephone
DS-CDMA	Direct-Sequence Code Division Multiple Access
FDMA	Frequency Division Multiple Access
GEO	Geostationary Earth Orbit
GRX	GPRS Roaming eXchange
GSM	Global System for Mobile communications
HA	Home Agent
HLR	Home Location Register
HR	Home Register
HUP	Hangup rate (iteration parameter)
IMS	IP-based Multimedia Systems
IP	Internet Protocol
IS-41	Interim Standard 41
LAN	Local Area Network
LEO	Low Earth Orbit
LR	Location Register
MR	Migration rate (iteration parameter)
MSC	Mobile Switching Center
MT	Mobile Terminal
NGCAC	Next Generation Call Admission Control
NGUA	Next Generation User Address
NGWS	Next Generation Wireless Systems
NUSR	Number of users (iteration parameter)
PCS	Personal Communication System
PLMN	Public Land Mobile Network
PSTN	Public Switched Telephone Network
QoS	Quality of Service
SIP	Session Initiation Protocol
SIR	Signal-to-Interference Ratio

SLA	Service Level Agreement
TDMA	Time Division Multiple Access
UMTS	Universal Mobile Telecommunications System
WAN	Wide Area Network
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network

1. INTRODUCTION

Wireless networks are voice and data networks that use radio transmission at their physical layer. The radio component of a wireless network is augmented with the core network, which typically utilizes wired connections. Wireless communication over the radio link is multiplexed on the core network. Wireless networks offer advantages like mobility, flexibility, reduced cost of management, scalability, and easy installation. Wireless technologies continue to advance and offer users higher standards by providing mobility. Wireless networks increase the ability to interact with each other as well as remove distance and time barriers.

Wireless networks include many technologies, systems, and services optimized for a variety of applications [1]. Wireless applications aim to provide similar services to their desktop counterparts. Examples of these applications include voice, e-mail, multimedia teleconferencing, electronic newspaper, location-based information services, and nomadic computing. These applications are expected to operate over a variety of environments such as *Wireless Local Area Networks (WLAN)*, *Personal Communication Systems (PCS)*, and satellite systems. WLAN provides wireless local access over unlicensed radio spectrum. Wireless access points, connected via a wireline network at the backbone, attach mobile users to the wired network. A cellular network is a radio network made up of a number of cells each served by a fixed transmitter. These cells provide radio coverage over area wider than one cell. Cellular networks offer a number of advantages such as increased capacity, reduced power usage, and better coverage.

In all kinds of wireless networks, radio spectrum constitutes the bottleneck. In the case of *Frequency Division Multiple Access (FDMA)* systems, scarce resource is the frequency band. For *Time Division Multiple Access (TDMA)* systems, the time slots constitute the scarce resource. In the case of *Code Division Multiple Access (CDMA)* systems, *Signal-to-Interference Ratio (SIR)* determines the number of simultaneously transmitting users. Accepting or rejecting decisions are determined by *access nodes*. In WLANs, the access nodes are *access points*, in PCS access nodes are *base stations*

and in the satellite networks access nodes are *transponders*.

It is important that a user's ongoing connection is not interrupted and service is available while the user is mobile. Wireless systems must use radio resources efficiently to maximize utilization and user satisfaction simultaneously. Achieving such a goal requires careful reallocation of resources, hence, importance of the call admission control arises. The duty of the call admission control is to use network resources in an optimized way to utilize the system with acceptable service quality.

1.1. Call Admission Control in Wireless Networks

The *Call Admission Control (CAC)* is the mechanism that decides whether a request is admitted into the system or not. In this section, we focus on the general CAC model. We examine the traditional approaches such as the threshold-based approach, pricing-based approach, mobility-based approach, and approaches based on estimation and prediction. We explain how these approaches strive to provide high utilization subject to user satisfaction.

1.1.1. General Model

The decision process of CAC can often be formulated by a high level representation. Whenever a user has a new connection request, the CAC policy takes the call request as input, and based on the system load and connection properties, decides whether or not to accept the request. CAC policy in Figure 1.1 illustrates the general CAC algorithm.

CAC policy determines system utilization, service quality, and user satisfaction. The concept of CAC policy is applicable to both hard capacity and soft capacity wireless systems. For hard capacity systems, a channel is defined as the time slot in TDMA systems or the frequency band in FDMA systems. In hard capacity systems, increase in resource usage by accepting an additional request can be determined by the number of channels required for the specific incoming call. On the other hand, for soft capacity

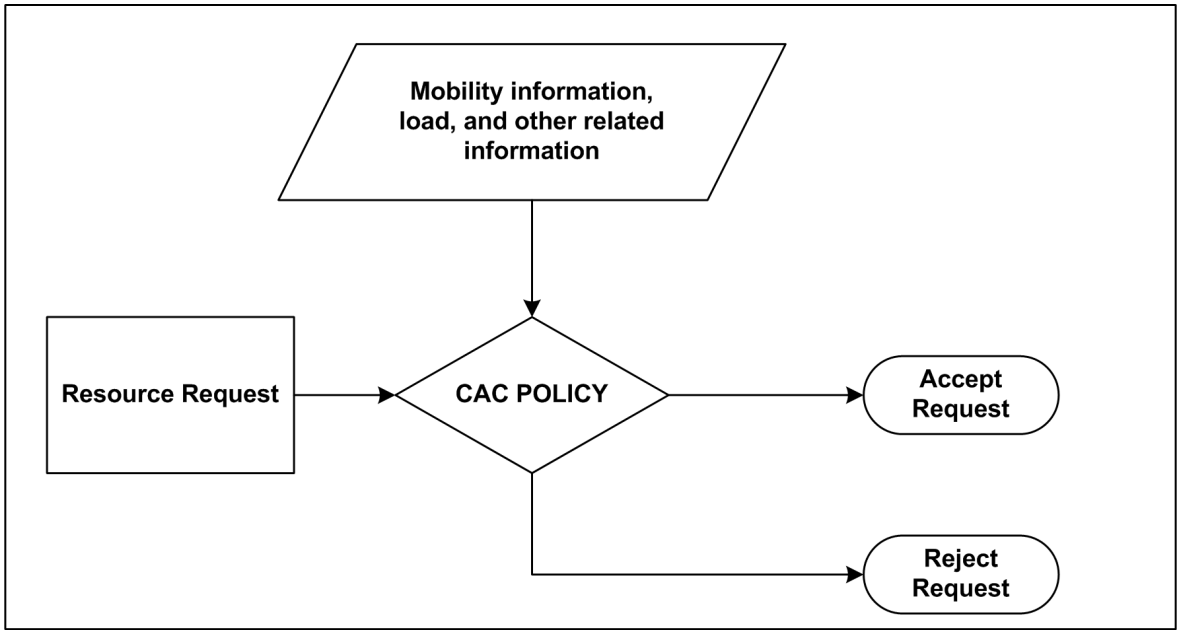


Figure 1.1. The CAC decision process

systems, there is no fixed capacity limit; the acceptance decision is given based on the current interference measurements (e.g., DS-CDMA systems).

In cellular systems, user requests can be classified in two classes: new calls and handoff calls from adjacent cells. From the user's point of view, being blocked in the beginning is less annoying than being dropped in the middle of a connection [2]. Hence, it is apparent that handoff calls should be given higher priority than new calls. However, there is always a tradeoff between new calls and handoff requests. If new calls are given a higher priority which means not reserving resources for handoff calls, then the system is highly utilized but probabilities of blocking and dropping increase. On the other hand, if handoff calls are given a higher priority, then the system is less utilized with lower probabilities of blocking and dropping. A highly utilized system that provides an acceptable level of *Quality of Service (QoS)* is a desired solution [3]. However, high resource utilization and QoS provisioning are always conflicting goals. As a result, the CAC policy has to be optimized, to provide a satisfactory level of QoS with the maximum utilization. Analytical models help comparing of the CAC policies and evaluation of QoS metrics to optimize the CAC policy.

Due to the characteristics of the mobile environment, the traffic load at an access node can change noticeably from time to time. As a result, providing consistent QoS satisfaction to all mobile users with heterogeneous requirements simultaneously is a complex job. CAC plays an important role in providing satisfactory QoS. Analytical modeling can be used as a tool to estimate values for QoS provisioning.

1.1.2. CAC Schemes in the Literature

Traditional CAC policy approaches in cellular networks vary from the simplest CAC policies to more complicated schemes. These algorithms are mostly developed for homogenous wireless networks. We first discuss threshold-based approaches, then we focus on collaborative approaches based on estimation, non-collaborative approaches based on prediction, mobility-based approaches, and pricing based approaches.

1.1.2.1. Threshold-based Approach. Threshold-based CAC approaches can be applied to both hard capacity and soft capacity wireless systems. Threshold-based CAC algorithms are based on the availability of resources, R [4]. R represents the number of currently allocated resources in hard capacity systems. In the case of soft capacity systems, it can be derived from SIR at the receiver. The objective of a threshold-based CAC algorithm is to keep R less than the threshold, R_{th} [5]. When a new resource request arrives, the algorithm estimates the increase ΔR caused by this new request. Generally, the CAC policy is based on the condition

$$R + \Delta R \leq R_{th}. \tag{1.1}$$

If Equation 1.1 is satisfied, the incoming request is accepted; otherwise, the request is rejected or queued [5]. In the simplest case, R_{th} is taken as the capacity, so Equation 1.1 reduces to

$$R + \Delta R \leq K, \tag{1.2}$$

where K stands for capacity. Thus, a request is granted if there are sufficient resources for the request. Thus, the system utilization is increased, but the service quality may suffer under high load.

To prioritize handoff calls over new calls, CAC policy does not use all available capacity. Some channels are reserved for high priority handoff calls, and low priority new calls may be rejected although there are available resources in the system. In this case, R_{th} is not equal to the capacity, so $K - R_{th}$ channels are reserved for handoff calls. This approach is called the *guard channel approach*. In this approach, a new call is accepted if allocation of ΔR channels do not exceed R_{th}, while a handoff call is accepted as long as there are enough channels. Thus, handoff calls are prioritized over new calls by utilizing R_{th}. To keep dropping rate and utilization at reasonable levels R_{th} must be chosen with care.

Compared to dynamic threshold, the guard channel scheme with static threshold is easier to implement, but it suffers from fluctuations in offered load. Higher utilization can be achieved by changing R_{th} according to the whole or partial network state. In [6], a more complex approach, the *fractional guard channel approach* is introduced. In this approach, the incoming request is accepted with a probability that depends on R. This approach helps keep the handoff call dropping probability lower and also avoids congestion while achieving better utilization. All guard channel approaches help lower the dropping rate. However, they generally decrease system utilization. Though there are more intelligent approaches in the literature, they are not used in practice due to high overhead.

1.1.2.2. Collaborative Approach Based on Estimation. In this approach, decisions are made locally with the aid of information gathered from neighboring cells. In [7], Epstein *et al.* propose a collaborative algorithm that uses estimates of call dropping and call blocking probabilities. The maximum number of ongoing calls N_c is estimated

from

$$P_{hd} = \frac{1}{2} \operatorname{erfc}\left(\frac{N_c - \overline{m}}{\sigma}\right), \tag{1.3}$$

where P_{hd} denotes the target call dropping probability, $\overline{m}$ and σ are the mean and the variance of the number of calls, respectively. The call blocking probability $P_{nb}(t)$ during time interval $[t-1, t)$ is estimated locally as

$$P_{nb}(t) = (1-w)P_{nb}(t-1) + w\frac{s(t)}{r(t)}, \tag{1.4}$$

where $s(t)$ and $r(t)$ are the number of blocked calls and the number of calls that arrived during $[t-1, t)$, respectively, and w is the weight used to calculate the exponential weighted moving average. The decision on an incoming call is made based on Equations 1.3 and 1.4 [5].

1.1.2.3. Non-Collaborative Approach Based on Prediction. In the case of high user mobility, information exchange with neighbor cells may cause significant control overhead. In such a situation, only local information is used to predict the resource requirement in the future [8]. Such a local predictive approach to CAC is shown to perform as good as a collaborative approach when the traffic fluctuation is moderate [5]. The prediction of future resource requirements is done without collaboration of neighboring cells.

1.1.2.4. Mobility-based Approach. The main idea in mobility-based approaches is to exploit user mobility information for better resource management. The *shadow cluster* concept is introduced in [9] based on user mobility information. In the *shadow cluster* technique, every mobile terminal with an active wireless connection exerts an influence on the cells in the vicinity of its current location according to call holding time, current direction, velocity, and position. In [2], the shadow cluster approach is extended from TDMA/FDMA systems to the soft capacity DS-CDMA systems.

Mobility-based approaches advance efficiency under average mobility, but calculating the amount of incoming traffic for a particular cell is nontrivial. This is even harder under high mobility conditions. Furthermore, real-time exchange of control messages among cells incurs a large communication overhead [5].

1.1.2.5. Pricing-based Approach. A pricing-based approach is proposed in [10], where the objective is to maximize the utility of wireless resources. The utility is defined as the user's level of satisfaction with perceived QoS. Maximizing the utility means maximizing user's level of satisfaction, which implies more resource allocation to each user. In contrast, for maximizing the revenue under flat rate pricing, it is necessary to accommodate more users. Therefore, for optimal CAC policy, the optimal operation point must be found.

In [10], the optimal point between utility and revenue is determined in terms of the new call arrival rate, and a pricing scheme is developed to achieve this optimal effective arrival rate in the network. QoS metric P_b referred to as the *grade of service* is defined as

$$P_b = \alpha P_{nb} + \beta P_{hd}, \tag{1.5}$$

where α and β are the weights for the new call blocking and handoff call dropping probabilities, respectively, and $\alpha + \beta = 1$.

The metric P_b can be defined as a function of new call arrival rate λ_n. The utility function depends on function of λ_n since utility is a function of P_b. Assuming flat rate pricing, the revenue depends on the new call arrival rate $f(\lambda_n)$. Hence, the total utility becomes

$$U(\lambda_n) = f(\lambda_n) \times h(P_b) = f(\lambda_n) \times h[g(\lambda_n)]. \tag{1.6}$$

The optimal value of the new call arrival rate λ_n^* that maximizes total utility can be

calculated by differentiating $U(\lambda_n)$ and considering the point where it is zero. Based on this optimal new call arrival rate, the pricing scheme is developed by changing the price of a call. By changing the price at the peak-hours, the new call arrival rate is adjusted. The reaction of users to the change in price is also modeled in [10].

1.1.3. Motivation

Call admission control plays an important role in resource utilization and user satisfaction. Many CAC algorithms are proposed for wireless and wired networks. In the case of *Next Generation Wireless Systems (NGWS)*, the heterogeneous structure imposes significant complexity in the design of the CAC policy and modeling. Our motivation in this thesis is to develop an analytical model that will help system designers calculate the dropping and blocking probabilities in NGWS. We use markov chains for modeling and solving the state probabilities. Since the state space explodes, we just focus on the states that are critical for the system designer. Focusing on the critical states reduces the calculations and space requirements to evaluate conditional state probabilities.

1.2. Heterogeneous Wireless Systems

Heterogeneous wireless systems merge multiple subsystems with different access technologies to provide high bandwidth access anytime, anywhere. Different subsystems with different access technologies have different constraints and capacities. Mobile users demand nomadic access to high-speed data and multimedia services from heterogeneous wireless systems.

Currently, there exist dissimilar wireless networks, such as Bluetooth for personal areas, WLANs for local areas, *Universal Mobile Telecommunications System (UMTS)* for metropolitan coverage, and satellite networks for global networking. These networks are designed for specific service needs and vary extensively in terms of bandwidth, latency, area of coverage, cost, and quality of service provisioning. For example, satellite networks can provide global coverage, but are limited to outdoors with high cost and

long propagation delays (from 20-25 ms for *Low Earth Orbit (LEO)* satellites to 250-280 ms for *Geostationary Earth Orbit (GEO)* satellites). *Third Generation (3G)* wireless systems like UMTS can deliver a maximum data rate of 2 Mbps at lower cost and have wide areas of coverage. WLANs have higher data rates, but can support only mobiles moving at lower speeds in smaller areas. Therefore, none of the existing wireless systems can simultaneously guarantee low latency, high bandwidth, and ubiquitous coverage needs of mobile users at low cost. This imposes a new direction in the design of NGWS [4].

1.2.1. Design Goals of Heterogeneous Wireless Systems

Next generation wireless systems are heterogeneous in structure. The needs of the mobile user and heterogeneity play an important role in the design of NGWS. The integrated NGWS keeps the best features of the individual networks: the global outdoors coverage of satellite networks, the wide mobility support of 3G systems, and the high speed and low cost of WLANs. At the same time, it eliminates the weaknesses of the individual systems. For example, limited data rate of 3G systems can be overcome when WLAN coverage is available. The basic idea is to use the best available network at any time [4].

We can summarize NGWS design goals as

- support for the best network selection based on user's service needs and profile,
- mechanism to ensure high-quality, security, and privacy,
- protocol to guarantee seamless inter-system mobility,
- scalable architecture (integration of any number of wireless systems of different service providers),
- QoS provisioning.

Design goals of NGWS shape the network architecture. There is a need for a new architecture to achieve roaming among heterogeneous networks of different service providers who may not necessarily have direct *Service Level Agreements (SLAs)*

with each other. The architecture should utilize the existing infrastructure as much as possible. This ensures economical and fast deployment. The architecture should be able to integrate any number of wireless systems of both existing and future service providers. The architecture should be transparent to different access technologies of different types of networks. The architecture should support seamless mobility management to eliminate connection interruption and QoS degradation during inter-system roaming [4].

1.2.2. Mobility Management in Heterogeneous Wireless Systems

One of the research challenges for next generation wireless systems is the design of intelligent mobility management techniques to achieve global roaming among various access technologies [11]. Mobility management contains two components: location management and handoff management [12]. In NGWS, there are two types of roaming for *Mobile Terminals (MTs)*: intrasystem and inter-system roaming. Intrasystem roaming refers to moving between different cells of the same system. Intrasystem mobility management techniques are based on similar network interfaces and protocols. Inter-system roaming, on the other hand, refers to switching between different backbones, protocols, techniques, or service providers [11].

Location management and resource management (which includes handoff management) are components of mobility management. Location management deals with tracking the locations of MTs between consecutive communications. On the other hand handoff management deals with keeping the connection active while MT is moving from one cell to another. Location management includes two major tasks. The first is *location registration*, where MT periodically informs the system to update relevant location databases. The second is *call delivery*, where the system determines the current location of the MT based on the information available at the system database when a communication for the MT is initiated [11].

1.2.3. Heterogeneous Wireless Systems in the Literature

The concept of integrating two or more communication systems to get better performance is already in use and there is ongoing research. The existing integration architectures address the following issues: integration of two specific systems, integration of two general systems, integration of networks of multiple operators but of the same technology [4].

In [13] and [14], specific pairs of different systems are integrated through an additional gateway, such as interworking of *Digital Enhanced Cordless Telephone (DECT)* with GSM and of IS41 with GSM. The additional gateway takes care of interworking and inter-operability issues such as transformation of signaling formats, authentication, and retrieval of user profiles. Similarly, the integration of satellite and terrestrial networks is studied in [14]. Interworking units are placed between the satellite and terrestrial systems. In addition, different architectures are proposed to integrate WLAN and 3G systems [15]. All of the architectures above integrate only specific pairs of systems, so they are not scalable to integrate multiple systems.

In [16], *Boundary Location Register (BLR)* approach is introduced to integrate any two adjacent networks with partially overlapping areas. However, this approach is designed for neighbor networks rather than overlapping networks and it is not scalable in the sense that one BLR gateway and SLAs between systems are needed for each pair of adjacent networks when integrating multiple networks. Furthermore, the GSM Association has proposed an *inter-Public Land Mobile Network (inter-PLMN)* backbone using *GPRS Roaming eXchange (GRX)* to globally integrate the GPRS networks deployed by various providers [17]. This architecture uses multiple peer GRX nodes for connecting several GPRS networks. This architecture is, however, limited only to GPRS technology.

In the SMART project, a new architecture is proposed to integrate heterogeneous wireless systems [18]. This architecture uses two distinct networks: *basic access network* and *common core network* for signaling and data traffic, respectively. This architecture

is scalable, but requires the development and deployment of new basic access and common core networks, and hence is not cost-effective.

In [19] and [20], heterogeneous network integration using mobile IP and *Session Initiation Protocol (SIP)* are proposed. However, these architectures do not have any mechanism to select the best available network. Although mobile IP and SIP are used to carry out inter-system handoff, seamless support of inter-system handoff is not always guaranteed [11].

Several NGWS architectures are summarized in [21]. Proposed architectures are examined: NTT DoCoMo proposes architecture for NGWS, which is an extension of the current 3G architecture [22]. A next generation wireless communication architecture that is comprised of old and new wireless communication architecture standards has been presented in [23]. Telefonica's NGWS architecture is composed of WLANs cellular networks, personal area networks, and distribution networks organized in a layered structure [24]. In the Wine Glass project, [25], WLANs and UMTS are merged into a next generation wireless network. Furthermore, Siemens [26] adopts 3GPP's *IP-based Multimedia Systems (IMS)* specifications [27] and defines its own next generation wireless network architecture. Integrating multiple subsystems into one NGWS brings many challenges ranging from interworking among inherently different wireless subsystems to QoS provisioning [25], [28]. To the best of our knowledge, none of the above architectures satisfy all the requirements of the NGWS simultaneously [4].

Architecture design goals of NGWS give rise to research areas like architecture design, CAC, mobility management, integration of different radio technologies, and network layer design. For inter-operation of different communication protocols, an adaptive protocol suite is required that will adapt itself to the characteristics of the underlying networks and provide optimal performance across a variety of wireless environments [29]. IP is recognized to become the core part of NGWS to support ubiquitous communications. For upcoming wireless systems, even the air interface will be packet-based. Furthermore, adaptive terminals in conjunction with "smart" base stations will support multiple air interfaces and allow users to seamlessly switch between different

access technologies [11].

1.3. Challenges in CAC for NGWS

The diverse QoS requirements and the presence of different wireless access technologies pose significant challenges in designing a CAC algorithm for NGWS [5].

1.3.1. Heterogeneous Networking

NGWS design goals imply that a call in one particular network must be able to roam and be handed over to another network transparently. The usual handoff initiation based on signal strength typically is not enough. Other subsystem parameters such as the congestion level at the network must be considered as well [2], [5]. When a call is handed over to another subsystem, the state of the NGWS changes. Hence system states in NGWS are more complicated than the states in ordinary homogeneous networks. The number of states grows exponentially as the number of subsystems increases. From the CAC point of view, a vertical handoff results in a new type of handoff call. A CAC algorithm must determine the priority of this type of call over new calls. A new performance metric, *vertical handoff call dropping probability* is defined in [5].

1.3.2. Multiple Classes of Services

A CAC algorithm should be designed to support multiple classes of service and each class should allow different QoS requirements. These service classes may be subject to different billing and prioritization. The service classes include information that can be used by the CAC algorithm. Connection of different classes may be prioritized by using multiple thresholds. Furthermore, properties of classes can be used for price-based CAC scheme as mentioned in Subsection 1.1.2.

1.3.3. Adaptive Bandwidth Allocation

Due to the diversity of applications and QoS requirements for mobile users and the dynamic nature of wireless channel quality, *Adaptive Bandwidth Allocation (ABA)* can be used to improve the utilization of wireless network resources [5]. According to the network load, resources assigned to connections can be increased or decreased to maintain the call dropping and blocking probabilities at the required level for QoS provisioning. In [30], an ABA algorithm that allocates a target level of bandwidth to a connection is proposed. ABA can also be used in vertical handoff. The acceptable bandwidth can be negotiated, and CAC algorithm can be based on the result of negotiation [5].

1.3.4. Cross-layer Design

For a wireless network, cross-layer optimization can lead to significant improvements in the performance of transmission protocol stack [31]. In contrast to a traditional voice-oriented circuit-switched network, QoS should be described in terms of both call-level and packet-level performance metrics in a purely packet-switched wireless network [5]. Therefore, a new request is admitted only if the QoS measurements of all ongoing calls including the incoming request in terms of packet-level and call-level performance can be maintained at the desired level [5].

1.4. Contribution of the Thesis

The contribution of our thesis can be summarized as follows:

- We propose an analytical model for CAC in NGWS. Up to our knowledge, this is the first analytical model for CAC in NGWS.
- Unlike the previous work in the literature, our NGWS architecture is not designed for specific subsystems.
- We point out the major challenges in modeling for NGWS.
- Due to the overlaid structure of NGWS, the state space explodes. We employ a

neat technique to calculate only the state probabilities of the important states.

1.5. Structure of the Thesis

So far we introduced CAC concepts in wireless networks and focused on related work and motivation. In Chapter 2, we examine NGWS and we give basic definitions of NGWS and focus on the architecture. In Chapter 3, we detail our CAC scheme for NGWS. We develop the analytical model of CAC scheme in Chapter 4. Here, we introduce the system definition and elementary events. We also present how the transition probabilities can be evaluated. In Chapter 5, we provide two techniques to calculate the state probabilities: the analytical approach and a practical approach, which is still an analytical approach with some assumptions. In Chapter 6, we present numerical results for an example topology. Finally in Chapter 7, we summarize the thesis and suggest future work.

2. NGWS DEFINITIONS AND ARCHITECTURE

Though CAC in wireless networks has been studied extensively, the heterogeneous structure of NGWS makes CAC very complex. Since it is not likely that a single emerging technology will be able to provide continuous coverage indoors and outdoors, NGWS will exploit the advantages of multiple wireless networks as subsystems to achieve all design goals mentioned in Subsection 1.2.1. Since the mobile terminal may have access to alternative subsystems simultaneously, the selection of the subsystem for connection setup plays an important role in system performance. In this thesis, we focus on properties related to QoS provisioning at connection setup level. The selection of the subsystem for connection establishment and handoff is a key factor in the performance of NGWS.

2.1. NGWS Introduction

Each subsystem in NGWS serves as an access network for the users. Since the service area of subsystems overlap, the MTs will generally have access to multiple subsystems simultaneously in overlapping regions. Physical position of the MT and signal propagation are the key factors in identifying the accessible subsystems. WLANs, PCS, satellite systems, and their future variations together with emerging new technologies like 4G Mobile, WiMAX, and IEEE 802.20 are candidates as subsystems of NGWS. In the literature, there are various proposals for NGWS architecture. Our architecture is flexible and is independent of the technologies of the individual subsystems [21]. There is no limit on the number of subsystems in our architecture; it supports multiple classes of data traffic and is extendible in the sense of underlying subsystems. Basic properties of NGWS will be as follows:

- support for voice, multimedia, and data traffic with QoS provisioning,
- support for the best network selection based on user's service needs and network state,
- backbone traffic carried over the Internet.

2.2. NGWS Basic Definitions

To proceed with the design of the analytical model, let us first establish some basic definitions.

Definition 2.2.1. The *set of subsystems* in NGWS, denoted by $\mathbb{S}$, contains the stand-alone subsystems that makeup NGWS.

Definition 2.2.2. $NGWS = (\mathbb{S}, HR)$ is a 2-tuple where the first component is the set of subsystems and the second component is the global home register, denoted by HR.

An instance of $NGWS$ can be

$$NGWS = (\mathbb{S}, HR), \tag{2.1}$$

where $\mathbb{S} = \{wl, pcs, sa, 4g\}$ is the set of subsystems and HR is the global home register for each element $s \in \mathbb{S}$. In the set of subsystems in Equation 2.1, wl, pcs, sa, and $4g$ correspond to WLAN, PCS, Satellite, and 4G Mobile subsystems, respectively. The reader should note that $\mathbb{S}$ can be expanded as needed to include other types of wireless subsystems. Such additions will not affect our admission control scheme.

2.3. NGWS Architecture

Each subsystem in $\mathbb{S}$ has its own cellular infrastructure. We use the term access node for the device that provides connectivity between MTs and the network. Access node corresponds to the *base transceiver station* in PCS subsystem, the *access point* in WLAN, the *transponder* in satellite subsystem, etc. We denote the i^{th} access node of subsystem s as b_i^s . For a given access node b_i^s , we define its cell as the set of locations from which it is possible to communicate with that access node. Thus, each subsystem s splits the service area into cells. If we consider just one subsystem, then the service area is split into cells. Clearly, there is a one-to-one correspondence between the cells and access nodes. The size, shape, and location of the cells depend on the location and power of the access node, and the terrain. Therefore, the cellular layouts of individual

subsystems differ. In NGWS, multiple subsystems are merged under a single system. The service area consists of overlapping cells, so the concept of cellular structure is not enough for the definition of NGWS.

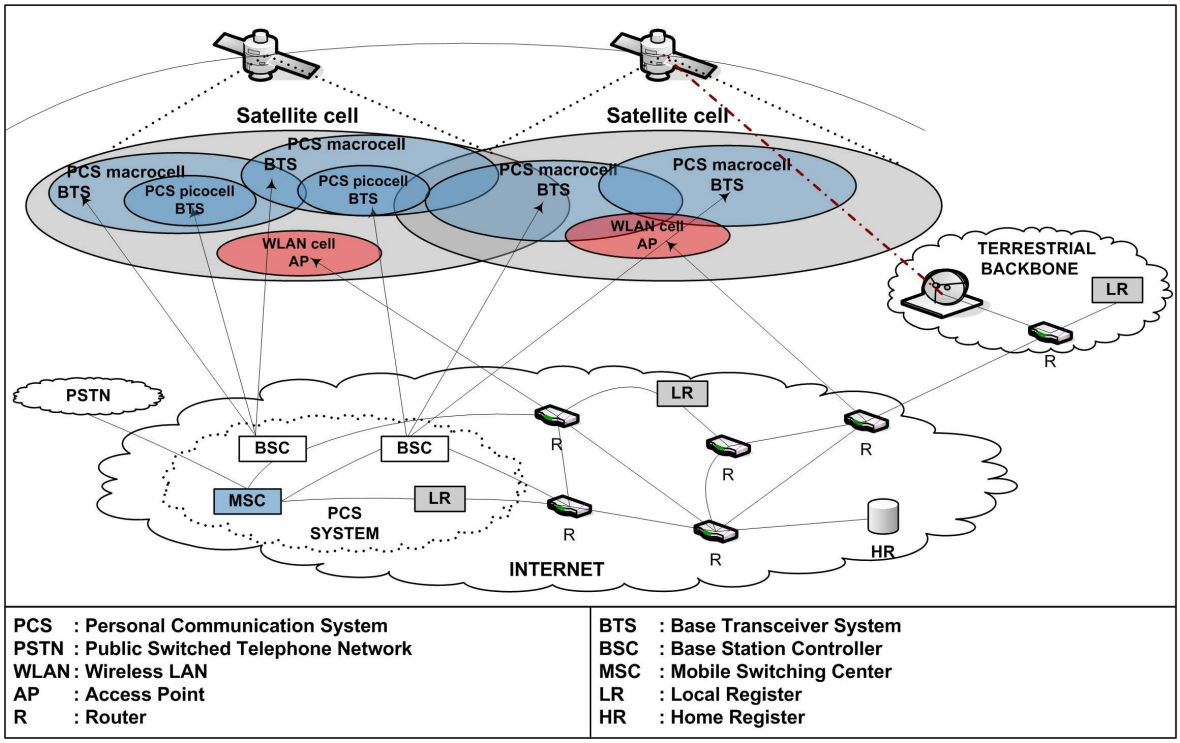


Figure 2.1. An example interconnection of subsystems in NGWS

The basic properties of NGWS listed above, state that NGWS provides various types of services such as voice, video, and multiple types of data. We denote the set of connection classes in NGWS as

$$\mathcal{C} = \{voice, video, low\,bandwidth\,data\,, high\,bandwidth\,data\,, ...\}. \qquad (2.2)$$

Each type of connection class has different requirements ranging from bandwidth to end-to-end latency. For the sake of simplicity, we consider only the bandwidth requirement and assume that reasonable values for the remaining requirements can be achieved if enough bandwidth is provided [21]. We denote the bandwidth requirement of a class k connection as $bw(k)$.

2.3.1. Home Register

In our NGWS architecture, there is a unique *home register* HR that serves all subsystems in mobility and connection management. The home register resides outside all subsystems, as a node in the Internet. The function of HR is to store static and dynamic information about all registered users [21]. In Figure 2.2, HR has a different interface for each type of subsystem and acts as the *Home Agent (HA)* for WLAN networks, as the *Home Location Register (HLR)* for PCS networks, etc.

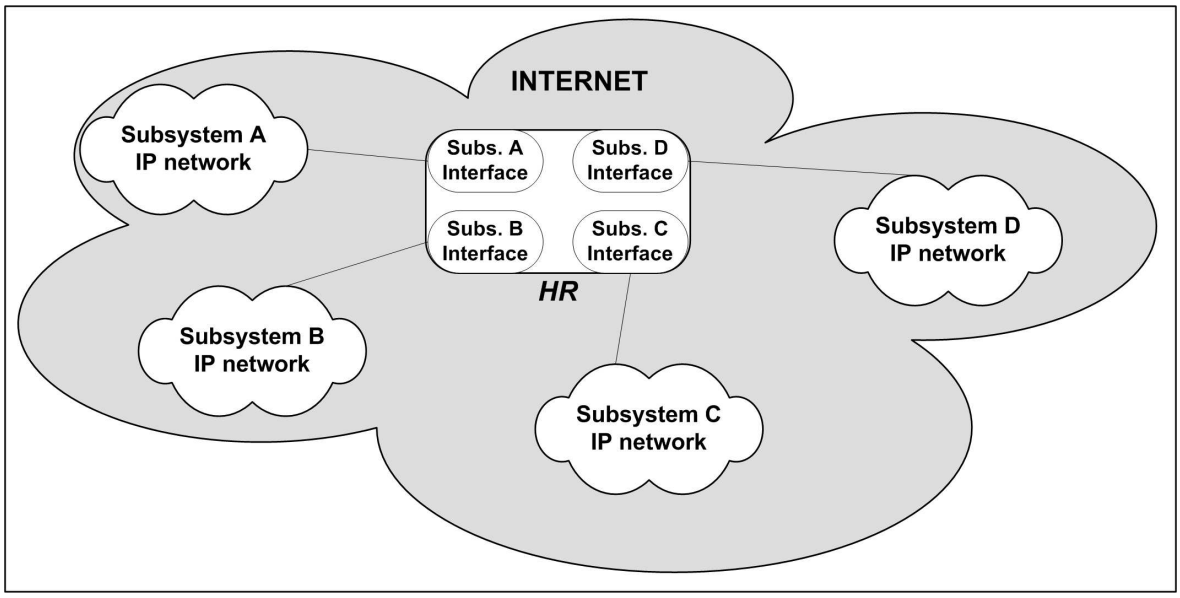


Figure 2.2. The structure of HR

An example configuration for NGWS consisting of satellite, PCS, and WLAN subsystems is depicted in Figure 2.1. The service area is covered by overlapping cells of different subsystems. The coverage area of individual subsystems may be discontinuous, as in the case of WLAN. Thus, the set of subsystems to which a mobile terminal can access at a given moment varies. Each subsystem has its own location register, LR, and the backbone traffic between the subsystems is carried over the Internet. The global HR serves all subsystems.

In current wireless systems, the central location register is queried for the user information with the user's specific identifier as the index. The mobile phone number

in PCS systems, and the home address in Mobile IPv6 are used for addressing the user in the corresponding subsystems. In NGWS, the information about a user can be retrieved from the HR by using the user's *Next Generation User Address (NGUA)* as the index. NGUA is the only address visible to the human users, independent of the subsystem in which the mobile terminal resides. It is the responsibility of HR to translate the NGUA specified by the caller party to the address (e.g., phone number, home IP address) in the relative subsystem.

2.3.2. Location Register

The cells in a subsystem are grouped into *registration areas* for efficient location management. Each registration area is assigned to a local register. We assume that there is a one-to-one correspondence between the local registers and registration areas. Local registers are not new components in our architecture. Instead, we propose utilizing the existing local registers. Each subsystem employs its own location management procedure internally, but LRs are in contact with the global home register [32].

3. NEXT GENERATION CONNECTION ADMISSION CONTROL

In a wireless network, it is the connection admission control scheme that decides whether new connection and handoff requests will be accepted. For new connection requests, NGWS must select the appropriate subsystem for connection establishment. Furthermore, the mobility of a user may require changes in the use of wireless resources, resulting in a handoff attempt for the user. It is the duty of the *Next Generation Connection Admission Control (NGCAC)* scheme to manage the connection requests and handoff attempts in a way that maximizes the network utilization, minimizes the resource outage, and distributes the load between subsystems.

The connection admission control scheme is triggered in three cases:

- Outgoing connection request: When a user initiates a connection request.
- Incoming connection request: When a remote (mobile or fixed) user initiates a connection request destined at the mobile user.
- Handoff request: When a mobile user with an active connection crosses a cell boundary.

3.1. NGCAC Problem Statement

Since MT has access to multiple subsystems simultaneously, NGWS must select one of the subsystems for connection. Among the accessible subsystems, one subsystem that can accommodate the connection request will be selected, subject to connection class and user preferences.

We define the vicinity of b_i^s as

$$\mathbb{V}(b_i^s) = \{b_j^t \mid t \neq s \; and \; b_i^s \cap b_j^t \neq \varnothing\}, \tag{3.1}$$

and call all access nodes in the vicinity as neighbors. Each access node b_i^s periodically transmits its load information l_i^s to all of its neighbors, and also keeps a record of their loads. We denote the recorded value of l_j^t at access node b_i^s as $l_j^{'t}$. Since the load information is exchanged between only a few access nodes in the vicinity, and the information exchange is performed over abundant wired links, this overhead is negligible. $l_j^{'t}$ may not be exactly up-to-date, but since the load in a cell does not fluctuate wildly, $l_j^{'t}$ will be reasonably close to l_j^t. If request rq is accepted, we denote the new load of b_i^s as $\widehat{l_i^s}\,(rq)$. if rq is accepted, we also denote the load of b_j^t based on the recorded value $l_j^{'t}$, as $\widehat{l_j^{'t}}\,(rq)$.

With each connection or handoff request rq, we associate an *ordered list* of accessible access nodes in which ordering criteria is the user's preferences for the connection class of rq. We denote the ordered list of access nodes specified in request rq as $\mathcal{L}_{ac}(rq)$. For outgoing connection setup and handoff requests, MT sends the request to the first access node, b_i^s, in $\mathcal{L}_{ac}(rq)$. However, for incoming connections the initiator (caller) is a remote node that is not aware of the subsystems accessible by MT, availability of the resources in the subsystems, and user preferences for MT. Furthermore, the paging process, which precedes connection establishment, need not be done through the subsystem over which the connection will be established. Therefore, we propose that in the paging reply message, MT specifies $\mathcal{L}_{ac}(rq)$ to be used in the connection admission. Then, the connection establishment is performed over the first access node, b_i^s, in $\mathcal{L}_{ac}(rq)$. Since $\mathcal{L}_{ac}(rq)$ contains the identifiers of a few access nodes, its overhead is negligible.

When b_i^s receives request rq, either directly from MT or from a remote node, it checks if the request can be accommodated. If $\widehat{l_i^s}\,(rq)$ remains below capacity c_i^s, b_i^s accepts the request, establishes the connection, and makes the necessary resource allocations. On the other hand, if $\widehat{l_i^s}\,(rq)$ exceeds c_i^s, b_i^s contacts, *on behalf of MT,* all access nodes b_j^t in $\mathcal{L}_{ac}(rq)$ for which $\widehat{l_j^{'t}}\,(rq) \leq c_j^t$. If there exists an access node b_j^t in $\mathcal{L}_{ac}(rq)$ that can accommodate request rq, the request will be transferred to b_j^t, and the connection will be established over that access node. On the other hand, if $\widehat{l_j^{'t}}\,(rq) > c_j^t$ for all b_j^t in $\mathcal{L}_{ac}(rq)$, request rq will be rejected. In this case, MT may either call off

the request or revise the connection class (connection requirements) and resubmit it as a new request. The algorithm for the proposed scheme is presented as a flowchart in Figure 3.1.

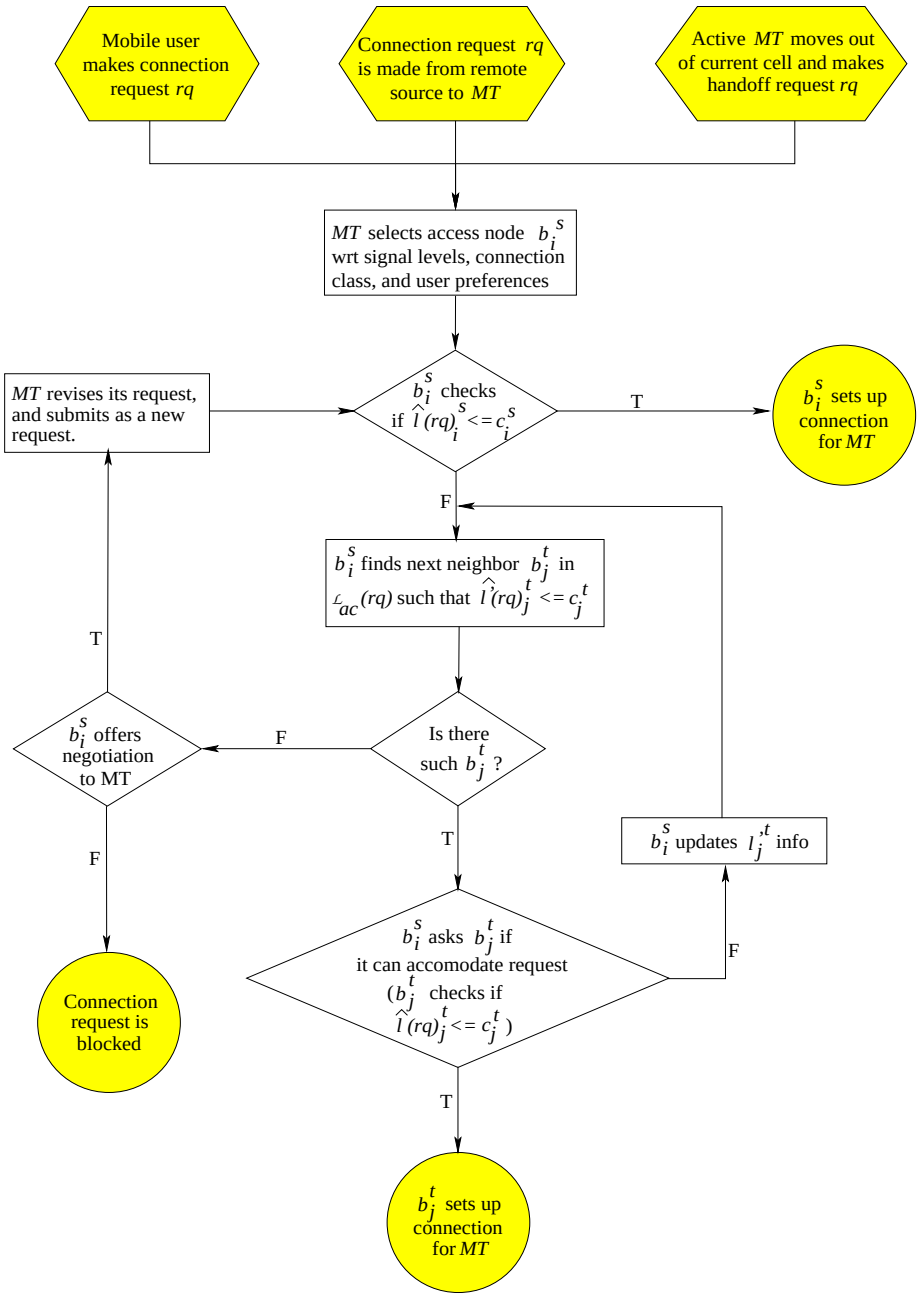


Figure 3.1. The algorithm of the connection admission control scheme

3.2. Subsystem Selection

The overlaid structure of NGWS implies that MT has access to multiple subsystems simultaneously. MT knows the set of subsystems it can access by scanning the pilot signal, from the access nodes. The selection of the subsystem s for a connection establishment depends on several factors:

Subsystem accessibility: The pilot signal of subsystem s must be strong enough for communication. Subsystem s is accessible if and only if MT is in a cell belonging to one of the access nodes of subsystem s .

Resource availability: The load of access node b_i^s , l_i^s , must not exceed the capacity c_i^s of b_i^s , if the request is admitted.

Service class and user preferences: The user is able to indicate which subsystems s/he prefers for each connection class. We denote the probability that MT prefers subsystem s for a connection of class k as $p(k, s)$ where $k \in \mathcal{C}$.

An example scenario is depicted in Figure 3.2 with three subsystems, sa, pcs, and wl. Although the service area is covered by three subsystems, there are *holes* where

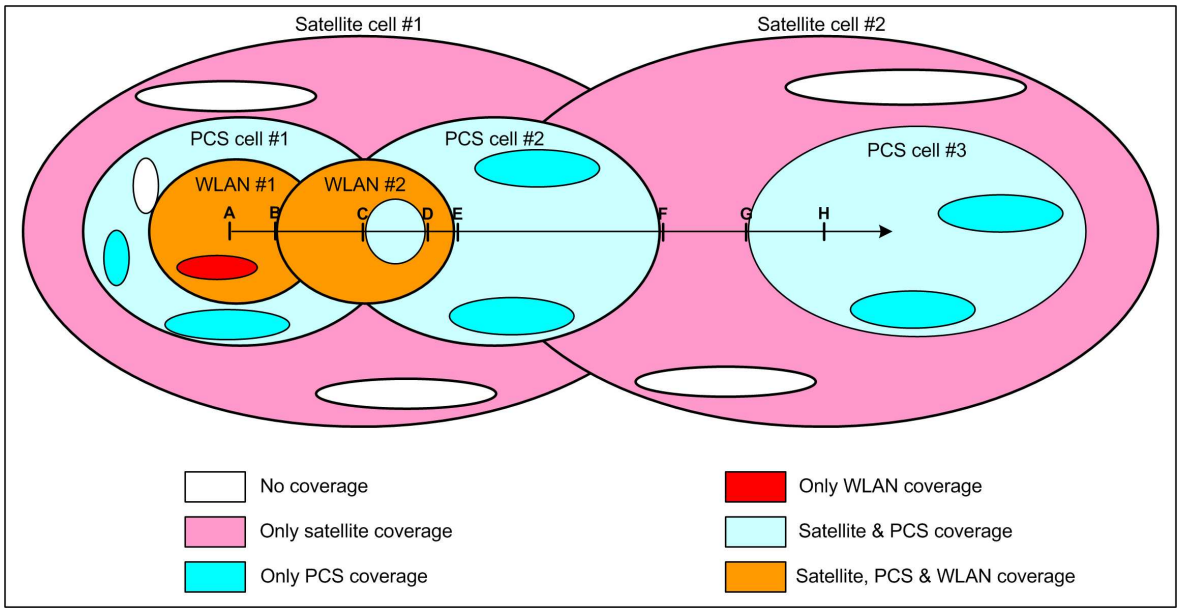


Figure 3.2. Connection admission scenario

no signal is received from one or more subsystems. Let's assume MT's subsystem

preference decreases in the order *wl*, *pcs*, and *sa*. *MT* initiates a connection at point *A* where it has access to all subsystems, and selects *wl* according to user preferences. As it moves on its trajectory, it encounters an intra-subsystem handoff between two WLAN cells at point *B*. At point *C*, *MT* enters a hole in the WLAN cell where access to *wl* is lost. Therefore, *MT* encounters an inter-subsystem handoff at point *C*, from *wl* to *pcs*, preferring *pcs* over *sa*. *MT* gets out of the hole at point *D*, but it will keep communicating over *pcs* until it reaches point *F*. At point *F*, *MT* leaves the PCS cell, so it will encounter another inter-subsystem handoff from *pcs* to *sa*. Finally at point *G*, *MT* successfully completes its connection.

4. ANALYTICAL MODEL OF NGCAC

In the analysis of wireless systems, the service area is typically split into cells since the partitioning criteria is the access node. However, in the case of NGWS, cellular granularity is too coarse to define a partition since there are multiple subsystems serving the same service area. Therefore, in our model, the service area is partitioned into smaller regions called physical areas.

4.1. System Definition

Let $\mathcal{B}$ be the set of all access nodes in all subsystems in the service area Ω. We denote the set of all subsets of $\mathcal{B}$ as $\mathcal{A} = 2^{\mathcal{B}}$.

Definition 4.1.1. An element of $\mathcal{A}$ is said to be an *area*.

Example 4.1.2. Let $\mathbb{S} = \{wl, pcs\}$ where wl represents WLAN and pcs represents PCS as subsystems. Let wl and pcs both have three access nodes. Hence, $\mathcal{B} = \{b_0^{wl}, b_1^{wl}, b_2^{wl}, b_0^{pcs}, b_1^{pcs}, b_2^{pcs}\}$. $\mathcal{A}$ has 2^6 elements which are called areas. These structures do not say anything about the topology of the service area Ω. So it is necessary to have a structure which has information about the topology.

We define $\widetilde{b} : \Omega \to \mathcal{A}$ such that $\widetilde{b}(x)$ is the set of all access nodes reachable from location x. This mapping introduces an equivalence relation $\sim$ on Ω such that $x_1 \sim x_2$ if $\widetilde{b}(x_1) = \widetilde{b}(x_2)$.

Theorem 4.1.3. *Relation introduced on Ω by $\widetilde{b}$ is an equivalence relation.*

Proof. We have to show $\sim$ is reflexive, symmetric, and transitive.
Reflexivity: Let $x \in \Omega$ be arbitrary. Which implies, $\widetilde{b}(x) = \widetilde{b}(x)$ since $\widetilde{b}$ is a well defined function. So we get, $x \sim x$. Therefore, $\sim$ is reflexive.
Symmetry: Suppose $x_1 \sim x_2$, which implies $\widetilde{b}(x_1) = \widetilde{b}(x_2)$. So $\widetilde{b}(x_2) = \widetilde{b}(x_1)$. From this, we get $x_2 \sim x_1$. Therefore, $\sim$ is symmetric.
Transitivity: Suppose $x_1 \sim x_2$ and $x_2 \sim x_3$, which implies $\widetilde{b}(x_1) = \widetilde{b}(x_2)$ and $\widetilde{b}(x_2) =$

$\widetilde{b}\,(x_3)$. So $\widetilde{b}\,(x_1) = \widetilde{b}\,(x_3)$. From this, we get $x_1 \sim x_3$. Therefore $\sim$ is transitive. Hence, $\sim$ is an equivalence relation. □

Definition 4.1.4. For each area $a \in \mathcal{A}$, physical area $\bar{a}$ is defined as

$$\bar{a} = \{x \in \Omega \mid \widetilde{b}\,(x) = a\}.$$

Clearly for each $a \in \mathcal{A}$, we have $\bar{a} \subseteq \Omega$. Physical areas are the equivalence classes introduced by $\sim$. Some of the equivalence classes may be empty. If $\bar{a} = \varnothing$, then it means there is no physical location in Ω such that all access nodes in a are reachable. So, $\sim$ partitions Ω according to reachable access nodes.

Proposition 4.1.5. *If* $\overline{a_1} = \overline{a_2} \neq \varnothing$, *then* $a_1 = a_2$

Proof. Assume , $\overline{a_1} = \overline{a_2} \neq \varnothing$.
Let $x \in \overline{a_1}$ be arbitrary ($\overline{a_1} \neq \varnothing$), which implies $x \in \overline{a_2}$ and $x \in \overline{a_1}$ (since $\overline{a_1} = \overline{a_2}$). So $\widetilde{b}\,(x) = a_2$ and $\widetilde{b}\,(x) = a_1$. From this we get $s \in a_1$ if and only if s reachable from x, and $s \in a_2$ if and only if s reachable from x. This implies $a_2 = a_1$ (since $s \in a_2$ if and only if $s \in a_1$). □

We find it more convenient to work with areas rather than with physical areas since it reduces the complexity of the description of the model. We call the areas for which corresponding physical areas are non-empty as *effective areas*. For obvious reasons, we are interested only in the effective areas. The upper bound on the number of effective areas is given by $\sum_{n=0}^{K} \binom{|\mathcal{B}|}{n}$ where K is the maximum number of access nodes reachable from one location.[1] The partitioning of the service area and the relationship between areas and cells according to the coverage in Figure 3.2 is shown in Figure 4.1. The number of area crossings on the trajectory of MT provides a reasonable explanation for not initiating a handoff procedure to the most preferred access node at every area boundary.

[1] From this point on, we will use the term area in the sense of effective areas.

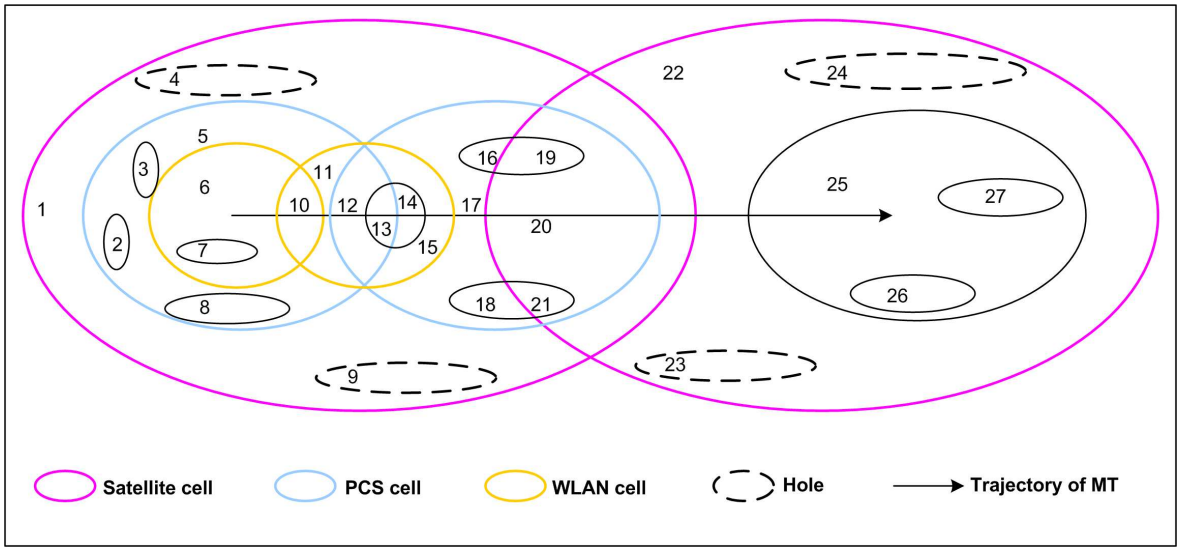


Figure 4.1. An example partitioning of the service area

For each area a_i , we have the following:

- $n_{a_i}(t)$: Number of users in area a_i at time t.
- $V^k_{a_j,a_i}(t)$: Migration rate of a class k connection from area a_j to a_i at time t. Hence, the number of class k migrating from a_j to a_i during the time period τ is equal to

$$\int_t^{t+\tau} n_{a_j} \cdot V^k_{a_j,a_i}(t)\, dt\,. \tag{4.1}$$

- $r^k_{a_i}(t)$: Connection generation rate of class k connections in area a_i at time t.
- $f(a_i,k,t)$: Connection profile of class k connections in area a_i at time t, i.e., the distribution of class k such as 0.60 for voice, 0.10 for multimedia, 0.30 for data.
- $p(a_i,k,s,t)$: Probability that a user in area a_i with a class k connection prefers subsystem SS^s at time t.

We denote the number of active connections of class k communication with access node b in area a as $x^k_a(b)$.

Definition 4.1.6. The state of the system is any particular distribution of values taken by the variables $x^k_a(b)$.Thus, the state of the system is the tuple

$$g = \begin{bmatrix} x_1^1(1), & x_1^1(2), & \cdots, & x_1^1(|\mathcal{B}|) \\ x_1^2(1), & x_1^2(2), & \cdots, & x_1^2(|\mathcal{B}|) \\ \cdots & & & \\ x_1^{|\mathcal{C}|}(1), & x_1^{|\mathcal{C}|}(2), & \cdots, & x_1^{|\mathcal{C}|}(|\mathcal{B}|) \\ x_2^1(1), & x_2^1(2), & \cdots, & x_2^1(|\mathcal{B}|) \\ \cdots & & & \\ x_2^{|\mathcal{C}|}(1), & x_2^{|\mathcal{C}|}(2), & \cdots, & x_2^{|\mathcal{C}|}(|\mathcal{B}|) \\ \vdots & & & \\ \vdots & & & \\ x_{|\mathcal{A}|}^1(1), & x_{|\mathcal{A}|}^1(2), & \cdots, & x_{|\mathcal{A}|}^1(|\mathcal{B}|) \\ \cdots & & & \\ x_{|\mathcal{A}|}^{|\mathcal{C}|}(1), & x_{|\mathcal{A}|}^{|\mathcal{C}|}(2), & \cdots, & x_{|\mathcal{A}|}^{|\mathcal{C}|}(|\mathcal{B}|) \end{bmatrix}. \tag{4.2}$$

Example 4.1.7. As example let's consider the following simple topology. In this

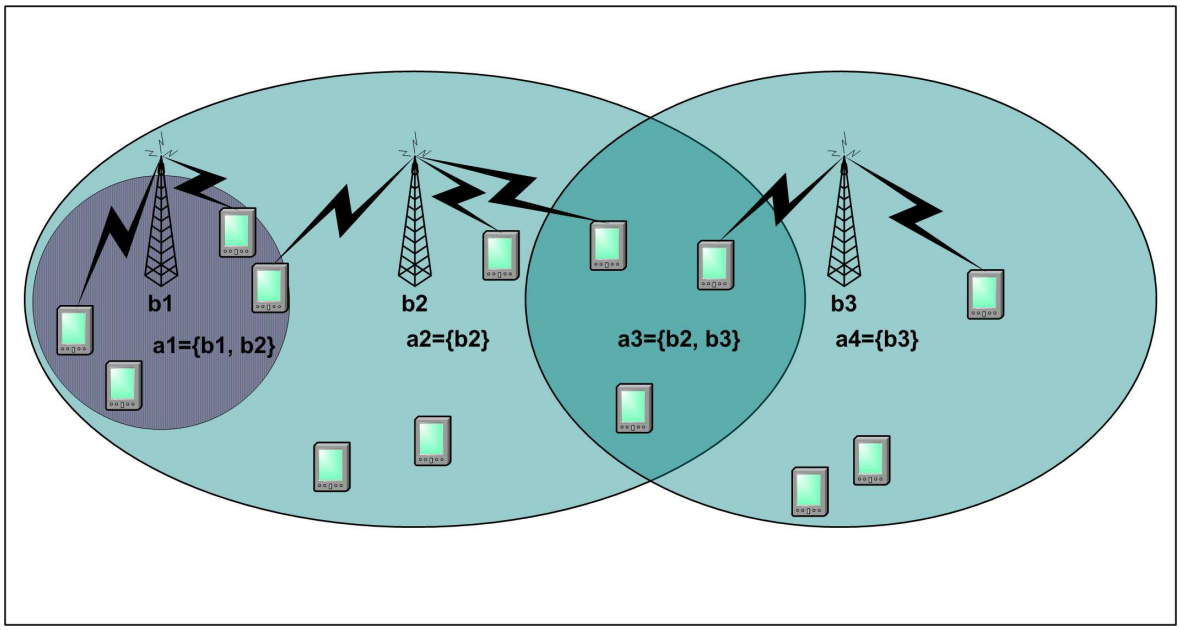


Figure 4.2. An example connection in areas

situation, there are three access nodes and two subsystems. Access nodes $b2$ and $b3$ belong to the PCS subsystem while $b1$ belongs to the WLAN subsystem. We assume

there is one connection class. So, the system definition is summarized as

$$\begin{aligned} \mathbb{S} &= \{wl, pcs\}, \\ \mathcal{C} &= \{c1\}, \\ \mathcal{B} &= \{b1, b2, b3\}, \\ \mathcal{A} &= \{\{b1, b2\}, \{b2\}, \{b2, b3\}, \{b3\}\} = \{a1, a2, a3, a4\}. \end{aligned} \tag{4.3}$$

We consider only the effective areas for $\mathcal{A}$. In $a1$, there are three active connections. Two of these connections are through $b1$ and the other active connection is through $b2$. This is possible since in $a1$, both $b1$ and $b2$ are reachable access nodes.

The following matrix consisting of x_a^k values denotes the state of the system shown in Figure 4.2.

$$g = \begin{bmatrix} x_{a1}(b1) & x_{a1}(b2) & x_{a1}(b3) \\ x_{a2}(b1) & x_{a2}(b2) & x_{a2}(b3) \\ x_{a3}(b1) & x_{a3}(b2) & x_{a3}(b3) \\ x_{a4}(b1) & x_{a4}(b2) & x_{a4}(b3) \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.4}$$

Definition 4.1.8. The state of the system can also be represented as a mapping

$$g : \mathcal{X} \rightarrow \mathbb{N} \cup \{0\} \tag{4.5}$$

where $\mathcal{X} = \left\{x_a^k(b)\right\}$. Thus, a state is any particular distribution of values taken by individual $x_a^k(b)$. We denote the set of all states as $\mathcal{G}$.

Although each variable $x_a^k(b_i^s)$ can assume values up to the capacity of corresponding access node b_i^s, we have the constraint

$$\sum_{a \in \mathcal{A}} \sum_{k \in \mathcal{C}} bw(k) \cdot x_a^k \leq c_i^s, \tag{4.6}$$

since the same wireless resources are shared by all areas covered by the same cell.

Therefore, many states in the state space are never visited.

4.2. Elementary Events

There are various elementary events that cause the system to switch from one state to another. Keeping in mind that a state of the system is a mapping as given by Equation 4.5, each elementary event causes a change in the mapping at one or two points. For example, let's assume that the current state of the system is defined by the mapping g_i, and access node b accepts a new connection request of class k from area a. This new connection causes $g_i\left(x_a^k(b)\right)$ to be incremented by one, defining a new mapping g_j. In Figure 4.3, an example in which $x_5^1(2)$ is incremented due to a new connection. Two mappings, g_i and g_j are the same except at the point marked with an arrow.

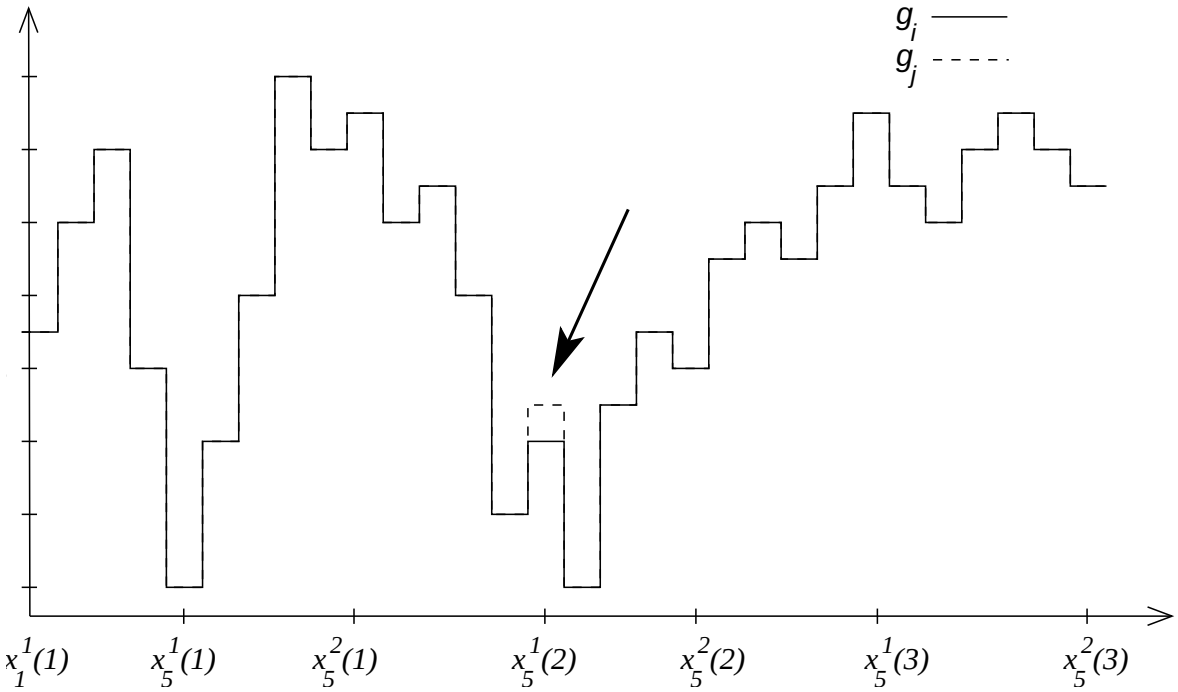


Figure 4.3. State transition due to a new connection of class 1 from area a_5 over access node 2

Elementary events, which cause the system to change its state, can be categorized into three main groups:

New Connection Events: A MT initiates or receives a new connection.

Migration Events: A MT with an active connection moves within the service area.

Hangup Event: A MT with an active connection terminates the connection.

4.2.1. New Connection Events

Example 4.2.1. While considering the effect of a new connection event on the system, we determine the state before the event. The state of the system is depicted as a function in Figure 4.4. For example, let's assume a new connection event occurs in

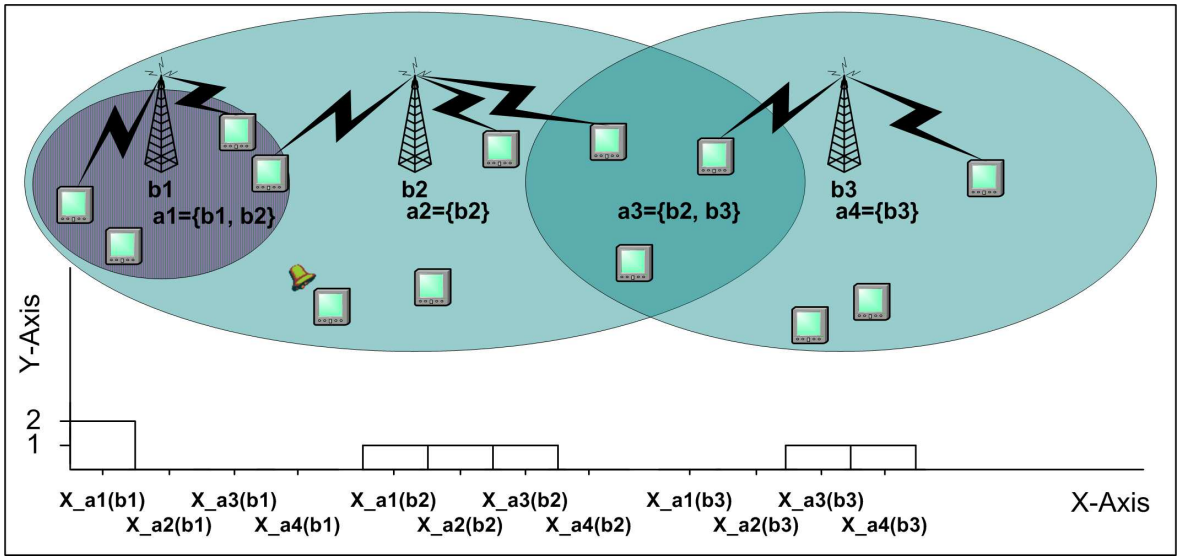


Figure 4.4. State before the new connection event

area $a2$ through access node $b2$. This event results in an increase in the number of active connections in $a2$ through $b2$, which means an increase in the value of $x^1_{a2}(b2)$. In Figure 4.4, MT marked with a ringing bell in area a_2 initiates a new connection and causes a new connection event. New connection event caused by MT changes the state of the system by increasing the value $x^1_{a2}(b2)$ by 1. The increased value is shown by the arrow in Figure 4.5.

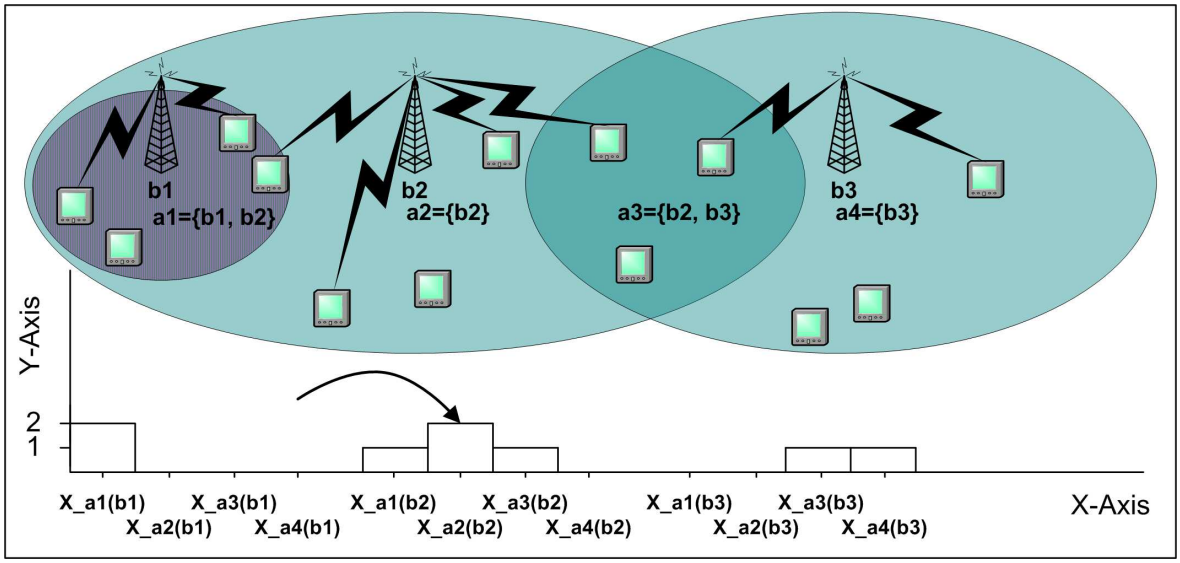


Figure 4.5. State after the new connection event

4.2.1.1. Outgoing New Connection. MT initiates a new connection request of class k in area a, and sends request rq directly to b_i^s, the first access node in $\mathcal{L}_{ac}(rq)$.

- Direct Outgoing New Connection:
  In this case, b_i^s accepts rq since $\widehat{l}_i^s(rq)$ does not exceed c_i^s. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_i^s)$:

$$g_j\left(x_a^k(b_i^s)\right) = g_i\left(x_a^k(b_i^s)\right) + 1\,. \tag{4.7}$$

- Indirect Outgoing New Connection:
  In this case, b_i^s cannot accommodate rq since $\widehat{l}_i^s(rq)$ exceeds c_i^s. b_i^s checks $\mathcal{L}_{ac}(rq)$ for the first access node b_j^t such that $\widehat{l'}_j^t(rq) \leq c_j^t$. The reader should note that, subsystems s and t may be equal or not. Thus, rq is *redirected* to b_j^t for connection setup. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_j^t)$:

$$g_j\left(x_a^k(b_j^t)\right) = g_i\left(x_a^k(b_j^t)\right) + 1\,. \tag{4.8}$$

4.2.1.2. Incoming New Connection. MT receives a paging request from a remote source for a class k connection while in area a. MT specifies $\mathcal{L}_{ac}(rq)$ in the paging reply, and the source initiates the connection setup procedure over b_i^s, the first access node in $\mathcal{L}_{ac}(rq)$.

- Direct Incoming New Connection:
  In this case, b_i^s accepts rq since $\widehat{l}_i^s(rq)$ does not exceed c_i^s. The system will switch from state g_i to g_j such that two states differ only at $x_a^k(b_i^s)$:

$$g_j\left(x_a^k(b_i^s)\right) = g_i\left(x_a^k(b_i^s)\right) + 1\,. \tag{4.9}$$

- Indirect Incoming New Connection:
  In this case, b_i^s cannot accommodate rq since $\widehat{l}_i^s(rq)$ exceeds c_i^s. b_i^s checks $\mathcal{L}_{ac}(rq)$ for the first access node b_j^t such that $\widehat{l'}_j^t(rq) \leq c_j^t$. The reader should note that, subsystems s and t may be equal or not. Thus, rq is *redirected* to b_j^t for connection setup. The system will switch from state g_i to g_j such that two states differ only at $x_a^k\left(b_j^t\right)$:

$$g_j\left(x_a^k\left(b_j^t\right)\right) = g_i\left(x_a^k\left(b_j^t\right)\right) + 1\,. \tag{4.10}$$

4.2.2. Migration Events

4.2.2.1. Intra-cell Movement. In this case, MT with an active connection of class k over access node b_i^s, in area a_u moves to area a_v, which is covered by the same access node. Therefore, handoff will not occur, but the system will switch from state g_i to g_j such that two states differ only at $x_{a_u}^k(b_i^s)$ and $x_{a_v}^k(b_i^s)$:

$$g_j\left(x_{a_u}^k(b_i^s)\right) = g_i\left(x_{a_u}^k(b_i^s)\right) - 1 \tag{4.11}$$

$$g_j\left(x_{a_v}^k(b_i^s)\right) = g_i\left(x_{a_v}^k(b_i^s)\right) + 1\,. \tag{4.12}$$

Example 4.2.2. In Figure 4.6, MT in area $a2$ with an arrow moves in the direction shown. MT communicates through access node $b2$. After changing the area, MT does not change the access node it connects through.

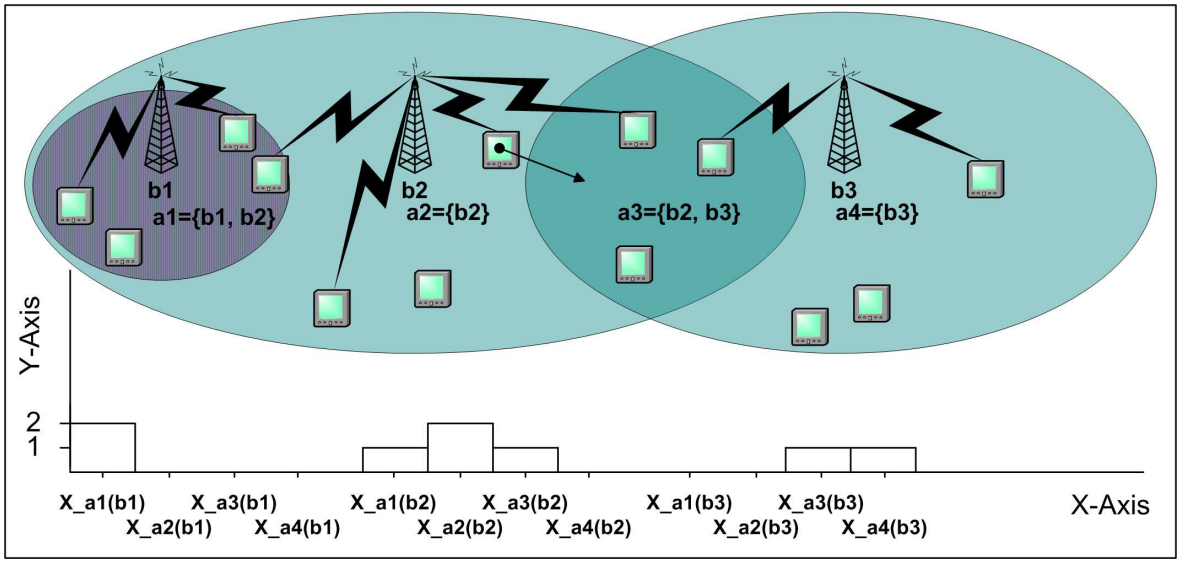


Figure 4.6. State before the intra-cell movement

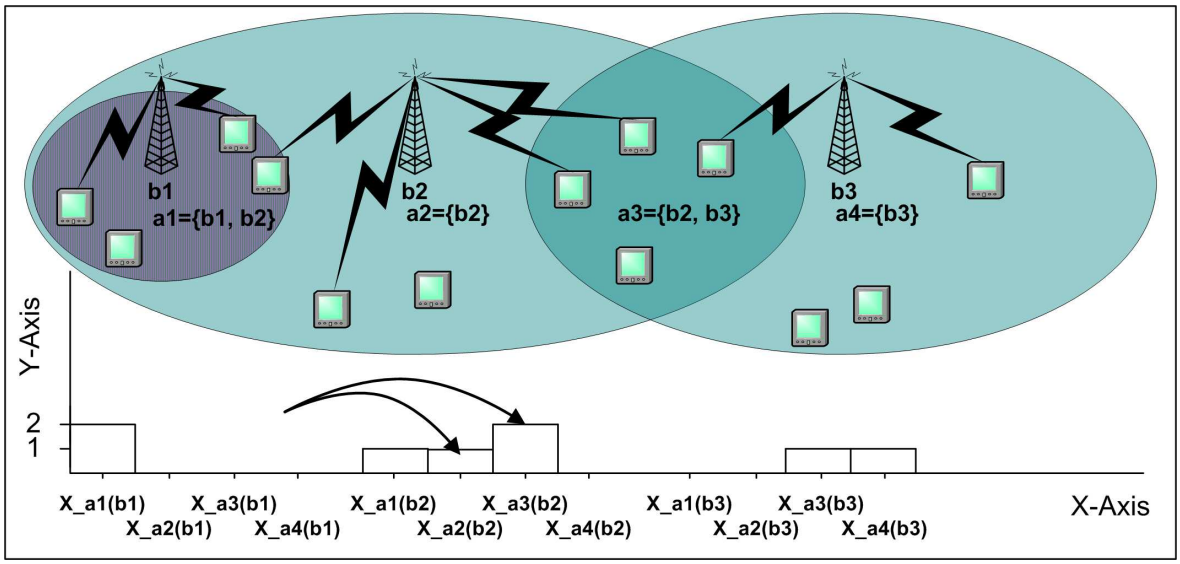


Figure 4.7. State after the intra-cell movement

System state changes as shown in Figure 4.7. After the movement, MT is in area $a3$ with an active connection through $b2$. The value of $x^1_{a2}(b2)$ is decreased by one and the value of $x^1_{a3}(b2)$ is increased by one. The state change is shown in Figure 4.7. If the capacity of $b2$ is four, then $b2$ does not accept any new connection and handoff requests.

4.2.2.2. Intra-subsystem Handoff. In this case, MT with an active connection of class k over access node b^s_i, in area a_u moves to area a_v, which is covered by b^s_j but not by b^s_i. Therefore, MT encounters a handoff from b^s_i to b^s_j, within subsystem s. The system will switch from state g_i to g_j such that two states differ only at $x^k_{a_u}\left(b^s_i\right)$ and $x^k_{a_v}\left(b^s_j\right)$:

$$g_j\left(x^k_{a_u}\left(b^s_i\right)\right) = g_i\left(x^k_{a_u}\left(b^s_i\right)\right) - 1 \tag{4.13}$$

$$g_j\left(x^k_{a_v}\left(b^s_j\right)\right) = g_i\left(x^k_{a_v}\left(b^s_j\right)\right) + 1\,. \tag{4.14}$$

Example 4.2.3. In Figure 4.8, MT in area $a3$ with an arrow on it moves in the direction shown. MT communicates through the access node $b2$. After changing the area MT changes the access node it connects through, since $b2$ coverage is not enough to communicate. MT initiates a handoff request to $b3$.

The state of the system changes as shown in Figures 4.8 and 4.9. After the movement, MT is in area $a4$ with an active connection through $b3$. The value of $x^1_{a3}(b2)$ is decreased by one and the value of $x^1_{a4}(b3)$ is increased by one. The state change is shown in Figure 4.9. When MT moves into area $a4 = \{b3\}$, the only reachable access node is $b3$. Since access node $b3$ is not fully loaded, it grants the handoff attempt. After handoff event, we have two MTs that are connected through $b3$ in area $a4 = \{b3\}$ and one MT that is connected through $b3$ in area $a3 = \{b2, b3\}$.

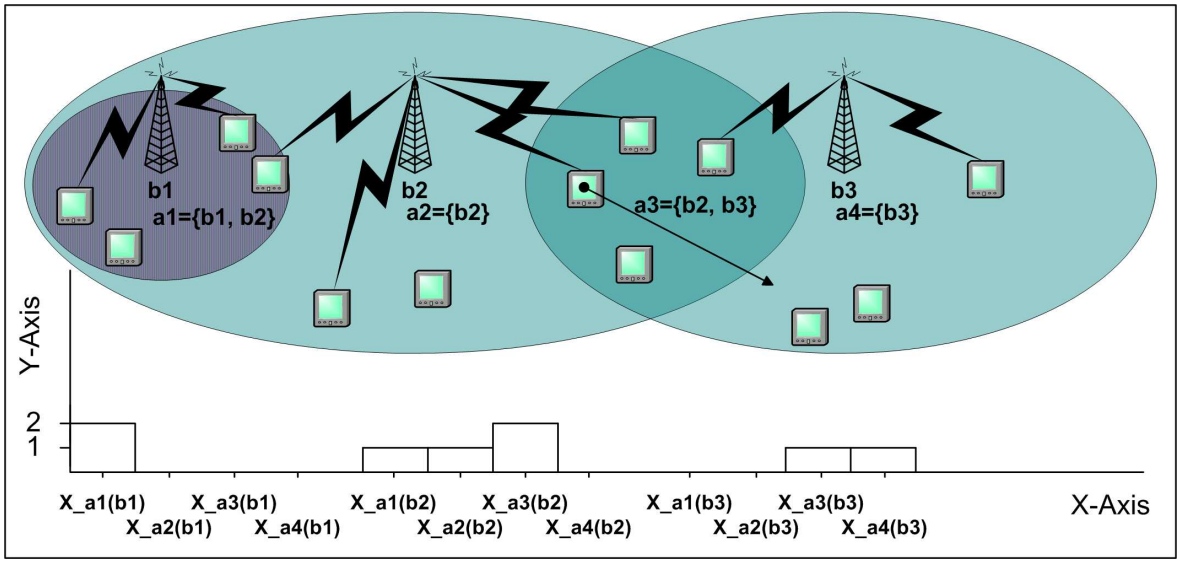


Figure 4.8. State before the intra-subsystem handoff

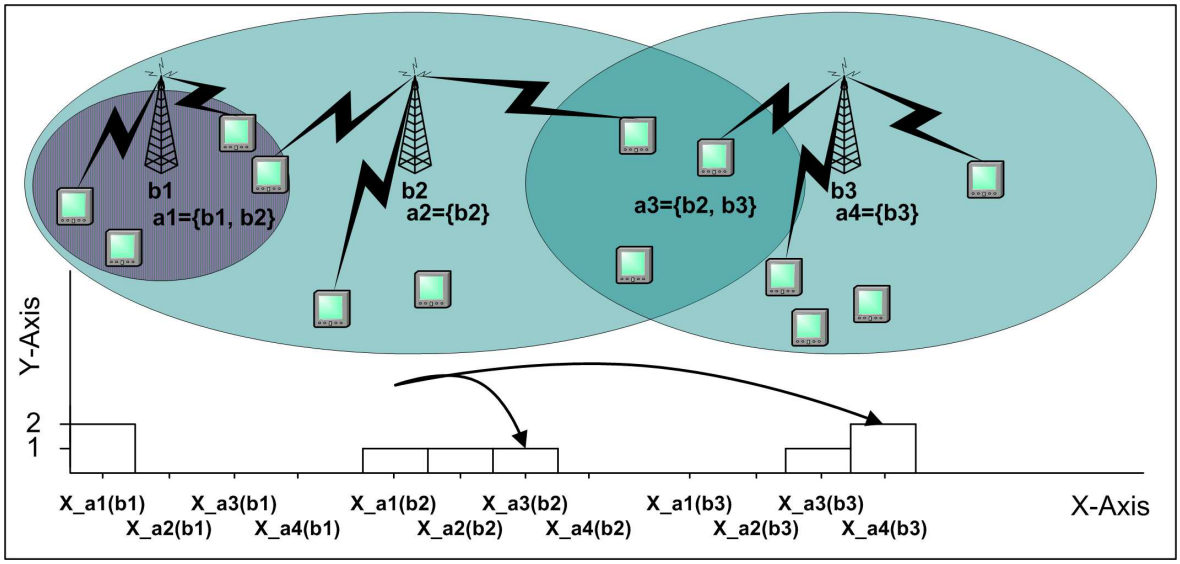


Figure 4.9. State after the intra-subsystem handoff

4.2.2.3. Inter-subsystem Handoff. In this case, MT with an active connection of class k over access node b_z^w, in area a_u moves to area a_v, which is not covered by any access node of subsystem w. To continue uninterrupted service, MT sends a handoff request directly to b_i^s, the first access node in $\mathcal{L}_{ac}(rq)$.

Example 4.2.4. In Figure 4.10, MT in area $a1$ with an arrow moves in the direction shown. MT communicates through the access node $b1$. After changing the area, MT changes the access node it connects through, since $b1$ coverage is not enough to communicate. MT initiates a handoff request to $b2$. The system state changes as

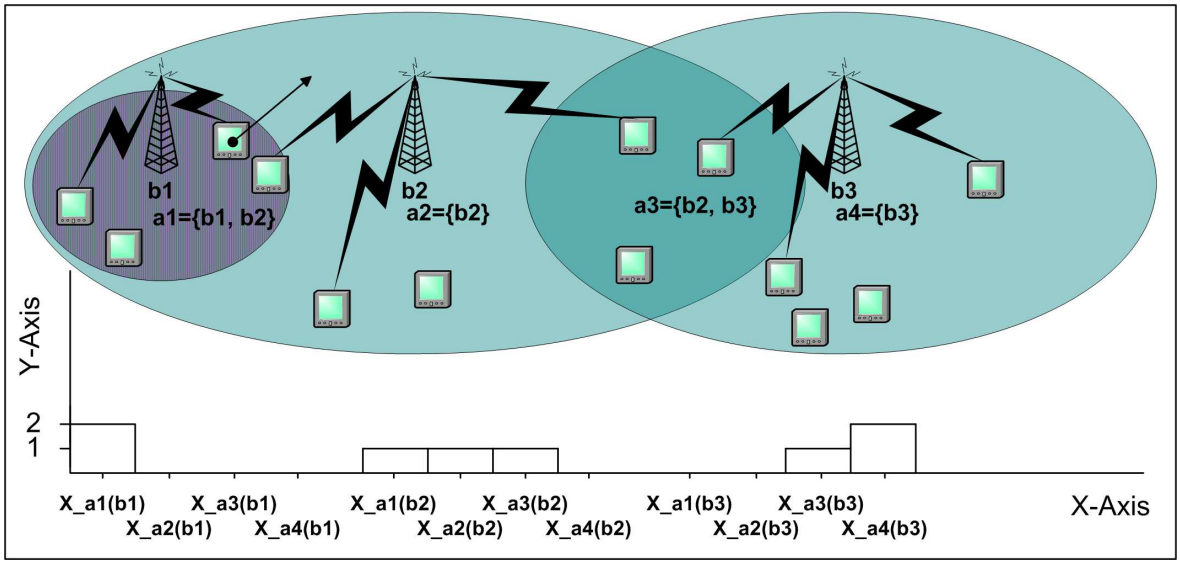


Figure 4.10. State before the inter-subsystem handoff

shown in Figure 4.11. After the movement, MT is in area $a2$ with an active connection through $b2$. The value of $x_{a1}^1(b1)$ is decreased by one and the value of $x_{a2}^1(b2)$ is increased by one.

- Direct Inter-subsystem Handoff:
  In this case, b_i^s accepts rq since $\widehat{l_i^s}(rq)$ does not exceed c_i^s. The system will switch from state g_i to g_j such that two states differ only at $x_{a_u}^k(b_z^w)$ and $x_{a_v}^k(b_i^s)$:

$$g_j\left(x_{a_u}^k(b_z^w)\right) = g_i\left(x_{a_u}^k(b_z^w)\right) - 1 \tag{4.15}$$

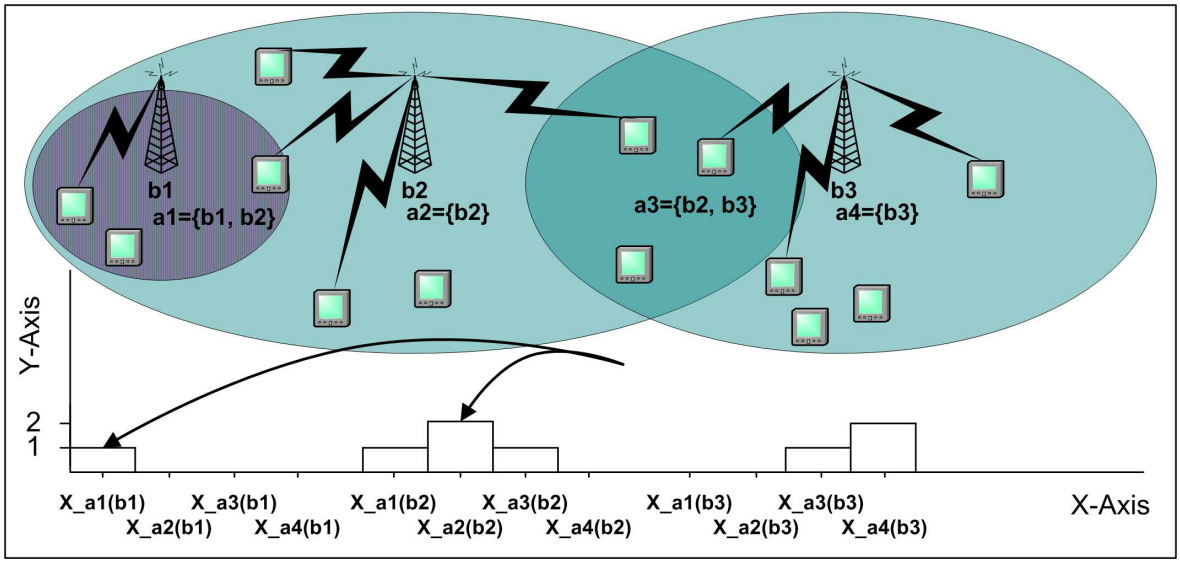


Figure 4.11. State after the inter-subsystem handoff

$$g_j\left(x^k_{a_v}\left(b^s_i\right)\right) = g_i\left(x^k_{a_v}\left(b^s_i\right)\right) + 1\,. \tag{4.16}$$

- Indirect Inter-subsystem Handoff:

  In this case, b^s_i cannot accommodate rq since $\widehat{l}^s_i\left(rq\right)$ exceeds c^s_i. b^s_i checks $\mathcal{L}_{ac}\left(rq\right)$ for the first access node b^t_j such that $\widehat{l'}^t_j\left(rq\right) \leq c^t_j$. The reader should note that, subsystems s and t may be equal or not. Thus, rq is *redirected* to b^t_j for handoff. The system will switch from state g_i to g_j such that two states differ only at $x^k_{a_u}\left(b^w_z\right)$ and $x^k_{a_v}\left(b^t_j\right)$:

$$g_j\left(x^k_{a_u}\left(b^w_z\right)\right) = g_i\left(x^k_{a_u}\left(b^w_z\right)\right) - 1 \tag{4.17}$$

$$g_j\left(x^k_{a_v}\left(b^t_j\right)\right) = g_i\left(x^k_{a_v}\left(b^t_j\right)\right) + 1\,. \tag{4.18}$$

4.2.3. Hangup Event

In this case, MT with an active connection of class k in area a over access node b^s_i terminates the connection voluntarily. The system will switch from state g_i to g_j

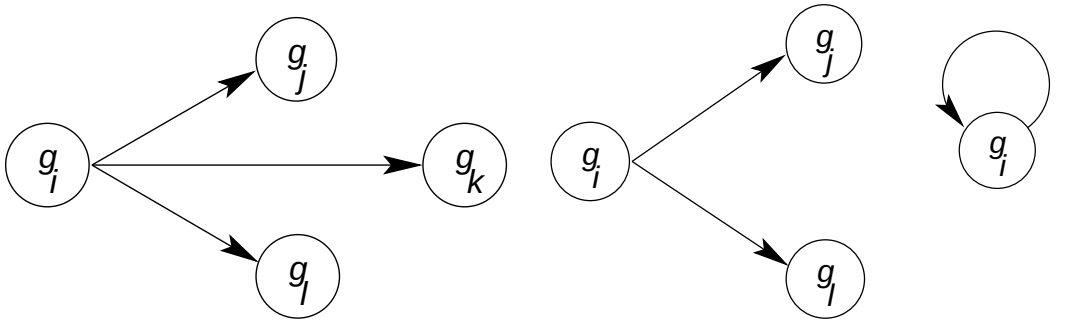
such that two states differ only at $x_a^k(b_i^s)$:

$$g_j\left(x_a^k(b_i^s)\right) = g_i\left(x_a^k(b_i^s)\right) - 1\,. \tag{4.19}$$

4.3. Transition Graphs

We introduce *transition graphs* to explain how the next state is found if the present state is given. There is an individual graph $\Gamma(e,s)$ for every elementary event e and every subsystem s. All transition graphs have the same vertex set $\mathcal{G}$, the set of all states. The set of arcs, $\mathcal{V}$, depends on e and s. We define the set of graphs associated with event e as $\Gamma_e : \mathcal{G}, \mathcal{V}, e, \mathcal{S}$.

We explain the transition graphs with an example in which subsystems s_1, s_2, and s_3 are accessible in area a. The received signal levels from the access nodes of these subsystems will be different at every point in a. We denote the probability that the received signal from the access node of subsystem s_n is the strongest for any randomly selected user in area a as q_n. We examine the transition graph for the *outgoing new connection* event. Depending on the state, we have the $\{s_1, s_2, s_3\}$, $\{s_1, s_2\}$, $\{s_1, s_3\}$, $\{s_2, s_3\}$, $\{s_1\}$, $\{s_2\}$, $\{s_3\}$, and $\varnothing$ cases where $\{x, y, z\}$ represents the case that corresponding cells of the subsystems x, y, and z have enough resources, and $\varnothing$ represents the case that none of the subsystems has enough resources. The outgoing arcs in the transition diagrams that correspond to the $\{s_1, s_2, s_3\}$, $\{s_1, s_3\}$, and $\varnothing$ cases are depicted in Figure 4.12(a-c), respectively. In Figure 4.12(a) for the $\{s_1, s_2, s_3\}$ case, the flow out of g_i due to outgoing new connections is split into three according to the ratios q_1, q_2, and q_3 toward the states g_j, g_k, and g_l. However, in Figure 4.12(b) for the $\{s_1, s_3\}$ case, the flow out of g_i is split into two according to the ratios $\frac{q_1}{q_1+q_3}$ and $\frac{q_3}{q_1+q_3}$ toward the states g_j and g_l.



(a) s_1,s_2,s_3 case (b) s_1,s_3 case (c) $\varnothing$ case

Figure 4.12. Outgoing arcs for state g_i in the transition graph

4.4. Transition Probabilities

Transitions between states occur when MT initiates connections, moves in the service area, and hangups. Due to the complexity of the NGCAC scheme, we examine each case separately.

4.4.1. New Connection Event Probabilities

4.4.1.1. Direct Outgoing NC Probability. The probability that a direct outgoing new connection attempt of class k occurs for access node b is

$$P_{no}(b,k,\Delta t) = \Delta t \cdot \sum_{a \in \mathcal{A}(b)} \{r_a(t) \cdot f(a,k,t) \cdot n_a(t) \cdot p(a,k,s(b),t)\} + o(\Delta t) \quad (4.20)$$

where $\mathcal{A}(b)$ is the set of areas constituting the cell of access node b, $r_a(t)$ is the connection generation rate of all connection classes in area a at time t, and $s(b)$ is the subsystem to which access node b belongs.

4.4.1.2. Indirect Outgoing NC Probability. The probability that an indirect outgoing new connection attempt of class k occurs for access node b is

$$\widetilde{P}_{no}(b,k,\Delta t) = \Delta t \cdot \sum_{a \in \mathcal{A}(b)} \{r_a(t) \cdot f(a,k,t) \cdot n_a(t) \cdot \alpha(b)\} + o(\Delta t) \quad (4.21)$$

where

$$\alpha(b) = \sum_{\substack{u=1 \\ u \neq b}}^{|\mathcal{B}|} \left(\sum_{\substack{v=1 \\ v \neq b}}^{u} R^k(v) \cdot p_v \right) \cdot \frac{p_b}{\sum_{w=1}^{|\mathcal{B}|} p_w - \sum_{v=1}^{u} p_v} \tag{4.22}$$

represents the total probability that MT prefers other access nodes serving the same area, but those access nodes cannot accommodate the request due to lack of resources, and $R^k(v)$ is the probability that access node v rejects a request of class k, i.e.,

$$R^k(v) = P\left(\left\{ \sum_{k \in \mathcal{C}} \sum_{a \in A(b)} x_a^k(v) \right\} + bw(k) > C(v) \right) \tag{4.23}$$

where $C(v)$ is the capacity allocated to access node v.

4.4.1.3. Direct Incoming NC Probability. The probability that a direct incoming new connection attempt of class k occurs for access node b is

$$P_{ni}(b, k, \Delta t) = \Delta t \cdot \sum_{a \in \mathcal{A}(b)} \left\{ r_a^k(t) \cdot n_a(t) \cdot p(a, k, s(b), t) \right\} + o(\Delta t) \ . \tag{4.24}$$

4.4.1.4. Indirect Incoming NC Probability. The probability that an indirect incoming new connection attempt of class k occurs for access node b is

$$\widetilde{P}_{ni}(b, k, \Delta t) = \Delta t \cdot \sum_{a \in \mathcal{A}(b)} \left\{ r_a^k(t) \cdot n_a(t) \cdot \alpha(b) \right\} + o(\Delta t) \ . \tag{4.25}$$

Thus, the probability that a new connection attempt of class k occurs for access

node b is

$$\begin{aligned} P_{new}(b,k,\Delta t) &= P_{no}(b,k,\Delta t) + \widetilde{P}_{no}(b,k,\Delta t) \\ &\quad + P_{ni}(b,k,\Delta t) + \widetilde{P}_{ni}(b,k,\Delta t) \\ &\quad + o(\Delta t)\ . \end{aligned} \tag{4.26}$$

4.4.2. Migration Event Probabilities

4.4.2.1. Intra-cell Movement Probability. The probability that a mobile that has an active connection of class k over access node b moves from one area to another in the same cell without changing its access node is

$$P_{intra}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j \in \mathcal{A}(b)} \sum_{a_i \in \mathcal{A}(b) a_i \neq a_j} V^k_{a_j,a_i}(t) \cdot x^k_{a_j}(b) + o(\Delta t)\ . \tag{4.27}$$

4.4.2.2. Intra-subsystem Handoff Probability. The probability that access node b receives a handoff request of class k from another cell in the same subsystem, $s(b)$, is

$$P_{intra}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j \in \mathcal{A}} \sum_{a_i \in \mathcal{A}(b)} \sum_{c \in \mathcal{B},\, s(c)=s(b),\, c \neq b} V^k_{a_j,a_i}(t) \cdot x^k_{a_j}(c) + o(\Delta t)\ . \tag{4.28}$$

4.4.2.3. Direct Inter-subsystem Handoff Probability. The probability that access node b receives a direct handoff request of class k from another cell of another subsystem, $s(b)$, is

$$P_{inter}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j \in \mathcal{A}} \sum_{a_i \in \mathcal{A}(b)} \sum_{c \in \mathcal{B},\, s(c) \neq s(b)} V^k_{a_j,a_i}(t) \cdot x^k_{a_j}(c) \cdot p(a_j,k,s(b)) + o(\Delta t)\ . \tag{4.29}$$

4.4.2.4. Indirect Inter-subsystem Handoff Probability. The probability that access node b receives an indirect handoff request of class k from another cell of another subsystem, $s(b)$, is

$$\widetilde{P}_{inter}(b,k,\Delta t) = \Delta t \cdot \sum_{a_j \in \mathcal{A}} \sum_{a_i \in \mathcal{A}(b)} \sum_{c \in \mathcal{B},\, s(c) \neq s(b)} V^k_{a_j,a_i}(t) \cdot x^k_{a_i}(c) \cdot \alpha(b) + o(\Delta t)\ . \quad (4.30)$$

Thus, the probability that a handoff attempt of class k occurs for access node b is

$$\begin{aligned} P_{handoff}(b,k,\Delta t) &= P_{intra}(b,k,\Delta t) \\ &\quad + P_{inter}(b,k,\Delta t) \\ &\quad + \widetilde{P}_{inter}(b,k,\Delta t) \\ &\quad + o(\Delta t)\,. \end{aligned} \quad (4.31)$$

4.4.3. Hangup Event Probability

The probability that MT terminates the connection voluntarily is

$$P_{hangup}(b,k,\Delta t) = \Delta t \cdot \sum_{a \in \mathcal{A}(b)} h^k_a(t) \cdot x^k_a(t) \cdot p(a,k,s(b),t) + o(\Delta t) \quad (4.32)$$

where $h^k_a(t)$ is the rate at which class k connections in area a hangup at time t.

5. CALCULATING STATE PROBABILITIES

5.1. Analytical Approach

Calculating the state probabilities for such a complex system analytically is a challenge on its own. The state probabilities in our dynamic system can be calculated by solving the eigenvalue problem of the transition matrix.

We denote the probability that the system is in state g_i at time t as $\mathbf{P}_i(t)$. Initially, the system starts in state g_0, in which there are no ongoing connections. Thus, $\mathbf{P}_0(0)$, the probability of being in state g_0 at time 0, is equal to 1.0 while $\mathbf{P}_i(0) = 0.0$ for all other states g_i. As time elapses, the elementary events mentioned in Chapter 4 occur, carrying the system from state to state. This causes the state probabilities to change in time.

Though we do not know the system state exactly at a given time, we can calculate the probability of being in each state g_i at time t, literally $\mathbf{P}_i(t)$. During an infinitesimally short time interval Δt, the system may switch to/from g_i due to an elementary event. Therefore, $\mathbf{P}_i(t + \Delta t)$ is different from $\mathbf{P}_i(t)$. The rate of change in the state probability of g_i during Δt, $\mathbf{P}_i'(t)$, is given by

$$\mathbf{P}_i'(t) = \mathbf{P}_i(t) \cdot \left(-\sum_{j \neq i} \lambda_{ij} \right) + \sum_{j \neq i} \mathbf{P}_j(t) \cdot \lambda_{ji} \tag{5.1}$$

where λ_{ij} is the rate of flow from state g_i to g_j. λ_{ij} values are calculated using Equations 4.20-4.32 depending on the state transitions triggered by each elementary event.

In theory, when the system reaches equilibrium we have

$$\forall g_i \in \mathcal{G} \quad \mathbf{P}_i'(t) = 0. \tag{5.2}$$

From Equation 5.2, we derive a method to solve the state probabilities with the assumption that the system is in equilibrium. We use the power method, which is an iterative process, for solving state probabilities. In each iteration, probabilities converge to equilibrium state probabilities that corresponds to the eigenvector of largest the eigenvalue of the transition matrix.

In the power method, we start with an initial probability vector that sums up to one. We represent the vector of state probabilities, i.e., the vector composed of state probabilities of all states at time t, as $[\mathbf{P}_i(t)]_{g_i \in \mathcal{G}}$. Then we start the iterations, correcting the state probabilities with the assistance of the transition matrix in each step. In each iteration, we evaluate new probabilities for the next iteration according to Equation 5.3. After sufficient iterations, state probabilities converge to the steady state probabilities. We use the metric $d(\cdot\,,\cdot)$ in Equation 5.4, to represent the difference between two probability vectors. For the rest of our work we use metric $d(\cdot\,,\cdot)$.

$$\mathbf{P}_i(t) = \frac{\sum_{j \neq i} \mathbf{P}_j(t) \cdot \lambda_{ji}}{\sum_{j \neq i} \lambda_{ij}} \tag{5.3}$$

$$d([\mathbf{P}_i(t)]_{g_i \in \mathcal{G}}, [\mathbf{P}_i(t + \Delta t)]_{g_i \in \mathcal{G}}) = \sum_{g_i \in \mathcal{G}} |\mathbf{P}_i(t) - \mathbf{P}_i(t + \Delta t)| \tag{5.4}$$

We utilize metric $d(\cdot\,,\cdot)$ to analyze the convergence of the state probability vector. While solving the system numerically, iteration is stopped when $d(\cdot\,,\cdot)$ evaluated between successive iterations is less than a threshold. In Equation 5.4, $d(\cdot\,,\cdot)$ is the metric derived from ℓ_1-norm. Here, we have two probability vectors, $[\mathbf{P}_i(t)]_{g_i \in \mathcal{G}}$ and $[\mathbf{P}_i(t + \Delta t)]_{g_i \in \mathcal{G}}$. We sum the absolute probability difference of all states in these probability vectors. Calculating $d(\cdot\,,\cdot)$ between identical vectors gives zero due to non-degeneracy of the metric. The final probability vector in the last iteration is close to the probability vector in equilibrium situation according to the metric in Equation 5.4.

5.2. Challenges in Analytical Approach

The analytical model presented in Section 5.1 provides a tool to calculate state probabilities using the power method. However, the size of the state space constitutes a major challenge. Any combination of values for the tuple given in Equation 4.2, as long as the access node capacities are not exceeded, constitutes a different state. To make the problem more tangible, we consider the simple cellular layout in Figure 5.1, where each access node is assigned only four channels. We also assume only one connection class for the sake of simplicity. Below, we show that even this example is a challenging scenario. The reader should note that some of the users in a_2 may communicate over

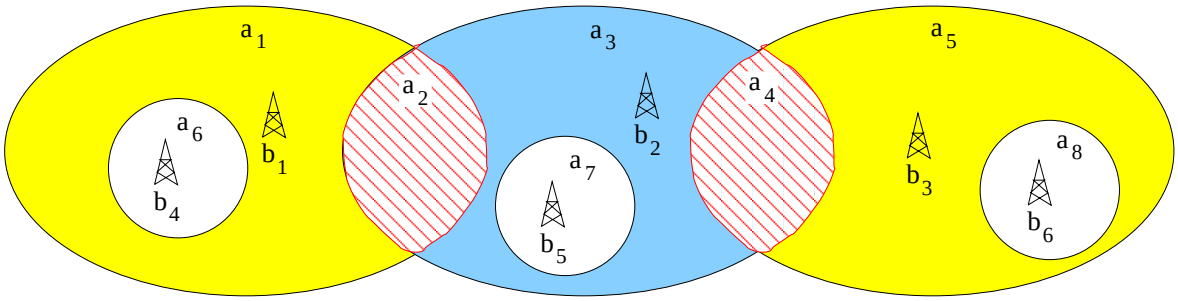


Figure 5.1. Example cellular layout

b_1 while others communicate over b_2. The sum of the variables $x^1_{a_1}(b_1)$, $x^1_{a_2}(b_1)$, and $x^1_{a_6}(b_1)$ should not exceed four, the capacity of b_1. Considering such constraints for all variables, we find that there are 201 different combinations for each access node. Since we have six access nodes, the number of states reaches $201^6 = 65\,944\,160\,601\,201$. However, some of these states are not effectively possible. For example, the value of the variable $x^1_{a_3}(b_1)$ should always be zero. When we exclude such states and consider only *effective states,* the number of states reduces down to $8\,962\,362\,486\,784$. Since it is practically impossible to evaluate a model with so many states, we need to find an intelligent way to consider only states that are relevant for modeling purposes.

5.3. Practical Approach

Due to the challenges mentioned in the previous subsection, we propose another method, which is practical in the number of states considered, subject to some assump-

tions.

The states that represent the cases where the system is far from high load are of no interest for research or system design. Therefore, we analyze the state probabilities given that the system is under particularly higher load. We evaluate the state probabilities for *Fr0 states* (no free channels in any cell), *Fr1 states* (only one free channel available in anyone of the cells), and *Fr2 states* (only two free channels available in the overall network). The reader should note that *Fr0* is not a single state; since there are multiple combinations for $x_a^k(b)$ variables under full load, all channels can be in use in different ways. All channels of a cell may be occupied in different areas of that cell. Similar idea applies for *Fr1* and *Fr2 states*. Assuming that the system is most probably in the states we consider at time t means

$$\sum_{i \in Fr0} \mathbf{P}_i(t) + \sum_{i \in Fr1} \mathbf{P}_i(t) + \sum_{i \in Fr2} \mathbf{P}_i(t) \backsimeq 1, \tag{5.5}$$

which implies the sum of the probabilities of the remaining states is very close to zero. This assumption is expressed by Equation 5.6.

$$\sum_{i \notin (Fr0 \cup Fr1 \cup Fr2)} \mathbf{P}_i(t) \backsimeq 0 \tag{5.6}$$

By considering only the states in *Fr0*, *Fr1*, and *Fr2*, we reduce the number of states down to 156 800, which is reasonable. *Using this approach, we are able to find the probability that the system goes to one of the critical states in Fr0, Fr1, or Fr2, given that it is operating in the range close to full capacity.* Results in the next chapter are produced using this approach.

6. NUMERICAL RESULTS

In this chapter, we analyze the effects of several parameters on the system performance. First, we discuss the convergence of the iterations and the dropping rate, then argue about the effect of migration rate on the system, and finally discuss the effect of connection generation rate on the system using the practical approach in Section 5.3.

In the rest of this section, we use the state definitions in Equations 4.2 and 4.5. We do not consider all states, but we consider only the states in *Fr0*, *Fr1*, and *Fr2*. Hence we do not find the exact state probabilities, but calculate the probabilities of states in *Fr0*, *Fr1*, and *Fr2* subject to Equation 5.5. We denote these *conditional probabilities* by $\mathbf{P_c}$.

6.1. Iteration Parameters

In our tests, we use the following parameters:

Table 6.1. Iteration parameters and their units

Parameter	Explanation	Unit
Δt	Time interval	sec
ε	Threshold for convergence	no unit
MR	Migration rate	$users/sec$
CGR	Connection generation rate	$conn/sec$
HUP	Hangup rate	$conn/sec$
$NUSR$	Number of users	$users/area$

To allow the system load to stabilize, we keep HUP equal to CGR in the tests.

6.2. Convergence with Respect to $d(\cdot\,,\cdot)$

We first analyze how the system converges to steady state. For the system to converge with respect to the metric $d(\cdot\,,\cdot)$, the change in state probabilities should decrease as the number of iterations increase.

In Figure 6.1, we analyze convergence with different values of Δt. The x-axis represents iterations and the y-axis represents the metric evaluated for two probability vectors formed in successive iterations. For larger time increments, the state probabilities change faster in the beginning. However, the system converges around $3500 - 4000$ iterations for all cases. With different time increments the system converges to the

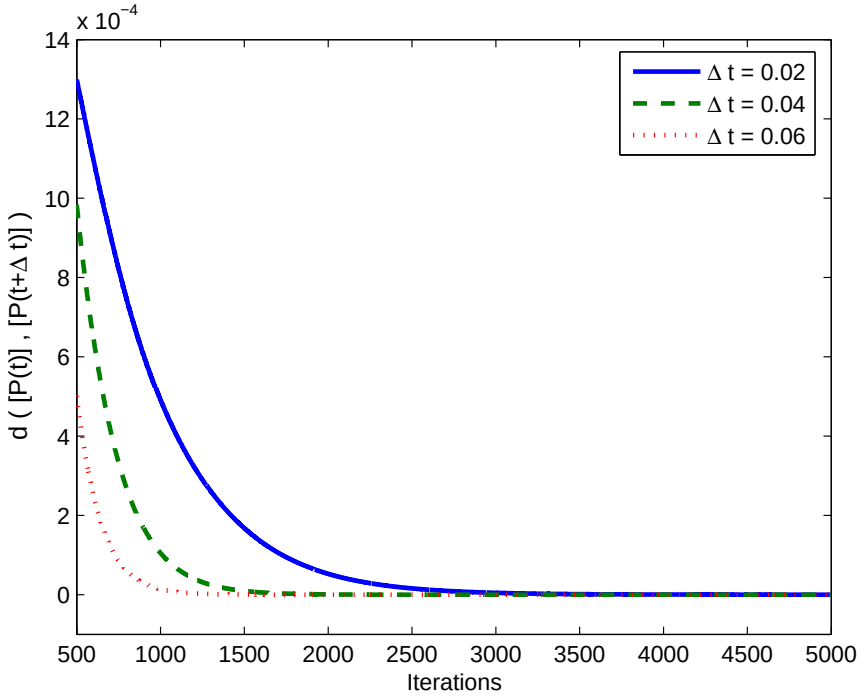


Figure 6.1. Effect of Δt on convergence with respect to $d(\cdot\,,\cdot)$; ($MR = 0.1$, $CGR = 0.25$, $NUSR = 50$, $HUP = 0.25$)

same $\mathbf{P_c}$ values. Table 6.2 illustrates that the system converges to approximately the same $\mathbf{P_c}$ value for different Δt. We have used a time increment of $0.05\,s$ in the rest of the tests.

Table 6.2. $\mathbf{P_c}$ values for different Δt.

	$\mathbf{P_c}[Fr0]$	$\mathbf{P_c}[Fr1]$	$\mathbf{P_c}[Fr2]$
$\Delta t = 0.02$	0.03559912	0.22212804	0.74231886
$\Delta t = 0.04$	0.03559909	0.22212781	0.74231946
$\Delta t = 0.06$	0.03559913	0.22212803	0.74231922

6.3. Convergence of $P_c(dropping)$

In addition to convergence with respect to metric $d(\cdot, \cdot)$, we also evaluate the conditional probability of dropping active connections, $\mathbf{P_c}(dropping)$. In Figure 6.2, the x-axis represents the number of iterations and the y-axis represents the conditional probability of dropping. The tests are performed using different values for migration rate, MR. It is apparent from the figure that increasing the migration rate also increases $\mathbf{P_c}(dropping)$. Furthermore, we observe that the tests converge around 1000 iterations for all values of MR.

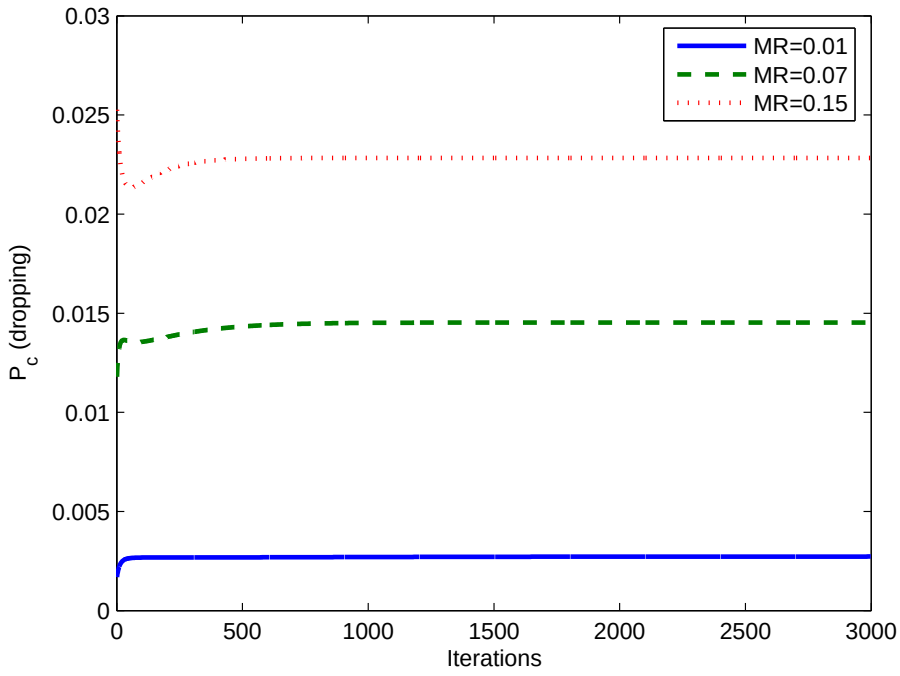


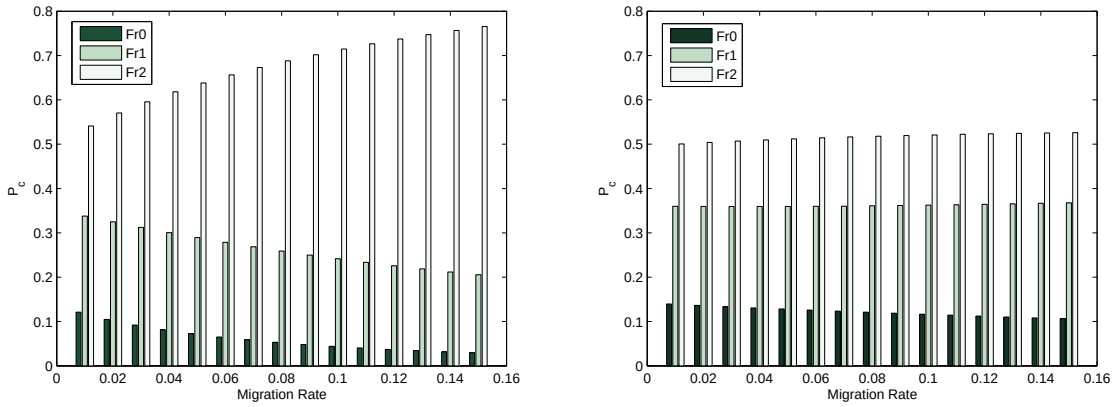
Figure 6.2. Effect of MR on convergence of $\mathbf{P_c}(dropping)$; ($\Delta t = 0.05$, $CGR = 0.3125$, $NUSR = 50$, $HUP = 0.3125$)

6.4. Effect of Migration Rate

Migration rate changes the dynamics of the system. To analyze the effect of migration rate, we apply the practical approach with different MR values. We sum $\mathbf{P_c}$ of states in *Fr0*, *Fr1*, and *Fr2* separately to analyze where the system inclines to. For example, higher $\mathbf{P_c}$ value for *Fr2* means that the system mostly operates in *Fr2* states, when it is highly loaded. In other words, having two channels available in the whole network is more probable than having no available channels subject to Equation 5.5. We examine the effects of MR using single and multiple CGR values in the following subsections.

6.4.1. Effect of MR for Single Connection Generation Rate

We analyze the behavior of the system with respect to MR for two different CGR values. We test the system performance for $CGR = 0.3125$ and $CGR = 3.125$.



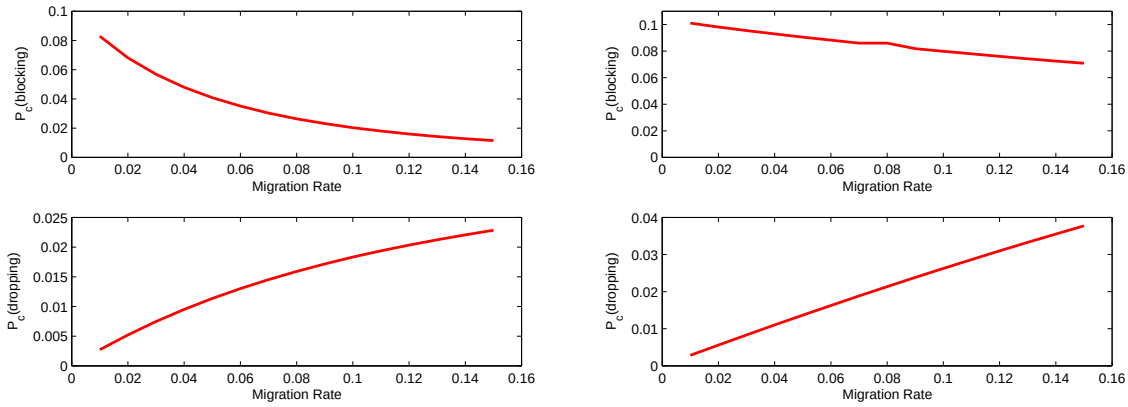
(a) $CGR = HUP = 0.3125$ (b) $CGR = HUP = 3.125$

Figure 6.3. Effect of migration rate on $\mathbf{P_c}$; ($\Delta t = 0.05$, $NUSR = 50$)

In Figure 6.3, the x-axis represents MR and the y-axis represents the sum of $\mathbf{P_c}$ for *Fr0*, *Fr1*, and *Fr2* states. For these tests, we fix the value of CGR and vary MR. For each MR value, probability vector converges to the equilibrium vector. Summing up $\mathbf{P_c}$ values for *Fr0*, *Fr1*, and *Fr2* allows us to grasp where the system operates. Desired operation mode is where $\mathbf{P_c}[Fr2]$ is higher than $\mathbf{P_c}[Fr1]$ and $\mathbf{P_c}[Fr0]$. If $\mathbf{P_c}[Fr2]$ value is the highest, then it means probability of being in one of the states in

Fr2 is higher than being in one of the states in *Fr1* or *Fr0*.

The effect of the increase in CGR can be better observed by comparing Figure 6.3(a) and (b). Since CGR value is relatively higher in Figure 6.3(b) it is apparent that the change in MR value does not affect the system as in Figure 6.3(a). The system reacts to the increase in MR value similarly but slower, since there are more connection attempts.



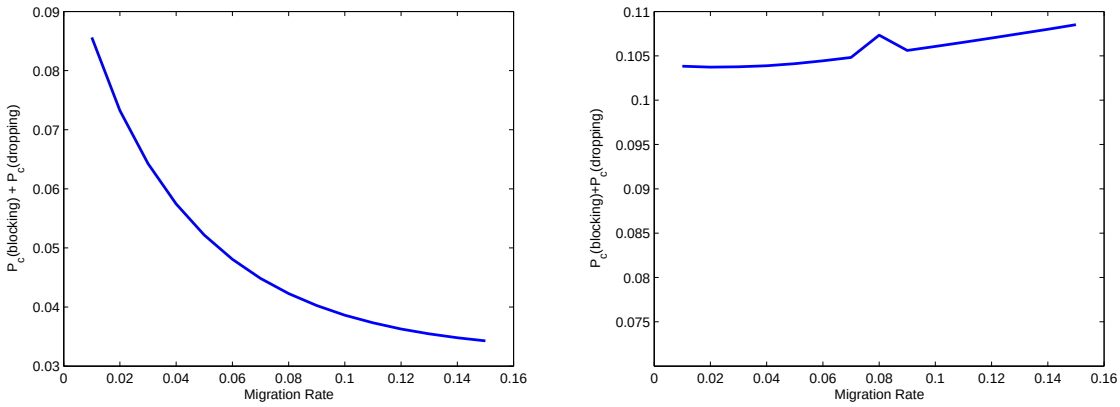
(a) $CGR = HUP = 0.3125$ (b) $CGR = HUP = 3.125$

Figure 6.4. Effect of migration rate on $\mathbf{P_c}(blocking)$ and $\mathbf{P_c}(dropping)$; ($\Delta t = 0.05$, $NUSR = 50$)

We focus on the case $CGR = 0.3125$ to analyze the effect of MR on $\mathbf{P_c}$. The inclination of the system towards *Fr2* can be analyzed by comparing Figures 6.3(a) and 6.4(a). When MR increases, $\mathbf{P_c}$ value of *Fr2* increases and $\mathbf{P_c}$ values of *Fr1* and *Fr0* decrease. This means the system is inclined towards *Fr2* states as MR values increase. Though this behavior seems strange at first, it can be explained by analyzing $\mathbf{P_c}(dropping)$. Figure 6.4(a) depicts dropping and blocking with respect to migration rate. As the migration rate increases, many calls are dropped since we operate close to full capacity. Thus, $\mathbf{P_c}(dropping)$ increases with increasing MR. With every dropped connection, a channel is released in the system. Since, there are more available channels, the system has room for new connection requests, resulting in decreasing $\mathbf{P_c}(blocking)$. We also observe that, when we increase the CGR value 10 times, we monitor higher blocking and dropping rates since released channels are occupied faster. In Figure 6.4(b), $\mathbf{P_c}(blocking)$ value is over 0.06. However, in Figure 6.4(a), we have

$\mathbf{P_c}(blocking)$ value is below 0.08.

Figure 6.5 depicts sum of dropping and blocking with respect to migration rate for both cases. As the migration rate increases, the sum of $\mathbf{P_c}(blocking)$ and $\mathbf{P_c}(dropping)$ decreases in Figure 6.5(a). However, in Figure 6.5(b), the sum increases with increasing MR. In Figure 6.4(b), since the load is 10 times higher, the decrease in $\mathbf{P_c}(blocking)$ is less than that in Figure 6.4(a). This situation also explains the increase in Figure 6.5(b).



(a) $CGR = HUP = 0.3125$ (b) $CGR = HUP = 3.125$

Figure 6.5. Effect of migration rate on $\mathbf{P_c}(blocking) + \mathbf{P_c}(dropping)$; ($\Delta t = 0.05$, $NUSR = 50$)

6.4.2. Effect of MR for Multiple Connection Generation Rates

We also analyze the effect of MR for different CGR values on $\mathbf{P_c}[Fr0]$, $\mathbf{P_c}[Fr1]$, and $\mathbf{P_c}[Fr2]$ individually. First we analyze $\mathbf{P_c}[Fr0]$ and $\mathbf{P_c}(blocking)$ together. In Figure 6.6, the x-axis represents MR values, the y-axis represents sum of $\mathbf{P_c}$ values of states in *Fr0*, and each curve corresponds to a different CGR value. Similar to Figure 6.3, as MR value increases, all the curves corresponding to different CGR values decrease. We also observe that tests with higher CGR values result in higher $\mathbf{P_c}[Fr0]$ since there are more connection attempts with higher CGR values. Thus, the system leans toward *Fr0* more than *Fr1* and *Fr2*. Hence we have higher $\mathbf{P_c}(blocking)$ for higher CGR values as seen in Figure 6.7.

Similar behavior is observed for *Fr1* in Figure 6.8. However, for *Fr2* the situation

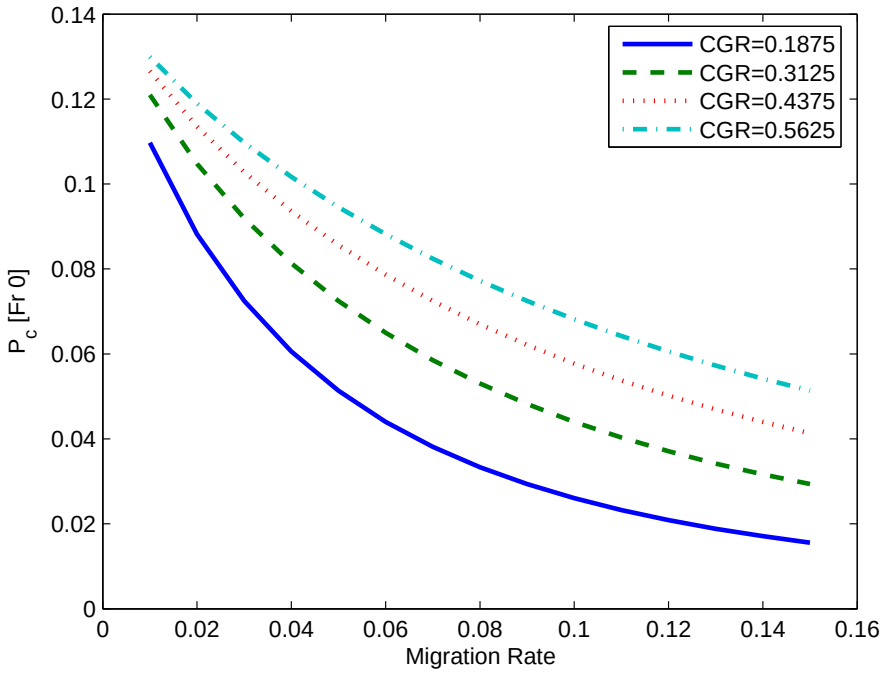


Figure 6.6. Effect of migration rate on $\mathbf{P_c}[Fr0]$ with multiple CGR values; ($\Delta t = 0.05$, $NUSR = 50$)

changes. In Figure 6.9, the x-axis represents MR values, the y-axis represents $\mathbf{P_c}[Fr2]$, and each curve corresponds to different CGR value. As MR value increases, all the curves corresponding to different CGR values increase. The increase in $\mathbf{P_c}[Fr2]$ can be explained by the increase in $\mathbf{P_c}(dropping)$, as shown in Figure 6.10. Dropping events occur more frequently as we increase MR value, so available channels increase frequently in the system.

In Figure 6.10, the x-axis represents MR values, the y-axis represents $\mathbf{P_c}(dropping)$, and each curve corresponds to different CGR value. For a fixed MR value we observe that lower CGR values imply less attempts to fill empty channels in the system. Hence, we expect that lower CGR values result in lower $\mathbf{P_c}(dropping)$ values. We realize this expectation in Figure 6.10.

We analyze the effect of MR for different CGR values, by considering the order of the curves in Figures 6.6-6.8. We observe that the conditional probability decreases

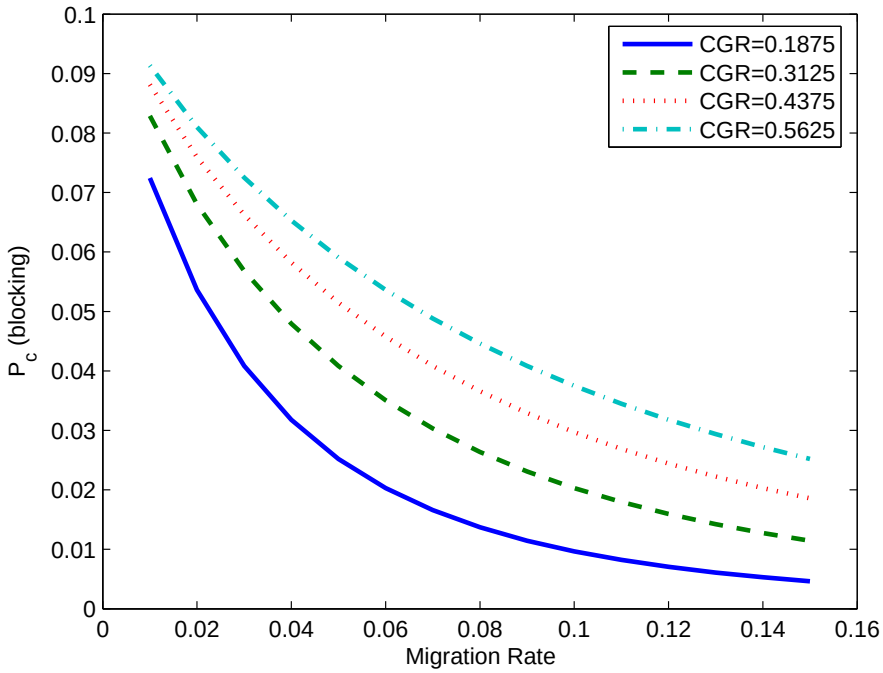


Figure 6.7. Effect of migration rate on $\mathbf{P_c}(blocking)$ with multiple CGR values; ($\Delta t = 0.05$, $NUSR = 50$)

with increasing MR and decreasing CGR. In figures, curves corresponding to smaller CGR values are below the others and decreasing. However, in Figure 6.9, the conditional probability increases with increasing MR and lower CGR. In the figure, the curves corresponding to smaller CGR values are above the others and increasing. When CGR value is lower we expect that $\mathbf{P_c}$ value of being in *Fr0* is lower. In Figures 6.6 and 6.8, curve corresponding to the smallest CGR value is below the others. Due to lower $\mathbf{P_c}$ value of being in *Fr0* and *Fr1*, we expect that $\mathbf{P_c}$ value of being in *Fr2* is higher than the others. We observe the positive effect of higher MR value on the system since *Fr2* states are better states than *Fr0* and *Fr1* states. However, we analyze the effect of MR in Figure 6.10 on $\mathbf{P_c}(dropping)$. We observe that increasing MR value results in higher $\mathbf{P_c}(dropping)$.

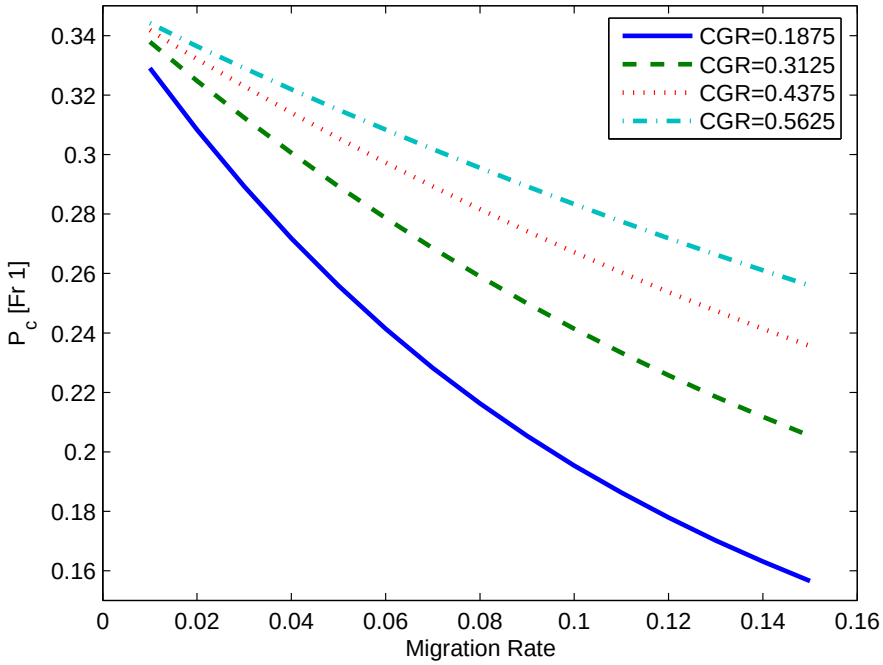


Figure 6.8. Effect of migration rate on $\mathbf{P_c}[Fr1]$ with multiple CGR values; ($\Delta t = 0.05$, $NUSR = 50$)

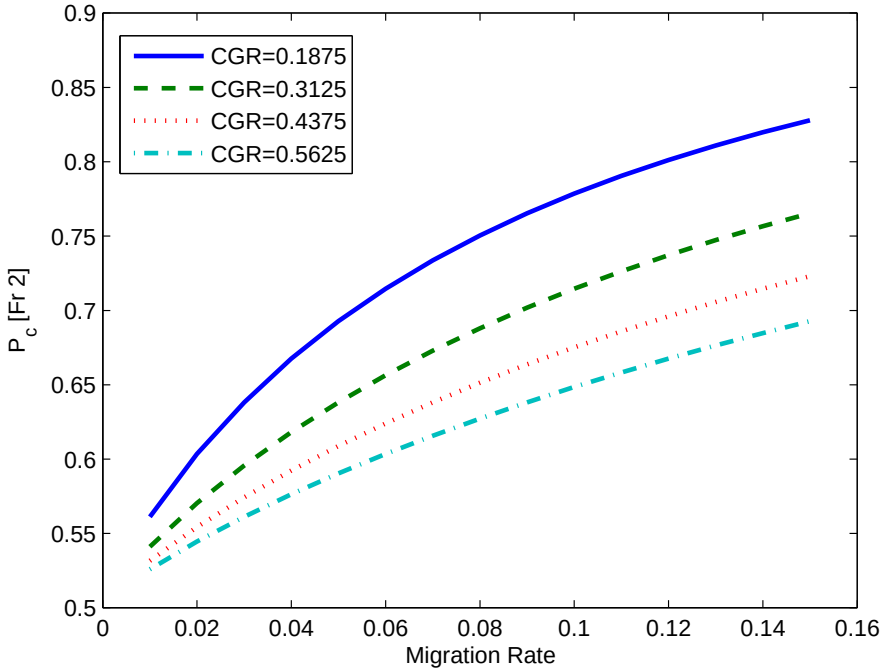


Figure 6.9. Effect of migration rate on $\mathbf{P_c}[Fr2]$ with multiple CGR values; ($\Delta t = 0.05$, $NUSR = 50$)

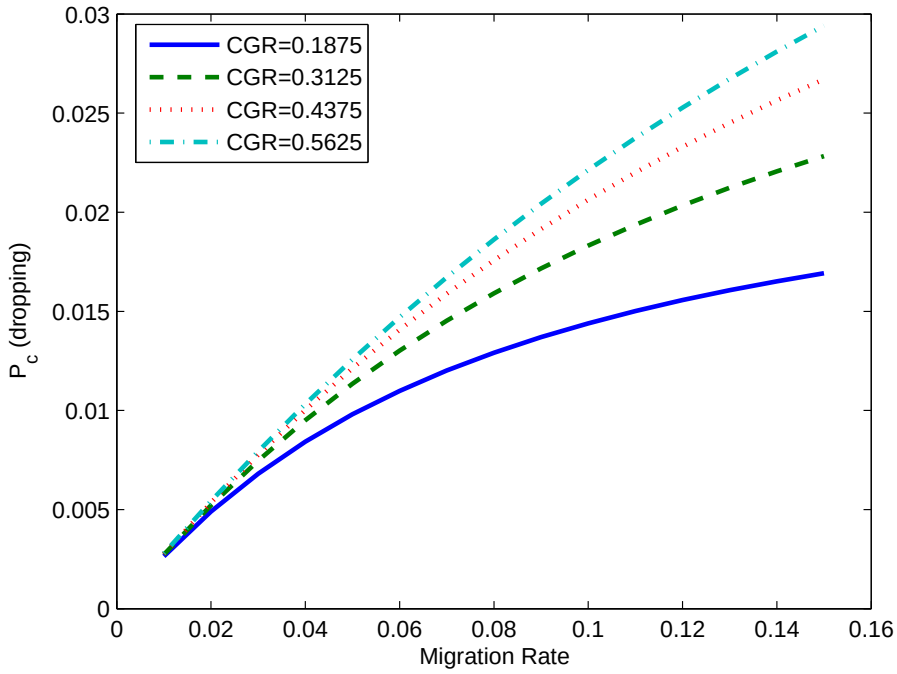


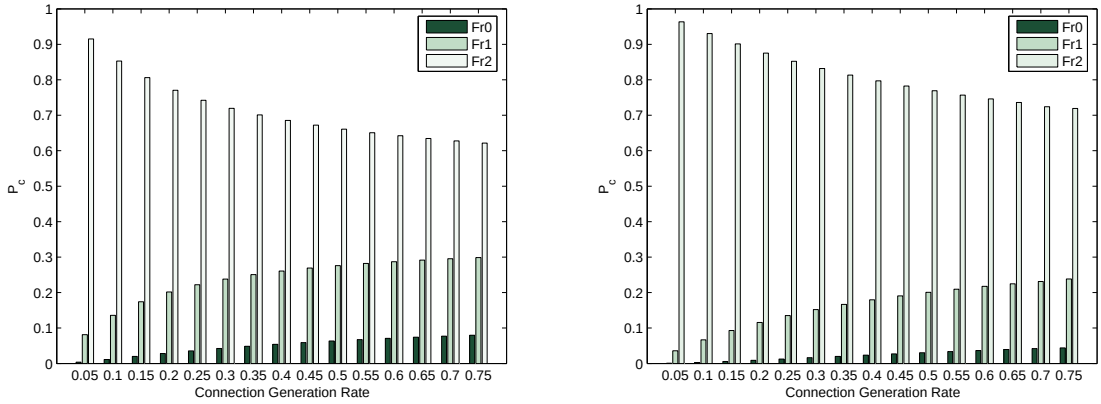
Figure 6.10. Effect of migration rate on $\mathbf{P_c}(dropping)$ with multiple CGR values; ($\Delta t = 0.05$, $NUSR = 50$)

6.5. Effect of Connection Generation Rate

Connection generation rate is another factor that changes the dynamics of the system. To analyze the effect of connection generation rate, we apply the practical approach with different CGR parameter values. We examine the effects of CGR using single and multiple MR values.

6.5.1. Effect of CGR for Single Migration Rate

We analyze the effect of CGR for single MR value with two different cases. We have two different MR values to understand how the system reacts to the change in CGR under different MR values. We test system performance with MR fixed at 0.1 and also ranging from 0.05 $users/sec$ to 0.15 $users/sec$.



(a) $MR = 0.1$ (b) $MR = \{0.05, 0.1, 0.15\}$

Figure 6.11. Effect of connection generation rate on $\mathbf{P_c}$; ($\Delta t = 0.05$, $NUSR = 50$)

In Figure 6.11, we analyze the effect of connection generation rate on $\mathbf{P_c}$. We vary the connection generation rate from 0.05 to 0.75 $conn/sec$, and keep MR constant. As the connection generation rate increases, we observe that the system shifts from *Fr2* states to *Fr1* and *Fr0* states. Thus, the analytical model provides a tool for the system designer to understand with what probability the system goes to *Fr0* for the given load.

In Figure 6.11(b), we analyze the effect of connection generation rate on $\mathbf{P_c}$ with variable MR value. Since MR values change from area to area, we get a non-uniform distribution of users over the service area when the system reaches equilibrium. Areas a_0, a_3, a_4 have 34 users, a_1, a_2 have 105 users, a_5, a_7 have 18 users, and a_6 has 52 users at the equilibrium point. The starting value for the number of users are 50 as in the previous cases. There are more connection attempts in the two areas where the users mostly accumulate. Other areas are less dense, implying higher $\mathbf{P_c}$ value in the *Fr2* states which have higher x_a^k (b) values for these areas.

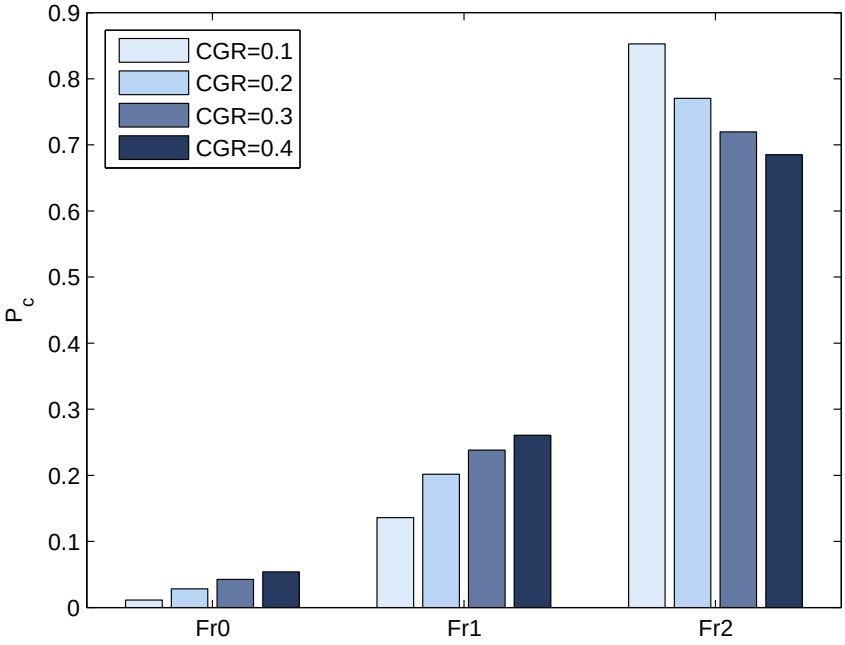


Figure 6.12. State type versus $\mathbf{P_c}$; ($\Delta t = 0.05$, $MR = 0.1$, $NUSR = 50$)

Figure 6.12 depicts the system behavior under different loads. $\mathbf{P_c}$ values for *Fr0*, *Fr1*, and *Fr2* are shown on the x-axis using separate bars for each CGR value. We observe from the figure that the system is mostly in *Fr2* states. As the load increases, the system spends more time in *Fr1* and *Fr0* states. Analyzing the change in $\mathbf{P_c}$ individually for each type of state shows that for *Fr0* and *Fr1* states, $\mathbf{P_c}$ increases as the load increases. However, for *Fr2* states, $\mathbf{P_c}$ decreases as the load increases since there is a shift towards *Fr0* states.

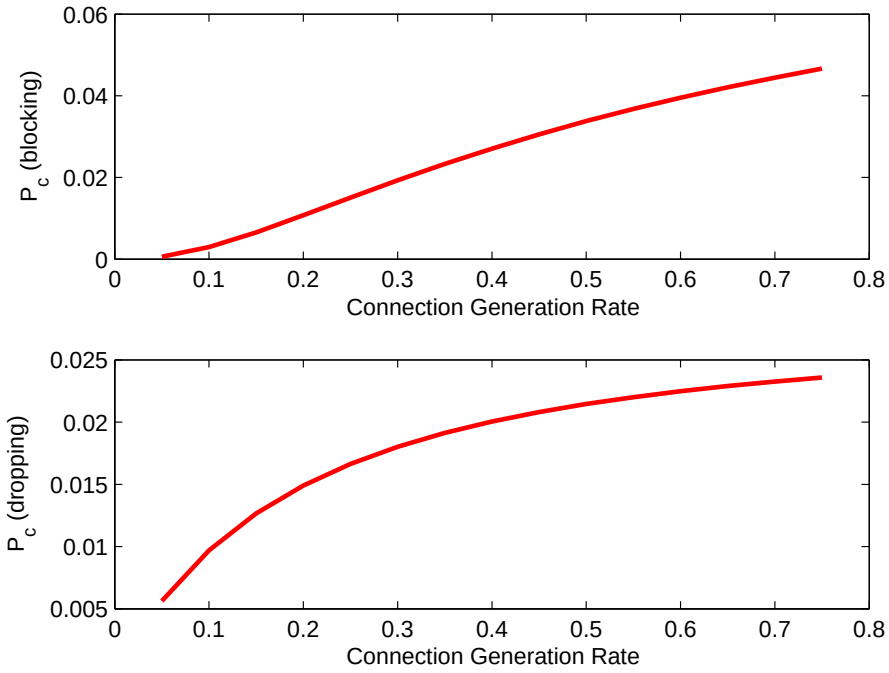


Figure 6.13. Effect of connection generation rate on $\mathbf{P_c}(blocking)$ and $\mathbf{P_c}(dropping)$; ($\Delta t = 0.05$, $MR = 0.1$, $NUSR = 50$)

In Figure 6.13, we analyze the effect of CGR on $\mathbf{P_c}(dropping)$ and $\mathbf{P_c}(blocking)$. The x-axis represents CGR, the y-axis represents $\mathbf{P_c}(blocking)$ and $\mathbf{P_c}(dropping)$, respectively. The figure demonstrates the link between $\mathbf{P_c}(blocking)$ and $\mathbf{P_c}(dropping)$ while MR is fixed. We observe that the increase in CGR causes an almost linear increase in $\mathbf{P_c}(blocking)$. However, the increase rate of $\mathbf{P_c}(dropping)$ slows down for higher CGR values. This behavior can be explained by the fact that we are operating close to full capacity. For higher CGR values, we know from Figure 6.11, that the probability of the system being in *Fr0* and *Fr1* states is higher. Hence, increasing CGR further does not add many more connections into the system since most of these attempts are blocked. Therefore, the increase rate of $\mathbf{P_c}(dropping)$ slows down as CGR increases further.

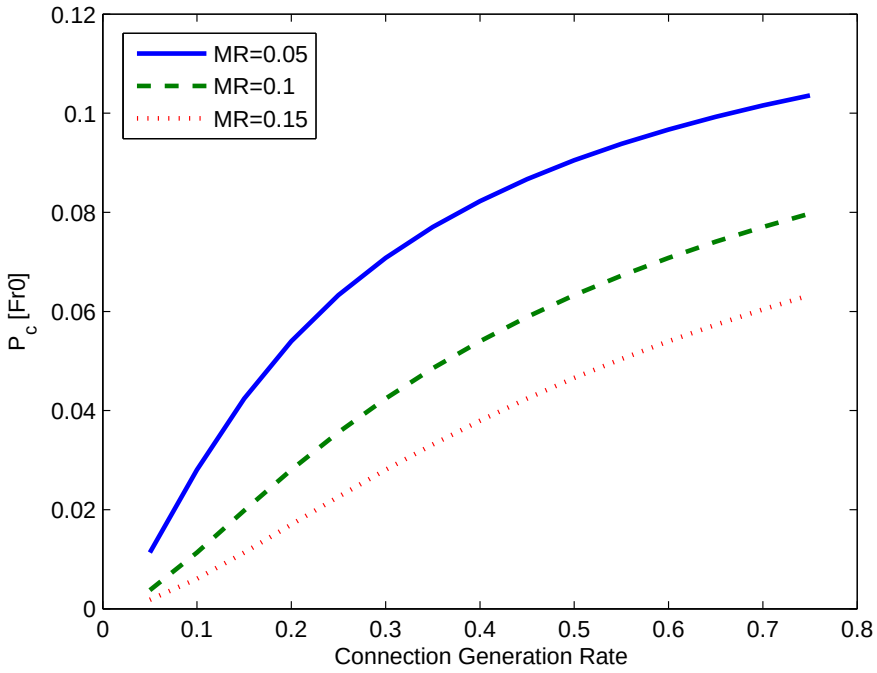


Figure 6.14. Effect of connection generation rate on $\mathbf{P_c}[Fr0]$ for multiple MR values; ($\Delta t = 0.05$, $NUSR = 50$)

6.5.2. Effect of CGR for Multiple Migration Rate

We also analyze effect of CGR for different MR values on $\mathbf{P_c}[Fr0]$, $\mathbf{P_c}[Fr1]$, and $\mathbf{P_c}[Fr2]$ individually. In Figure 6.14, the x-axis represents CGR values, the y-axis represents sum of $\mathbf{P_c}$ values of states in *Fr0*, and each curve corresponds to a different MR value. The increase in CGR value results in higher $\mathbf{P_c}[Fr0]$ since there are more connection attempts. We also observe that tests with higher MR values result in lower $\mathbf{P_c}[Fr0]$, as explained previously in Section 6.4.

In Figure 6.15, the x-axis represents CGR values, the y-axis represents $\mathbf{P_c}(blocking)$, and each curve corresponds to a different MR value. The increase in CGR value results in higher $\mathbf{P_c}(blocking)$ since more connection attempts are blocked. We also observe that tests with higher MR values result in lower $\mathbf{P_c}(blocking)$, as explained previously in Section 6.4. There is a close relation between $\mathbf{P_c}[Fr0]$ and $\mathbf{P_c}(blocking)$ since blocking events occur more frequently when the system is in *Fr0* states. So

$\mathbf{P_c}(blocking)$ increases as CGR value increases due to increase in probability of being in *Fr0*. For higher MR values $\mathbf{P_c}(dropping)$ is higher which implies lower $\mathbf{P_c}[Fr0]$ value. Due to lower $\mathbf{P_c}[Fr0]$ value, $\mathbf{P_c}(blocking)$ is lower for higher MR values.

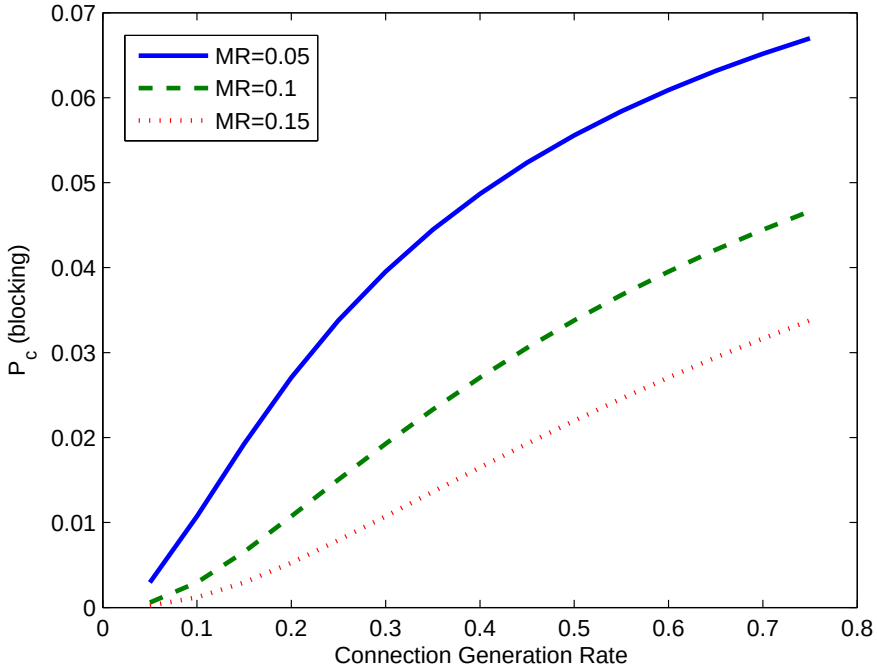


Figure 6.15. Effect of connection generation rate on $\mathbf{P_c}(blocking)$ for multiple MR values; ($\Delta t = 0.05$, $NUSR = 50$)

Similar behavior is observed for *Fr1* in Figure 6.16. However, for *Fr2* the situation changes. In Figure 6.17, the x-axis represents CGR values, the y-axis represents $\mathbf{P_c}[Fr2]$, and each curve corresponds to different MR value. As CGR value increases, $\mathbf{P_c}[Fr2]$ decreases for all curves. We also observe that tests with higher MR values result in higher $\mathbf{P_c}[Fr2]$ since higher MR values imply higher $\mathbf{P_c}(dropping)$, as explained previously in Section 6.4. We observe the negative effect of higher CGR value on the system, since *Fr2* states are better better states than *Fr0* and *Fr1* states.

In Figure 6.18, the x-axis represents CGR values, the y-axis represents $\mathbf{P_c}(dropping)$, and each curve corresponds to different MR value. Dropping events occur more frequently as CGR increases, because channels released due to dropping events are occupied faster. Since the channels are occupied faster, new dropping events are more

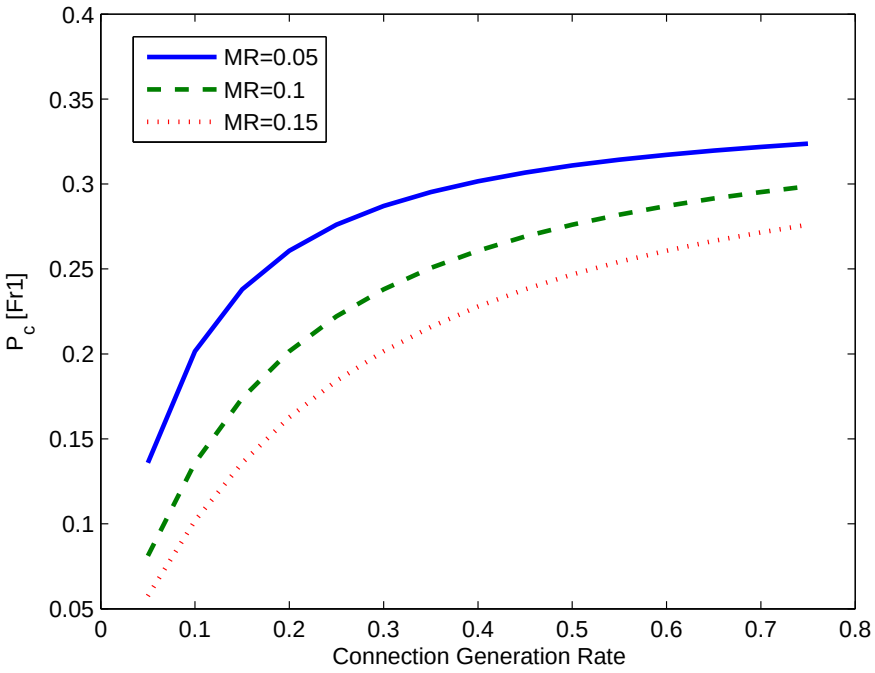


Figure 6.16. Effect of connection generation rate on $\mathbf{P_c}[Fr1]$ for multiple MR values; ($\Delta t = 0.05$, $NUSR = 50$)

probable. However, at high CGR values, the increase rate of $\mathbf{P_c}(dropping)$ slows down as explained in Figure 6.13. For a fixed CGR value we observe that higher MR values imply, higher $\mathbf{P_c}(dropping)$ as explained previously in Section 6.4.

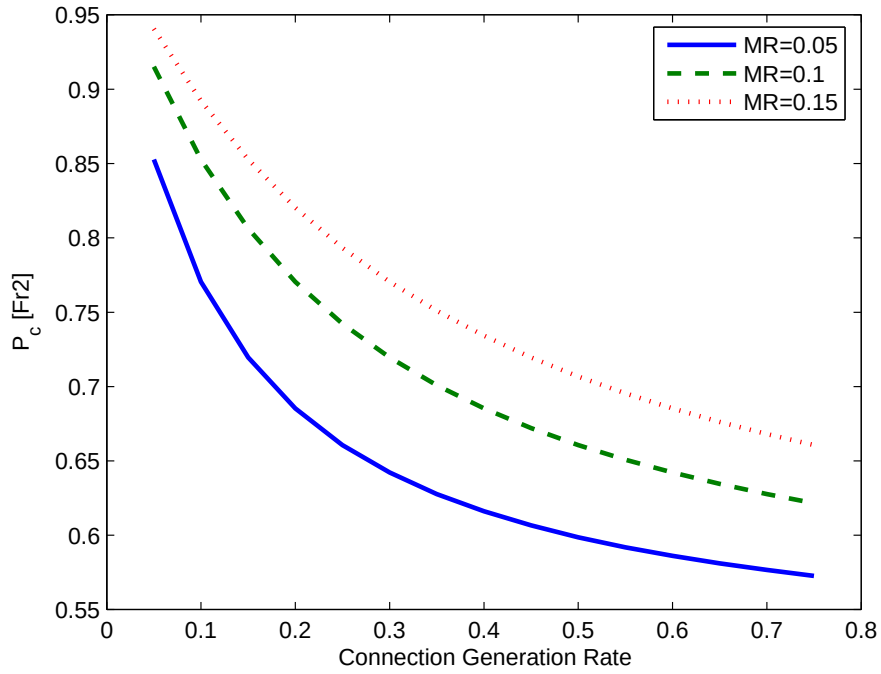


Figure 6.17. Effect of connection generation rate on $\mathbf{P_c}[Fr2]$ for multiple MR values; ($\Delta t = 0.05$, $NUSR = 50$)

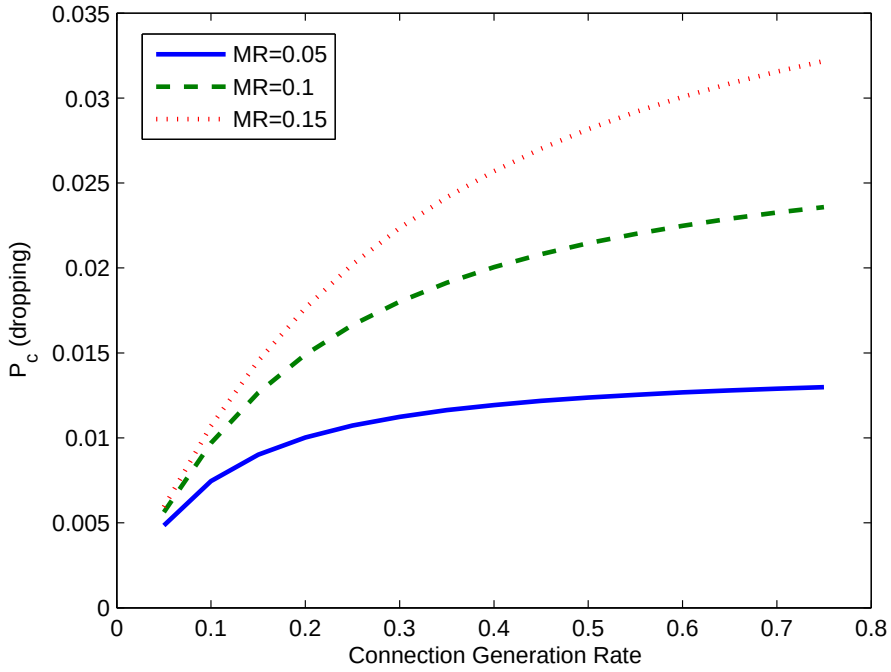


Figure 6.18. Effect of connection generation rate on $\mathbf{P_c}(dropping)$ for multiple MR values; ($\Delta t = 0.05$, $NUSR = 50$)

7. CONCLUSIONS AND FUTURE WORK

NGWS will be composed of multiple subsystems. The selection of the appropriate subsystem for connection setup is a crucial issue for the overall system performance. In this thesis, after defining the network architecture, we proposed a novel connection admission control scheme. The proposed NGCAC scheme considers the accessibility of the subsystems and the availability of the resources in those subsystems in addition to connection class of the connection request and the user's preferences. We also provided an analytical model of the proposed scheme. We pointed out major challenges in analytically modeling NGWS and provided a practical approach as a workaround.

With the practical approach we examine some test cases. First, we analyze how the system converges to equilibrium. Then, we focus on the effect of migration rate. We observe that, the system inclines towards *Fr2* states when we increase MR value. *Fr2* states are more preferable than *Fr1* and *Fr0* states for the system designer. We analyze the system behavior and calculate $\mathbf{P_c}(blocking)$ and $\mathbf{P_c}(dropping)$ corresponding to the given MR values. Inclination toward *Fr2* states can be explained by the increase in $\mathbf{P_c}(dropping)$ for higher MR values. Since we consider states close to full capacity, many calls are dropped due to higher MR value. Therefore, the system is inclined towards *Fr2* states as MR increases. This causes a decrease in $\mathbf{P_c}(blocking)$ values. We observe a similar effect for MR with different CGR values. Furthermore, as the connection generation rate increases, we observe that the system shifts from *Fr2* states to *Fr1* and *Fr0* states.

As future work, we will consider more complex scenarios and evaluate the effects of more parameters on the performance of NGWS and consider implementing the negotiation part of the NGCAC scheme. We will revise the analytical model for load sharing. We will work on other methods for evaluating state probabilities to make evaluation of more complex scenarios feasible to consider. We will also consider revising the analytical model to allow packet-based air interface.

APPENDIX A: IMPLEMENTATION DETAILS

We developed a computer program, named *NGWS-Solver*, to obtain the results presented in Chapter 6. NGWS-Solver uses the ideas presented in Section 5.1 and 5.3 in an iterative way. In each iteration, NGWS-Solver evaluates the new transition probabilities and then uses Equation 5.3 to evaluate new state probabilities. In each iteration, state probabilities of the desired states converge to $\mathbf{P_c}$. Therefore, we evaluate the conditional state probabilities iteratively.

A.1. Implementation Platform

We have run NGWS-Solver on a computer with the specifications given in Table A.1: We have run NGWS-Solver nearly 300 times in total, for different conditions and results. Each run implements on the average 20 000 iterations in the last phase.

Table A.1. Execution platform

Operating System	Linux
Kernel	2.4.20-8
Compiler	GCC
Processor	3.6GHz dual Intel processors
RAM	6 GB

A.2. Pseudo Codes

NGWS-Solver is implemented in the following four phases.

A.2.1. Phase-0

In phase-0, we prepare the parameter file and the definition file for effective areas.

Table A.2. Pseudo code of phase-0

1. Check arguments;
2. Prepare parameter file `params.txt`;
3. Prepare effective area file `SYS_effAreas.txt`;

A.2.2. Phase-1

In phase-1, we prepare the state definition file according to input files and files prepared by phase-0.

Table A.3. Pseudo code of phase-1

1. Read subsystems from file `SYS_subsys.txt`;
2. Read access nodes from file `SYS_accNodes.txt`;
3. Read effective areas from file `SYS_effAreas.txt` (prepared by phase-0);
4. Read connection classes from file `SYS_connClasses.txt`;
5. Evaluate total capacity;
6. DistributeLoad(totalcapacity) (forming L0-states);
7. DistributeLoad(totalcapacity - 1) (forming L1-states);
8. DistributeLoad(totalcapacity - 2) (forming L2-states);
9. WriteStates2File(0, `stL0.txt`);
10. WriteStates2File(1, `stL1.txt`);
11. WriteStates2File(2, `stL2.txt`);

A.2.3. Phase-2

In phase-2, we prepare the transition probabilities for the next phase.

Table A.4. Pseudo code of phase-2

```
1.  Read subsystems from file SYS_subsys.txt;
2.  Read access nodes from file SYS_accNodes.txt;
3.  Read effective areas from file SYS_effAreas.txt (prepared by phase-0);
4.  Read connection classes from file SYS_connClasses.txt;
5.  Evaluate total capacity;
6.  Evaluate number of states;
7.  Allocate space for states;
8.  ReadStates(L0) (prepared by phase-1);
9.  ReadStates(L1) (prepared by phase-1);
10. ReadStates(L2) (prepared by phase-1);
11. Allocate space for transitions;
12. For each state st do
      12.a  If (IsNewConnectionApplicable(st))
            then EvaluateNewConnProbs(st);
      12.b  If (IsMigrationEventApplicable(st))
            then {
              EvaluateMigProbs(st);
              EvaluateDropProbs(st);
            }
      12.c  If (IsHangupApplicable(st))
            then EvaluateHangupProbs(st);
13. WriteTransitions2File(trStates.txt);
```

A.2.4. Phase-3

In phase-3, we evaluate state probabilities iteratively according to the transition probabilities.

Table A.5. Pseudo code of phase-3

```
1.  Read parameters from file params.txt;
2.  Read effective areas from file SYS_effAreas.txt (prepared by phase-0);
3.  Read transition probabilities from file trStates.txt (prepared by phase-2);
4.  Add loop probabilities;
5.  Normalize transition probabilities;
6.  Initialize probability arrays;
    /* Iteration Starts Here */
7.  for i=0 to i=MAX_ITERATION do
      7.a   for each state st do
              7.a.1   Evaluate inflow for st;
              7.a.2   Evaluate dropping probability for st;
      7.b   Normalize probability arrays;
      7.c   Evaluate user migration for areas;
      7.d   Evaluate new transition probabilities for next iteration;
      7.e   Adjust loop probabilities for next iteration;
      7.f   AppendResults2Files();
      7.g   If converged then break;
8.  Close files;
```

A.3. Data Structures

In phase-0 we do not use any data structure, since it is just a preparation phase. Phase-2 extends the data structures of phase-1. We present the extended data structures in Tables A.6-A.11.

A.3.1. AccNode Structure

AccNode structure stores the access node definitions and parameters specific to the access node.

Table A.6. Data structure: AccNode

struct	AccNode
int	index; (access node id)
int	ssindex; (subsystem id)
int	capacity; (access node capacity)
int	areasInCnt; (number of areas in access node coverage)
int*	areasIn; (link-list of area ids in access node coverage)

A.3.2. Area Structure

Area structure stores information about the areas defined in Section 4.1. It contains the area definition and parameters.

Table A.7. Data structure: Area

struct	Area
int	index; (area id)
int	accNodeCnt; (number of access nodes heard)
int*	accNodesHeard; (link-list of access nodes heard)
int	nhbdAreaCnt; (number of neighbor areas)
int*	nhbdAreas; (link-list of neighbor areas)
float*	nhbdMigrates; (link-list of migration rates to neighbor areas)
float	conGenRate; (connection generation rate in area)
float	hangupRate; (hangup rate in area)
float*	userPrefs; (user preferences)
int	userCnt; (number of users in area)

A.3.3. ConnClass Structure

ConnClass structure stores information about the connection class definition.

Table A.8. Data structure: ConnClass

struct	ConnClass
int	index; (connection class id)
int	weight; (weight of the connection class)

A.3.4. LListNode Structure

LListNode structure defines the nodes stored in the hash table.

Table A.9. Data structure: LListNode

struct	ConnClass
State	data;
int	keyT;
int	keyACC[`ACCN_CNT`]
int	keyF1;
int	keyF2;
struct LListNode*	next;

A.3.5. State Structure

State structure stores information about the state defined in Definition 4.1.6.

A.3.6. TrState Structure

This data structure is the most essential part of the program. Transitions from each state are considered and these transitions are kept in a matrix-like structure. Row

Table A.10. Data structure: State

struct	State
int	id;
int	state2DMatrix[ACCN_CNT][AREA_CNT]
int	totalLoad;

represents the "from state" and column represents the "to state".

Table A.11. Data structure: TrState

struct	TrState
int	toID; (id of the state this transition is directed to)
int	fromID; (id of the state this transition is directed from)
float	currentProb; (multiplied by number of users if necessary)
float	trProb; (raw probability)
int	multByAreaUser; (-1 means not multiply)
int	isPD; (0 means not PD)
int	isDirect; (0 means not direct)
struct TrState*	rowNext;
struct TrState*	colNext;

A.4. Scripts and Output Files

We use shell scripts to run program phases repeatedly to get output. In the shell scripts, we run phase-0 and phase-1 first. Then, we run phase-0, phase-2 and phase-3 repeatedly with different parameters. We run this loop once for every value on the x-axis. One of the scripts is depicted. Other script files are similar in structure. This is an example script file named `01convergence_effDT.scrpt`:

```
#!/bin/sh

if [ $# -ne 8 ] then
  echo "USAGE:$0 start incr n SM_DL MIGRATE CONGENRATE NUSERS HUPRATE"
  exit 1
else
  echo Start Date: `date +"%A %d in %B %Y (%r)"`
  start=$1
  incr=$2
  n=$3
  SM_DL=$4
  MIGRATE=$5
  CONGENRATE=$6
  NUSERS=$7
  HUPRATE=$8

  i=1
  curr_DT=`echo $start-$incr+$i*$incr | bc`
  echo -e "PREPARING FOR SIMULATION"
  echo -e "PH0\t[ACTIVE]"
  echo -e "************************************************"
  ./zezus04_ph0 $SM_DL $curr_DT $MIGRATE $CONGENRATE $NUSERS $HUPRATE
  retCode=$?
```

```
echo -e "RETURN CODE\t=$retCode\t(0 means no error)"
echo -e "************************************************"
echo -e "PH0\t[DONE]\n\n"

echo -e "PH1\t[ACTIVE]"
echo -e "************************************************"
./zezus04_ph1
retCode=$?
echo -e "RETURN CODE\t=$retCode\t(0 means no error)"
./clearPH1.scrpt
echo -e "************************************************"
echo -e "PH1\t[DONE]\n\n"

for j in `seq 1 $n`;
do
  curr_DT=`echo $start-$incr+$j*$incr | bc`

  echo -e "RUNNING SIMULATION WITH --> Delta T = $curr_DT"
  echo -e "************************************************"
  echo -e "PH0\t[ACTIVE]"
  echo -e "************************************************"
  ./zezus04_ph0 $SM_DL $curr_DT $MIGRATE $CONGENRATE $NUSERS $HUPRATE
  retCode=$?
  echo -e "RETURN CODE\t=$retCode\t(0 means no error)"
  echo -e "************************************************"
  echo -e "PH0\t[DONE]\n\n"

  echo -e "PH2\t[ACTIVE]"
  echo -e "************************************************"
  ./zezus04_ph2
  retCode=$?
  echo -e "RETURN CODE\t=$retCode\t(0 means no error)"
```

```
    echo -e "\n**************************************************"
    echo -e "PH2\t[DONE]\n\n"

    echo -e "PH3\t[ACTIVE]"
    echo -e "**************************************************"
    ./zezus04_ph3
    retCode=$?
    echo -e "RETURN CODE\t=$retCode\t(0 means no error)"
    echo -e "\n**************************************************"
    echo -e "PH3\t[DONE]\n\n"

    echo -e "PH4\t[ACTIVE]"
    echo -e "**************************************************"
    ./zezus04_ph4
    retCode=$?
    echo -e "RETURN CODE\t=$retCode\t(0 means no error)"
    echo -e "\n**************************************************"
    echo -e "PH4\t[DONE]\n\n"

  done

  echo -e "CLEANING\t[ACTIVE]"
  echo -e "**************************************************"
  rm -f stL*.txt
  rm -f trStates.txt
  echo -e "**************************************************"
  echo -e "CLEANING\t[DONE]\n\n"

  echo -e "<!!!! FINISHED SIMULATION (01Convergence_effDT)!!!!>"
  echo End Date: `date +"%A %d in %B %Y (%r)"`
fi
```

Main output files are listed below.

- `params.txt`: Output file generated by phase-0, which contains ε and Δt values. Approximate size is 40 bytes.
- `SYS_effAreas.txt`: Output file generated by phase-0, which contains effective area data. Approximate size is 1 KB.
- `stL0.txt`: Output file generated by phase-1, which contains state information of $[Fr0]$ states. Approximate size is 853 KB.
- `stL1.txt`: Output file generated by phase-1, which contains state information of $[Fr1]$ states. Approximate size is 4.2 MB.
- `stL2.txt`: Output file generated by phase-1, which contains state information of $[Fr2]$ states. Approximate size is 13 MB.
- `trStates.txt`: Output file generated by phase-2, which contains state transition probability information. Approximate size is 50 MB.
- `stateRes_%f_%f_%f_%d_%f_%d.txt`: Output file generated by phase-3, which contains state probabilities. Approximate size is 3 MB.
- `stateResSP_%f_%f_%f_%d_%f_%d.txt`: Output file generated by phase-3, which contains accumulated state probabilities of $[Fr0]$, $[Fr1]$ and $[Fr2]$ states separately. Approximate size is 30 bytes.

REFERENCES

1. Cox, D., "Wireless Personal Communications: What is it", *IEEE Personal Communications Magazine*, Vol. 2, No. 2, pp. 20-35, 1995.

2. Tugcu, T. and C. Ersoy, "A New Call Admission Control Scheme Based on Mobile Position Estimation in DS-CDMA Systems", *ACM/Kluwer Journal of Wireless Networks*, Vol. 11, No. 3, pp. 341-351, 2005.

3. Beigy, H. and M. R. Meybodi, "A General Call Admission Policy for Next Generation Wireless Networks", *Journal of Computer Communications, Elsevier Publishing Company*, Vol. 28, No. 16, pp. 1798-1813, October 2005.

4. Akyildiz, I. F., S. Mohanty, and J. Xie, "A Ubiquitous Mobile Communication Architecture for Next-Generation Heterogeneous Wireless Systems", *IEEE Radio Communications Magazine*, Vol. 43, No.6, pp. S29-S36, June 2005.

5. Niyato, D. and E. Hossain, "Call Admission Control for QoS Provisioning in 4G Wireless Networks: Issues and Approaches", *Special Issue of IEEE Network on 4G Network Technologies for Mobile Telecommunications*, Vol. 19, No. 5, pp. 5-11, September 2005.

6. Ramjee, R., R. Nagarayan, and D. Towsley, "On Optimal Call Admission Control in Cellular Networks", *Proc. IEEE INFOCOM '96*, Vol. 1, pp. 43-50, March 1996.

7. Epstein, B. and M. Schwartz, "Predictive QoS-based Admission Control for Multiclass Traffic in Cellular Wireless Networks", *IEEE JSAC*, Vol. 18, pp. 523-534, March 2000.

8. Zhang, T., E. van den Berg, J. Chennikara, P. Agrawal, J. C. Chen, and T. Kodama, "Local Predictive Resource Reservation for Handoff in Multimedia Wireless IP Networks", *IEEE JSAC*, Vol. 19, No.10, pp. 1931-1941, October 2001.

9. Levine, D. A., I. F. Akyildiz, and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept", *IEEE/ACM Trans. Net.*, Vol. 5, pp. 1-12, February 1997.

10. Hou, J., J. Yang, and S. Papavassiliou, "Integration of Pricing With Call Admission Control to Meet QoS Requirements in Cellular Networks", *IEE Trans. Parallel and Distrib. Sys.*, Vol. 19, No. 9, pp. 898-910, September 2002.

11. Akyildiz, I. F., J. Xie, and S. Mohanty, "A Survey of Mobility Management in Next-generation All-IP-based Wireless Systems", *IEEE Wireless Commun.*, Vol. 11, No. 4, pp. 16-28, August 2004.

12. Akyildiz, I. F., J. McNair, J. Ho, H. Uzunalioglu, and W. Wang, "Mobility Management for Next Generation Wireless Systems", *Proc. IEEE*, Vol. 87, No. 8, pp. 1347-1384, August 1999.

13. Ghaheri-Niri, S. and R. Tafazolli, "Cordless-cellular Network Integration for the 3rd Generation Personal Communication Systems", *Proc. IEEE VTC*, Vol. 1, pp. 402-408, 1998.

14. Priscoli, F. D., "Interworking of a Satellite System for Mobile Multimedia Applications With the Terrestrial Networks", *IEEE JSAC*, Vol. 17, No. 2, pp. 385-394, February 1999.

15. Buddhikot, M., G. Chandranmenon, S. Han, Y. W. Lee, S. Miller, and L. Salgarelli, "Design and Implementation of a WLAN/CDMA2000 Interworking Architecture", *IEEE Commun. Mag.*, Vol. 41, No. 11, pp. 90-100, November 2003.

16. Akyildiz, I. F. and W. Wang, "A Dynamic Location Management Scheme for Next Generation Multitier PCS Systems", *IEEE Trans. Wireless Commun.*, Vol. 1, No. 1, pp. 178-189, January 2002.

17. "Inter-PLMN Backbone Guidelines", *GSM Assn. classification*, v. 3.4.0, March

2003.

18. Havinga, P. J. M., G. J. M. Smit, G. Wu, and L. K. Vogniild, "The SMART Project: Exploiting the Heterogeneous Mobile World", *Proc. 2nd Int'l. Conf. Internet Comp.*, Las Vegas, NV, pp. 346-352, June 2001.

19. Glass, S., T. Hiller, S. Jacobs, and C. Perkins, "Mobile IP Authentication, Authorization and Accounting Requirements", *IETF RFC 2977*, October 2000.

20. Rosenbeg, J., H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol", *IETF RFC 3261*, June 2002.

21. Tugcu, T., H. B. Yilmaz, and F. Vainstein, "Analytical Modelling of CAC in Next Generation Wireless Systems", accepted to *Elsevier Computer Networks Journal*.

22. NTT-DoCoMo, "Outline of Fourth-generation Mobile Communications", in *http://www.nttdocomo.co.jp/corporate/rd/new_e/4gen01_e.html*, 2002.

23. Berezdivin, R., R. Breinig, and R. Topp, "Next Generation Wireless Communications Concepts and Technologies", *IEEE Communications Magazine*, Vol. 40, pp. 108-116, March 2002.

24. Jimenez, J., "Towards to 4G", in *http://research.ac.upc.es/conferencies/ITCSS/jimenez.pdf*, 2002.

25. "WINEGLASS - Wireless IP Network as a Generic Platform for Location Aware Service Support", in *http://domo-bili.cselt.it/WineGlass*, 2002.

26. Lu, W. W., "Compact Multidimensional Broadband Wireless: The Convergence of Wireless Mobile and Access", *IEEE Communications Magazine*, pp. 119-123, November 2000.

27. 3rd Generation Partnership Project, "IP Based Multimedia Services Framework

Report", *3GPP Technical Specification*, 23.228 V5.3.0.

28. Robles, T., A. Kadelka, H. Velayos, A. Lappetelainen, A. Kasler, H. Li, D. Mandato, J. Ojala, and B. Wegmann, "QoS Support for an All-IP System Beyond 3G", *IEEE Communications Magazine*, Vol. 39, pp. 64-72, August 2001.

29. Akyildiz, I. F., Y. Altunbasak, F. Fekri, and R. Sivakumar, "AdaptNet: An Adaptive Protocol Suite for the Next-generation Wireless Internet", *IEEE Communications Magazine*, Vol. 42, No. 3, pp. 128-136, March 2004.

30. Chou, C. T. and K. G. Shin, "Analysis of Adaptive Bandwidth Allocation in Wireless Networks with Multilevel Degradable Quality of Service", *IEEE Trans. Mobile Comp.*, Vol. 3, No. 1, pp. 5-17, January-March 2004.

31. Carneiro, G., J. Ruela, and M. Ricordo, "Cross-layer Design in 4G Wireless Terminals", *IEEE Wireless Commun.*, Vol. 11, No. 2, pp. 7-13, April 2004.

32. Tugcu, T., I. F. Akyildiz, and E. Ekici, "Location Management Framework for Next Generation Wireless Systems", *Proceedings of IEEE International Conference on Communications*, Vol. 7, pp. 3926-3931, June 2004.