Efficient Admission Control Schemes in Cellular IP Networks

T. Giang

(Master of Engineering)

2005

RMIT

Efficient Admission Control Schemes in Cellular IP Networks

A thesis submitted in fulfilment of the requirements for the degree of Master of Engineering

Triet Giang
B.Eng.

School of Electrical and Computer Engineering
Science, Engineering and Technology Portfolio
RMIT University
Sep 2005

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Triet Giang

30 September 2005

ACKNOWLEDGEMENT

I am greatly indebted to the individuals whose guidance, support and encouragement were essential to this thesis.

I would like to give my sincere thanks to my supervisor, Dr. Jidong Wang, for his enthusiasm, advice, guidance and support over the last two years. He is not only my supervisor but also a life mentor.

This thesis is dedicated to my family. I would like to thank my parents, sister and relatives for their encouragement and support; special thanks to my wife and son. I would like to thank my uncle's family who always welcomed me when I came in Melbourne.

My sincere thanks also go to Professor Neil Furlong (RMIT Pro-Vice Chancellor Research and Innovation), Professor Nguyen Xuan Thu, Dr David Fraser, Dr Michael Waters, Professor Robert Snow, Ms Roberta Abell, Dr Andrew Bean, Mr Nguyen Phuong Lam, Professor Andrew Scown (RMIT International University, Vietnam), Mr Mark Handley (Operations Manager in School of Electrical and Computer Engineering), Dr Dung Nguyen (Institute of Applied Mechanics, Vietnam), Ms Kathryn Thomas, Ms Lynda Gilbert (RMIT administrative staff) and other RMIT staff in Melbourne and Vietnam.

I am grateful to the Atlantic Philanthropies who sponsored this and other programs to develop Vietnam. I also thank Ms Mona El-Kadi (University of Trier, Germany) for her simulation codes and support when I first got stuck with my simulation.

Finally I thank my friends and students in RMIT International University Vietnam, who made this study program memorable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

2G	Second Generation of cellular networks
3G	Third Generation of cellular networks
3GPP	Third Generation Partnership Project
ABB	Actual Borrowable Bandwidth
AD	Adaptivity
AF	Assured Forwarding
AMPS	Advanced Mobile Phone System
AUC	Authentication Centre
BLT	Bandwidth Loss Tolerance
BSC	Base Station Controller
BTS	Base Transceiver Station
BU	Bandwidth Unit
CAC	Call Admission Control
CDG	CDMA Development Group
CDMA	Code Division Multiple Access
CIR	Carrier to Interference Ratio
DAMPS	Digital Advanced Mobile Phone System
DCA	Dynamic Channel Assignment
DiffServ	Differentiated Services
DRNC	Drift Radio Network Controller
DSCP	DiffServ Code Point
EF	Expedited Forwarding
EIR	Equipment Identity Register
FCA	Fixed Channel Assignment
FCFS	First Come First Served
FCLS	First Come Last Served
FDM	Frequency Division Multiplexing
FIFO	First In First Out
FILO	First In Last Out
FTP	File Transfer Protocol
GGSN	Gateway GPRS Support Node
GloMoSim	Global Mobile Information Systems Simulation

GPRS	General Packetised Radio Service
GSM	Global Service Mobile
HCA	Hybrid Channel Assignment
HLR	Home Location Register
HSCSD	High Speed Circuit-Switch Data
IMSI	International Mobile Subscriber Identity
IntServ	Integrated Service
IP	Internet Protocol
iTAS	Improved Threshold Access Sharing
ITU	International Telecommunication Union
LR	Loss Ratio
MAHO	Mobile-Assisted Handoff
Matlab	MATrix Laboratory
MCHO	Mobile-Controlled Handoff
ME	Mobile Equipment
MEX	Minimum EXpected
MMS	Multimedia Messaging Service
MPLS	MultiProtocol Label Switching
MSC	Mobile Switching Centre
NADC/USDC	North American/United States Digital Communication
NCHO	Network-Controlled Handoff
NMT	Nordic Mobile Telephone
NS-2	Network Simulator version 2
OMC	Operations & Maintenance Centre
OMNeT++	Objective Modular Network Testbed in C++
OPNET	OPtimised Network Engineering Tools
OSI	Open System Interconnection
PDA	Personal Data Assistant
PDC	Personal Data Communication
PDN	Public Data Network
PHB	Per Hop Behaviour
PSTN	Public Switch Telephone Network
QoS	Quality of Service
RR	Round Robin
RSVP	Resource Reservation Protocol

SGSN	Serving GPRS Support Node
SIM	Subscriber Information Module
SLA	Service Level Agreement
SMS	Short Messaging Service
SNR	Signal to Noise Ratio
SPT	Shortest Processing Time
SRNC	Serving Radio Network Controller
SRPT	Shortest Remaining Process Time
SS	Signal Strength
TACS	Total Access Communications System
TAS	Threshold Access Sharing
TC	Traffic Class
TCA	Traffic Conditioning Agreement
TCP	Transmission Control Protocol
TCU	Transcoder Unit
TDMA	Time Division Multiple Access
ToS	Type of Service
UE	User Equipment
UMTS	Universal Mobile Telecommunications Service
USIM	UMTS Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
UWCC	Universal Wireless Communications Consortium
VIP	Very Important Person
VLR	Visitor Location Register
VoIP	Voice over IP
WAP	Wireless Application Protocol
WB	Weight-Based
WCDMA-FDD	Wideband CDMA – Frequency Division Duplex
WCDMA-TDD	Wideband CDMA – Time Division Duplex

SUMMARY

The rapid growth of real-time multimedia applications over IP (Internet Protocol) networks has made the Quality of Service (QoS) a critical issue. One important factor affecting the QoS in the overall IP networks is the admission control in the fast expanding wireless IP networks. Due to the limitations of wireless bandwidth, wireless IP networks (cellular IP networks in particular) are generally considered to be the bottlenecks of the global IP networks. Admission control is to maintain the QoS level for the services admitted. It determines whether to admit or reject a new call request in the mobile cell based on the availability of the bandwidth. In this thesis, the term "call" is for general IP services including voice calls (VoIP) and the term "wireless IP" is used interchangeably with "cellular IP", which means "cellular or mobile networks supporting IP applications". In the wireless IP networks, apart from new calls, there are handoff (handover) calls which are calls moving from one cell to another. The general admission control includes the new call admission control and handoff call admission control. The desired admission control schemes should have the QoS maintained in specified levels and network resources (i.e. bandwidth in this case) are utilised efficiently. The study conducted in this thesis is on reviewing current admission control schemes and developing new schemes.

Threshold Access Sharing (TAS) scheme is one of the existing schemes with good performance on general call admission. Our work started with enhancing TAS. We have proposed an improved Threshold Access Sharing (iTAS) scheme with the simplified rate-based borrowing which is an adaptive mechanism. The iTAS aims to lower handoff call dropping probability and to maximise the resource utilisation. The scheme works at the cell level (i.e. it is applied at the base station), on the basis of reserving a fixed amount of bandwidth for handoff calls. Prioritised calls can be admitted by "borrowing" bandwidth from other ongoing calls. Our simulation has shown that the new scheme has outperformed the original TAS in terms of handoff prioritisation and handling, especially for bandwidth adaptive calls. However, in iTAS, the admission decision is made solely based on bandwidth related criteria. All calls of same class are assumed having similar behaviour. In the real situation, many factors can be referred in decision making of the admission control, especially the handoff call handling. We have proposed a novice scheme, which considered multiple criteria with different weights. The total weights are used to make a decision for a handoff. These criteria are hard to be modelled in the traditional admission models. Our simulated result has demonstrated that this scheme yields better performance in terms of handoff call

dropping compared with iTAS. We further expand the coverage of the admission control from a cell level to a system level in the hierarchical networks. A new admission control model was built, aiming to optimise bandwidth utilisation by separating the signalling channels and traffic channels in different tiers. In the new model, handoff calls are also prioritised using call classification and admission levels. Calls belonging to a certain class follow a pre-defined admission rule. The admission levels can be adjusted to suit the traffic situation in the system. Our simulated results show that this model works better than the normal 2-tier hierarchical networks in terms of handoff calls. The model settings are adjustable to reflect real situation. Finally we conclude our research and suggest some possible future work.

Chapter 1: INTRODUCTION

1.1 Background

The Internet is growing rapidly. One of the fastest growing sections is in wireless networks. Mobile (cellular) networks are evolving into IP (Internet Protocol) networks to overcome the limitations of traditional digital wireless networks and to expand its services. Wireless IP networks are more suitable for supporting the rapidly growing mobile data and multimedia applications. IP-based protocols, which are independent of the underlying access technologies, are better suited for supporting seamless services over heterogeneous radio technologies and for achieving integration with the fixed-line IP networks.

In the wireless network evolution, research works are concentrated on a few major fronts. Radio access technologies and Quality of Service are two of them. Radio access systems target the higher system capacity on bandwidth resources and the Quality-of-Service (QoS) focuses on the efficient utilisation of available resources to provide satisfactory services for diversified IP-based applications. This research is in line with IP QoS in cellular networks. The topic of the research is on admission control for new calls and handoff calls. The objective is to achieve low handoff call dropping rate for cellular networks supporting IP-based applications. In the following sections, the broad background of IP networks, IP based applications; cellular networks and the current status in call admission control are reviewed. The structure of the thesis is also outlined.

1.2 IP Based Applications

Together with Transmission Control Protocol (TCP), Internet Protocol (IP) forms the fundamental protocol suit of the Internet, which is the dominant media for today's information access, exchange, and distribution around the world [1]. IP applications running on wired networks have been migrated to wireless as a true solution for "in anywhere, doing anything and by anybody".

Broad applications are offered over IP networks. The typical ones are email, ftp, World Wide Web, e-commerce, online gaming and online entertainment. The early IP based applications are generally data based. With advent of services such as Voice over IP (VoIP), we have seen the shifting of traditional applications in telecommunication network towards IP networks. The boundary between telecommunication networks and computer networks became blur and

blur. The real-time multimedia applications have driven the two networks gradually merging together.

In today's mobile networks, voice is still the main service. However, in the 2.5G and 3G wireless networks, WAP, email access, web browsing, SMS (Short Messaging Service) – and MMS (Multimedia Messaging Service) are already in place. With the push for 3G and beyond 3G, all applications will be based on IP in the future. To summarise the possible applications over IP, Table 1-1 gives a better picture of the IP's impact in our daily life.

Table 1-1: Possible wireless applications

Sector	Services	Sector	Services
Business	mobile office virtual workgroups	Entertainment	games video clips virtual sightseeing gambling
Communication	video telephony video conference personal location	Finance	virtual banking online billing credit card
Community	emergency call administration services polling	Information	interactive shopping online newspaper online translation
Education	online library interactive distance learning	Telemactic	road transport logistics remote Control

Some applications require a large amount of bandwidth, while others can run smoothly with a small bandwidth share. Conventional online multimedia applications often occur between a user and a server. However applications running on wireless terminals do not always need a connection to a server, for example when two wireless terminals' users play games via Bluetooth.

IP technology's expansion into wireless networks has provided bigger space for potential IP based applications. The research on wireless access and mobility management, such as that proposed in IPv6 is paving the way on large scale of applications and services. We are going to see the explosion of versatile wireless IP based applications including the popular multimedia applications will happen in the near future. This depends on the Internet and wireless access systems with better quality and usability [2].

For any application, from users' point of view, it does not matter which network technology the application runs on as long as its performance is satisfactory. Therefore in the evolution to IP technology and wireless IP networks, the quality of service (QoS) is an important issue which needs to be addressed. We will look at the importance of QoS on IP-based applications in the next chapter.

Wireless networks are implemented on a number of different topologies. In this thesis, we focus on the cellular network, where handoff remains one of the concerns.

1.3 Cellular networks

1.3.1 The evolution of cellular networks

The first generation of wireless telephones are radio telephones. This system has a number of base stations which are connected by wired links. Each base station covers a large area. The capacity is low as the usable radio spectrum allows only a number of ongoing calls in an area.

After digital switches were developed in 1970s and the concept of cellular networks was introduced, cellular networks became feasible. Since then, they have gone through three generations with different technologies: analogue voice, digital voice and digital voice plus data [3].

The first generation (1G) started in 1979 in Scandinavia with Nordic Mobile Telephone (NMT), followed by Advanced Mobile Phone System (AMPS) in the United States, Total Access Communications System (TACS) in Europe and MCS-L1 in Japan. 1G systems used Frequency Division Multiplexing (FDM) with the spectrum of 824 - 849 MHz for transmission and 869-894 MHz for reception, divided into about 45 channels (30 kHz each) per cell with cell diameters of 10-20 km. Theoretically the system could support up to 2400-bps data communications; however they were mostly used for voice communication, at the rate of 1200 bps. Transmission type was analogue with a large amount of overhead signals.

The second generation (2G) was developed from the digital technology to cope with the needs of further security and bandwidth usage. The most popular standards are Global Service Mobile (GSM) in Europe and some countries in Asia, Personal Data Communication (PDC) in Japan, North American/United States Digital Communication (NADC/USDC), Digital Advanced Mobile Phone System (DAMPS), and cdmaOne (using Code Division Multiple Access -CDMA) in the United States and other Asian countries [4].

They were based on Time Division Multiple Access (TDMA) with up to 9.6 Kbps data transmission, the same as a conventional modem. GSM became the most widespread so far and other standards (e.g. Wireless Application Protocol (WAP) and i-Mode by NTT DoCoMo in Japan) were developed to provide easier access and more services, especially access to the Internet. Users' services include ticket booking, news reading, mobile banking, game playing, road traffic monitoring and so on.

In the fixed networks, data traffic grows very fast whereas voice traffic does not. The Internet becomes a popular information source. The trend to use Internet service on mobile terminals

grows stronger. The networks now need to support more bandwidth with less interruption. With additional services to 2G systems and as a transition generation to 3G, 2.5G systems provide higher data rate with high speed circuit-switch data (HSCSD) at 28.8 Kbps, general packetised radio service (GPRS) at 100 Kbps, and enhanced data rates for GSM evolution (EDGE) at 300 Kbps. HSCSD users are charged in the same way as traditional GSM users, whereas GPRS (packet-switch oriented) users are charged per data amount.

The tendency to multimedia applications on wireless terminals boosts another evolution step to the latest generation, 3G. The International Telecommunication Union (ITU) proposed the IMT-2000 scheme, which is deployed as Universal Mobile Telecommunications Service (UMTS) in Europe. Using wideband CDMA, there are four main technologies: American cdma2000 by CDMA Development Group (CDG), Wideband CDMA – time division duplex (WCDMA-TDD) and Wideband CDMA – frequency division duplex (WCDMA-FDD) by Third Generation Partnership Project (3GPP) and UWC-136HS by the Universal Wireless Communications Consortium (UWCC). Transmission rate is varied from 144 Kbps to 2 Mbps, depending on the speed of the terminal.

Some technical aspects of different network generation are compared in Table 1-2.

Table 1-2: Comparison of technical specifications in different cellular generations

	1G	**2G, 2.5G**	**3G**
Examples	NMT, AMPS, TACS, MCS-L1	GSM, D-AMPS, PDC, cdmaOne, EDGE, GPRS	UMTS, CDMA2000
Uplink channels (MHz)	824-849	1850-1910	1920 - 1980 (FDD)
Downlink channels (MHz)	869-894	1930-1990	2110 – 2170 (FDD)
Multiplexing technologies	FDM	FDM, TDM , CDM	FDM and TDM
Data rate	1200 bps	8,000 bps (D-AMPS) 9,600 bps (GSM)	144 kbps- 2 Mbps
Technology	Analogue, voice	Digital, voice, data	Digital, voice, data
Handoff	Hard	Hard	Soft

In the next section we will discuss the hierarchical structure, the trend of cellular networks.

1.3.2 Hierarchical cellular networks

To increase the wireless system capacity, while the number of channels per cell is fixed, it's necessary to increase the number of cells or reduce the cell size. The cell capacity is known to be inversely proportional to the square of the cell radius [5]. However the small cell size leads to more frequent movement between cells, hence more handoffs are required. Every time a handoff occurs, it is possible that the call may be dropped or some data may be lost.

To maximise the number of calls in a cell while minimise the network control associated with handoff, the overlaid cellular architecture is introduced. Smaller cells are built inside large

cells. There are two types of overlaid networks: homogeneous overlaid networks (or hierarchical cellular networks: a network divided into many layers with the same random access technique) and heterogeneous overlaid networks (with many layers, each has different random access technique). We only consider the first type.

The most common arrangement is the macrocell/ microcell network with two layers and two different frequency bands for macrocells and microcells [6]. The larger cells are the macrocells (with the cell radius from 1 to 10 km), the smaller ones are microcells (with the cell radius from 0.2 to 1 km) [7]. At any time, a mobile is in the coverage of a macrocell and a microcell. Macrocells use a lower frequency band due to better propagation characteristics of microwave through the atmosphere [8]. Microcells use higher frequency bands. This arrangement provides better spatial reuse of microcell frequencies, hence more bandwidth, better utilisation and inherent load balancing [9]. In GSM networks, microcells can operate at 1800MHz while macrocells can operate at 900MHz. Cells are controlled by different base stations. A mobile terminal attaches itself to a microcell if it does not move or its speed is considered low. In the other hand if the terminal is moving in a high speed, it is connected to a macrocell. The use of microcells satisfies the first goal (maximising the network utilisation) and that of macrocells meets the second goal (minimising the network control associated with handoff). Low-mobility or slow-moving terminals experience handoff at a microcell level whereas high-mobility or fast-moving terminals at a macrocell level. This results in an acceptable handoff rate for all calls. It is possible to have another lower layer for the picocells (with the cell radius from 10 – 200 m) mostly for indoor wireless communications or a higher layer for satellite cells. This structure is referred to as hierarchical cellular multi-hop networks [10], illustrated in Figure 1-1.

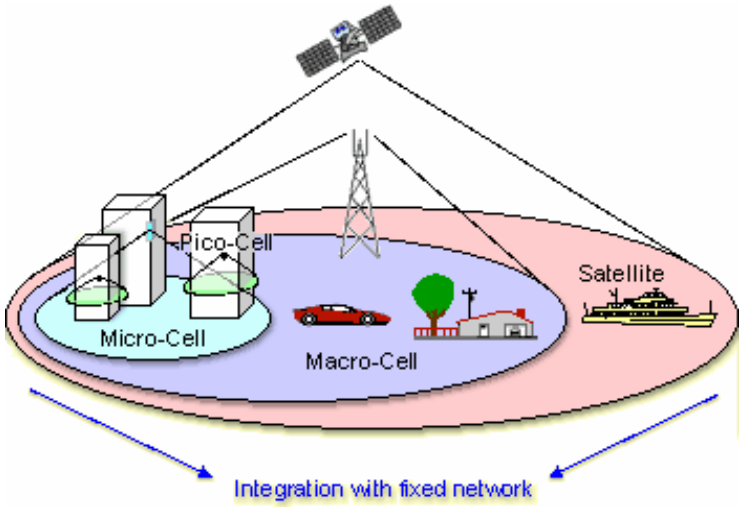


Figure 1-1: Hierarchical cell structure to offer global radio coverage [11]

In this structure, calls are handed off vertically (between cell layers) or horizontally (between cells in the same level). Horizontal handoffs often occur when the mobile terminal changes its physical speed.

To ensure satisfactory QoS requirements to cellular users, it is essential to apply an admission control scheme. The importance increases if IP-based applications are running on the wireless terminals.

1.4 Call Admission Control (CAC) in cellular networks

In the large IP networks, the access points in the wireless IP networks are considered as the bottleneck due to the limited wireless bandwidth resources. Admission control is aimed to use the precious bandwidth efficiently. The concept of Call Admission Control (CAC), which exists in the traditional wireless communication networks, still applies to wireless IP networks. The main difference is that in traditional communication networks, the service is unique, i.e. voice. In IP networks, the services become versatile.

In general, call admission control processes at an access point are the channel allocation and admission algorithm. The channel allocation controls the bandwidth usage in a cell to ensure that the resource is shared fairly. The admission algorithm prioritises calls. The mutual task of both processes is to define how calls are admitted.

The channel allocation organises the bandwidth in a cell so that it is most efficiently used. It must work with the admission algorithm to balance the bandwidth usage between different services. Most people would rather be rejected at the service request than be dropped in the middle of the call. Hence handoff calls should have priority to use bandwidth over new calls. This is normally done by reserving a portion of bandwidth for high priority calls. Estimating how much bandwidth should be reserved for high priority calls is the task of this control process.

The admission algorithm decides to accept or reject a call. The algorithm is based on a number of criteria, such as: the current load, the predicted load, the traffic class of the call and the call's type (a new or handoff call).

In cellular networks, handoff is the critical point where most of the forced terminated calls occur. A common situation occurs when a call is admitted in one cell but in a handoff it is forcefully terminated because of insufficient bandwidth in the adjacent cell. We would like to keep the probability of such termination to a minimum. At the same time, we would maintain the QoS requirement in other existing calls.

1.5 Motivation and research questions

There is no doubt that real-time IP multimedia applications are the driven forces behind the push for IP QoS. In case of the wireless IP networks, Call Admission Control (i.e. channel allocation, handoff control and admission control) is a unique component which has big impact on the overall IP QoS. Imagine, if the networks over admit services, their QoS level would be hard to meet due to limited resource. If a network is under admitting services, the network resources will be wasted. This research aims to review up-to-date proposed solutions on Call Admission Control. From those solutions, we draw a better admission control solution with reduced handoff call dropping probability. Furthermore, we derive a novel admission control to take all environmental criteria into consideration when making an admission decision. Finally we propose an admission control model for hierarchical networks.

We limit ourselves to the admission control in wireless networks with cellular and hierarchical structure. The applications studied are IP-based and adaptive bandwidth.

1.6 Thesis outline

This thesis is organised into six chapters. In this chapter, i.e. Chapter 1, wireless networks, their evolution and call admission control, as one of their QoS problems have been presented. The main research topics on the admission control, channel allocation and handoff control are identified and briefly discussed. The rest of the thesis is organized as in the following. Chapter 2 reviews the admission control background and most common admission control schemes up-to-date, in details. In chapter, 3, some outstanding schemes are analysed. A so called iTAS scheme is proposed to improve the handoff performance. In chapter 4, we initiate another admission scheme considering multiple criteria in making admission decisions. Chapter 5 targets hierarchical networks with a complete admission model proposed, prioritision of handoff calls is also adopted in the new model. In the last chapter, we summarise all findings and discussed possible further improvement on the new schemes.

Chapter 2: ADMISSION CONTROL – Theory and Practice

In this chapter, the principle and practice of admission control in traditional telecommunication networks as well as the cellular networks are reviewed. Our review will start with discussion in Quality of Service (QoS), the big umbrella, under which the admission control is positioned. In the last sections of this chapter, a various admission control schemes are presented and summarised.

2.1 Review of QoS on IP-based applications

QoS has become an important factor as the traditional telephone network and the IP network are gradually merging. Invented by Alexander Graham Bell in 1876, the telephone networks were designed for realtime voice communication, while the IP networks were designed to carry data.

The former networks use circuit-switching technology, in which the call occupies the whole line during the call. There are two main measurements for the performance of a telephone network: the call blocking probability, which is the probability that a call attempt is blocked due to insufficient lines (1% is normally the ideal value); and the voice quality of the call after the connection is established. The voice quality is assessed by distortion, delay, noise and echo.

The IP networks are initially designed for data (i.e. non-realtime) services. Data can be stored, forwarded or retransmitted with delay. Data transmission is packet based. Data are transferred along the path from its source to its destination, being forwarded node by node. When realtime applications, such as voice, are moved on to IP networks, the IP networks' packet loss, packet delay and packet jittering has to meet certain requirement to be able to provide satisfactory services.

In the context of general telecommunication networks, QoS includes many components such as the quality of voice, network support, maintenance, billing and general services. However, in VoIP, all components of the traditional QoS are not significant issues except the quality of voice, which is directly affected by the packet jittering, packet latency and packet loss in the IP networks. As IP networks were initially designed for pure data services and applications, for which, jittering and delay are acceptable. The Transmission Control Protocol (TCP) on top of IP can solve the problem of packet loss. However, applications such as realtime multimedia services would require the networks to guarantee the quality of a service by setting minimum bandwidth committed to a connection, and limits on data loss and delay. The current Internet

is providing the best effort service, i.e. all IP packets are treated equally on the basis of first come, first serve policy. It works well with non-realtime applications such as email, ftp, web browsing, but it fails to deliver satisfactory QoS for real-time applications such as VoIP, IP-based video conferences. IP QoS remains a hot issue in the converged telecommunication industry.

Different applications may have different QoS requirements. The IP networks should be able to provide services with different QoS level. From the network's point of views, QoS means the ability to provide a service with pre-defined standard. From the user's point of view, QoS is the quality that the person receives from the network for the service he/she subscribed.

The factors affecting the QoS include packet loss, jitter and delay. An acceptable toll quality could tolerate certain levels of data loss, jittering and delay. Transmission errors (noise and other disturbances in the network) may cause data loss. Applications used by human beings (e.g. voice) can tolerate accidental short disturbances; whereas, in computer communications, the application can not afford too many errors in transmission. If a data frame has a few erroneous bits, it may be destroyed and retransmitted. It is wasteful but necessary for the precise operation of computers. Most systems are designed to be able to retransmit except voice systems.

The ability to provide a service at a certain standard includes the allocation of bandwidth to a request and the ensuring of quality during the call duration. The bandwidth requirement can be further divided into the minimum acceptable bandwidth (the level which the application operates with acceptable performance) and the desired bandwidth (the level which the application wishes to operate at). A normal voice call does not need as much bandwidth as a call running a video-on-demand application. High bandwidth consumed applications may need an adaptive bandwidth mechanism to balance between their operation and fair resource management in the system.

QoS on IP-based applications in wireless platform

In cellular networks, QoS concepts expand to cell loss ratio, cell delay variation, cell transfer delay, handoff dropping rate, call blocking rate and so on. QoS in wireless communication is more challenging than that in wired line networks because of the limited bandwidth, location-dependence and user mobility. The bandwidth limitation requires an efficient bandwidth control. The location-dependence needs a proper network design while user mobility invokes many handoffs.

QoS requirements are to guarantee the quality of ongoing calls and to keep a low new call blocking probability. The quality of service for ongoing calls includes stable bandwidth allowance and low handoff dropping probability. From users' point of view, once connected to the network, their connections should be guaranteed with the required QoS. When the bandwidth is all allocated, new calls will be blocked. Therefore, to keep their service level up, network operators need to provide sufficient resources which mean more investment to them. As operators are focusing on revenue and profit, they need to compromise between the cost and the QoS commitments. However, well designed admission control and handoff control can help operators to maximise the efficiency of network resource utilisation. From another angle, the optimal admission control and handoff control can help to provide users with maximum QoS.

To deliver a good QoS solution for general IP networks, many network components and protocols need to be involved. IP QoS schemes outline the big picture of the QoS framework. Its implementation relies on the concrete protocols, algorithms deployed at different layers and components. Listed below are three main QoS schemes for IP technology: Integrated Services (IntServ), Label Switching and MultiProtocol Label Switching (MPLS) and Differentiated Services (DiffServ). We will look in depth DiffServ concept because we will apply a similar idea in our admission model.

Integrated Service (IntServ)

The formal name is "flow-based algorithms" applies resource reservation and admission control mechanisms to provide a guaranteed performance level [12]. It explicitly reserves flows for each traffic source. The scheme works well in a small network and is not scalable. This limitation is because all routers have to update its routing table information for each flow. The scheme introduces overheads when advanced setting up for each flow. It is suitable to audio and video applications. The main protocol in IntServ is the Resource Reservation Protocol (RSVP) to reserve resources. Other protocols are used to send data.

Label Switching and MultiProtocol Label Switching (MPLS)

In this scheme, packets are labelled. Routers use information on labels rather than the destination address [4]. MPLS is the standardised protocol in label switching. It is not in layer 2 nor 3 and its header is added in front of a TCP/IP header. Individual connections do not require individual setup. When a packet arrives, the first router will work with the next router to generate a label for the flow. This is just one of the techniques to route traffic.

Differentiated Service (DiffServ)

Opposed to "flow-based" in IntServ, this is referred to as "class-based". DiffServ is used to overcome the scalability problem [13]. In DiffServ, network resources (including bandwidth) are allocated per traffic class, rather than per flow as in IntServ. It is hard to absolutely differentiate traffic, most of the time, the classification is relative. To guarantee that QoS requirements are met, admission control may be necessary at the access points of DiffServ networks. Routers only need to know the information to route that bunch i.e. less routing information than in IntServ. It does not need advanced setup per-flow admission control or resource reservation.

Operators can define their own traffic classes and assign different handling mechanisms. A general structure of a DiffServ network can be found in Figure 2-1. Abbreviations and concepts are discussed later.

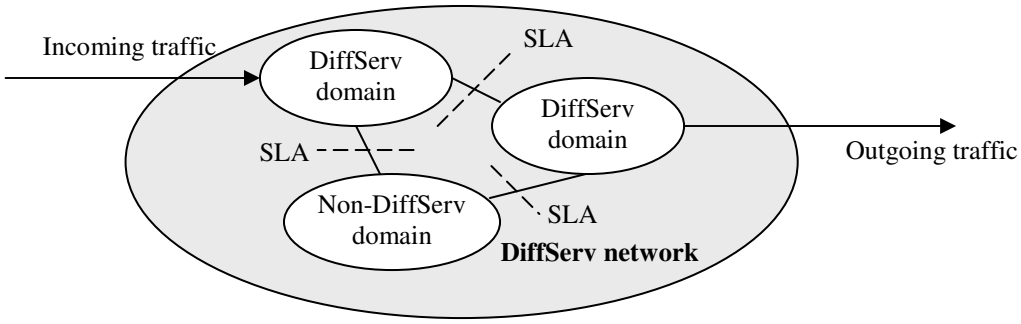


Figure 2-1: DiffServ network structure

With the inputs from its customer, the operator defines a mutual Service Level Agreement (SLA). This agreement clarifies the Traffic Conditioning Agreement (TCA) with the performance measurements such as delay, total traffic throughput, service level, traffic profile, shaping and marking of the traffic. An SLA can be static or dynamic depending on the customer service requirement. Traffic from one DiffServ domain to another must conform to the SLA.

At the packet level, there are not any additional fields defined for DiffServ use. To mark packets, DiffServ used the 8-bit Type of Service (ToS) field in IPv4 header or the Traffic Class (TC) field in IPv6 header. If the node does not support DiffServ, those fields keep their original functions. Two last bits of the field are reserved, only six bits of the field (known as DiffServ Code Point – DSCP) are used for packet marking. This 6-bit field yields 64 possible DSCP, which are grouped to 3 pools by IETF RFC 2474. Pool 1 has 32 DSCPs, available for standard actions. Pool 2 and 3 each has 16 DSCPs, reserved for local use or experimental, with an exception that Pool 3 DSCP may carry standard actions.

In a DiffServ network, packets sent from one node to another must follow a set of disciplines, known as Per Hop Behaviour (PHB). PHBs are router dependent, so the priority level definition is relatively local to that particular router. Each manufacturer has different ways to implement PHBs, but in general, there are two types of PHBs: Expedited Forwarding (EF) and Assured Forwarding (AF). In Expedited forwarding, some bandwidth is reserved for expedited traffic. To expedited traffic packets, the link is always free. Regular packets are queued and sent as normal. Assured forwarding scheme has 12 classes. Packets are classified, marked and sharp according to their type. The technical details of PHBs are not in the scope of this research. More can be found in [14, 15].

A DiffServ network consists of a number of DiffServ domains. At each domain, a bandwidth broker is applied to manage the bandwidth. The broker has traffic conditioners such as meters, markers, shapers and dropper. A meter measures the traffic to check if it complies with the agreement. A mark sets the DSCP in the packets. A shaper delays traffic to suit its traffic profile. A dropper discards packets that violate its profile.

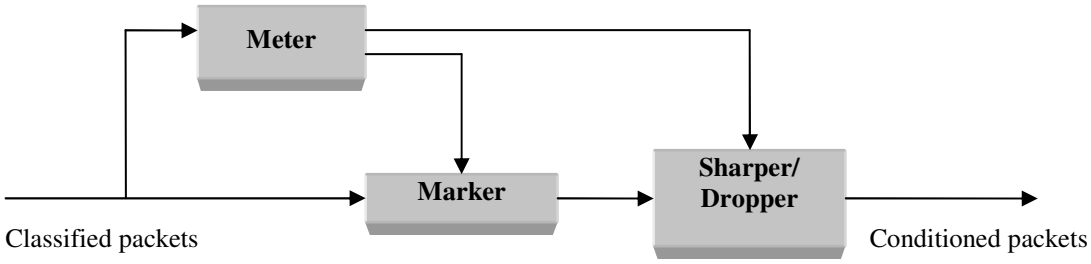


Figure 2-2: Functional diagram of a bandwidth broker [16]

In this research, we focus on the call admission control at the wireless access points, particularly from cellular phones to the network. At these points, we apply the DiffServ classification concepts to group traffic. In the backbone network, we assume that DiffServ is used to route traffic to the other end (wireless access point). We talk about our traffic classes in the following section.

Common traffic classifications

In QoS study for wireless IP networks, the traffic classification is widely accepted. IP multimedia applications can be divided into two groups according to their transmission type: symmetric or asymmetric. Symmetric applications have the same bandwidth requirement for sending and receiving directions, e.g. messaging or chatting applications. On the contrary, asymmetric applications have different bandwidth requirement in each transmission direction, e.g. FTP downloads use more bandwidth in the downstream than in the upstream. In this case the upstream transmits acknowledgements and control information only.

In general, applications can be grouped by their QoS requirements: realtime and non-realtime. Realtime applications are sensitive in delay but quite tolerant to data loss. For example, in a video conferencing call, the quality of the images can be a little distorted but the users should not experience too much delay as well as echoes. Non-realtime applications in the other hand can afford relatively longed delay but less tolerant to data loss.

It is probably more common to classify applications into their QoS requirement classes. As an example, we can take a look at the traffic classes in UMTS/ WCDMA systems: conversational, streaming, interactive and background class. Each class is described as below:

Conversational class: this is for conversational applications which involve person-to-person communication in realtime traffic with low delay (much less than 1 second) and stringent requirements, e.g. voice calls, video conferencing calls, interactive video games. Apart from low delay, low jitter, little echo and clarity are the basic QoS for voice calls. With multimedia applications other control information such as timing of the different media streams [17]. The performance is not seriously affected by reasonable loss, especially in voice calls. Applications in this class are symmetric.

Streaming class: it is also for realtime traffic. The information is sent to users but the communication does not involve human interaction. It is usually unidirectional with low delay requirement (around 1 second), low jitter and media synchronisation, e.g. video-on-demand, news streams, online radios. Applications are asymmetric with more bandwidth required in the downstream. Traffic in this class can tolerate reasonable amount of errors. Buffers can be used to reduce the delay variation [17].

Interactive class: applications in this class are quite error-rate-sensitive and require response within a certain time (less than 10 seconds), e.g. Internet browsing, server access. Although this class is not delay sensitive, applications have a timeout period to which a response is expected.

Background class: data transmission in this class works on best effort. The applications do not need an immediate response. This is the lowest QoS class, so the receiver does not expect to have data within a certain time (delay can be greater than 10 seconds), e.g. emails, FTP services. The error rate is expected to be as low as possible.

We have seen four UMTS traffic classes defined according to the service requirement. In most research on admission control, services are often categorised into two classes. This widely adopted assumption is defined as: Class I and Class II. Class I is realtime traffic with high priority and delay-sensitiveness, e.g. video, audio conferences or multimedia streaming

applications. This covers both conversational and streaming classes in UMTS. Class II is non-realtime traffic, e.g. e-mail, web browsing or ftp services. Interactive and background UMTS QoS class are included here. We will use this classification through out this thesis.

In the next section we will look at some latest admission control schemes, their characteristics, advantages and disadvantages.

2.2 Current proposed call admission control schemes

In cellular networks, the coverage of a geographical area is divided into cells to implement frequency reuse and increase the system bandwidth. Each cell is assigned an amount of bandwidth, which is divided to a number of channels. Because this number is limited per cell, our aim is to utilise the bandwidth by controlling the admission.

Call Admission Control has the following common purposes:

- Maintain stable QoS for ongoing calls by:
    - Guaranteeing the transmission rate
    - Minimising the handoff call dropping probability
    - Treating calls with different traffic classes in different ways
    - Keeping a low latency and disruption time during the handoff
    - Minimising the impact on the overall network, especially neighbouring cells
- Maximise revenue by:
    - Providing value-added services by giving priority to certain customers
    - Minimising the new call blocking probability
    - Sharing resources in a fair way

When a user makes a new call or moves from one cell to another cell, a call request is generated. The admission control involved at an access point is essential because it can prioritise the call and controls the bandwidth usage in a cell to ensure that the resource is shared fairly.

Before discussing the admission control in details, we will look at some channel assignment concepts and schemes. The operation ideas in these schemes are also applied in admission control schemes.

2.2.1 Channel assignment schemes

The process of assigning a set of frequencies to a cell is known as “channel assignment”. We introduce the concept of “channel allocation”, which is the process of provisioning channels in a cell to different types of traffic. The two processes are similar in term of dividing channels. Channel assignment assigns available channels in a coverage area to a number of cells, while channel allocation assigns those channels of that cell to different traffic types.

Channel assignment schemes are implemented centrally or distributed. In the centralised schemes, the assignment is done by a central controller. In the distributed schemes, the serving base station or the mobile decides which channel to use. If the serving base station makes the decision, it has to maintain information about the available channels in its neighbour cells. If the mobile makes the decision, it solely uses its CIR (Carrier to Interference Ratio) measurements.

We will review a few main channel assignment schemes in the following. They fall into three main categories: Fixed Channel Assignment (FCA) and Dynamic Channel Assignment (DCA) and Hybrid Channel Assignment (HCA). We will also look at the analogous versions in channel allocation schemes.

COMMON CHANNEL ASSIGNMENT SCHEMES

2.2.1.1 Fixed Channel Assignment (FCA)

In FCA, a cell is permanently assigned a set of frequencies. These schemes are very simple but because the numbers are fixed, the schemes can not adjust to the change of traffic conditions as well as user distribution in the area.

In simple FCA, the schemes allocate all cells the same number of channels. It works well if the traffic is uniformly distributed. However the traffic is not always uniform and this causes poor channel utilisation i.e. some cells are highly used while others have many free channels.

To improve the utilisation caused by non-uniform traffic, Zhang [18, 19] and Oh [18, 19] introduced non-uniform channel assignment schemes. Each cell has a profile keeping the local traffic history. Cells with high load (according to the profile) are allocated with more channels. The scheme [18] is referred to as “non-uniform compact pattern assignment”, which allocates channels to cells so that the average blocking probability of the system is the smallest. The allocation pattern is the way in which channels are allocated to co-channel cells. The allocation pattern with minimum physical distance between cells is the compact

allocation pattern. The scheme finds the compatible compact patterns that give the smallest average blocking probability of the system.

Using a different approach, Engel [20, 21] and Andreson [20, 21] proposed to statically borrow free channels from a cell and give to high load cells. Cells must be sufficiently physically apart to avoid CIR. At the start-up phase, the system periodically permanently assigns different numbers of channels to different cells based on the history load of the cell. These numbers can change periodically or predicatively to adapt to the traffic situation.

In contrary to static borrowing is channel borrowing. The number of borrowed channels in channel borrowing schemes changes accordingly because channels are borrowed temporarily. A borrowed channel will be returned to the original cell once the cell completes. Current channel borrowing schemes use different methods to select a free channel to borrow.

Channel borrowing schemes

Two types of channel borrowing are simple and hybrid. Simple schemes allow any channel in a cell to be borrowed while hybrid schemes divides its channels into two sets: the standard set is allocated to all cells; the borrowable set is available for borrowing. Note that an acceptor cell is borrowing a channel and a donor cell is lending a channel. When a channel is borrowed, it can not be used in neighbour cells of the acceptor, this is called “channel locking”.

Simple channel borrowing schemes [20-23]: a number of channels is assigned to a cell, after all channels are used, the call attempts to borrow a free channel from a neighbour cell. Lower

blocking probability under light and moderate traffic is achieved but it can raise the interference level in the donor cells, which affect the future calls in those [24]. Under heavy load, the schemes cause so much channel locking that the channel utilisation is lowered, hence the blocking probability is higher [25]. The algorithm of how a channel is selected to be borrowed can prevent channel locking. This is also the difference in the proposed schemes.

- Borrow first available [21]: all channels are divided into sets, each set has a sequential number. Then a set is assigned to many cells at a certain reuse distance. When there is a need to borrow a channel, the system simply searches in the sequence for the first available channel.
- Borrow from the richest: the cell borrows from the neighbour cell with the most channels available for borrowing [21]. This scheme does not take care of channel locking into account.

- Basic algorithm: similar to the Borrow from the richest but when selecting a channel, it tries to reduce the effect of channel locking [20, 21] to minimum.
- Basic algorithm with reassignment: improved from the Basic algorithm, this scheme will move a call using a borrowed channel to its normal channel (assigned to its residing cell) whenever one is free [20].

Surprisingly the *borrow-first-available* scheme gives an approximately good result as other more complex schemes. The complexity of the *borrow-first-available* scheme is also less than others. Note that the complexity of the algorithms depends on the network load as well. These are verified by simulation in [21].

Hybrid channel borrowing schemes:

- Simple hybrid channel borrowing strategy [18, 22, 26]: this just applies straight forward the generic idea of the hybrid scheme. The number of channels should be in the standard set and the borrowable set is estimated from the information about the traffic situation. However this can be dynamically adjusted periodically or predicatively [22].
- Borrowing with channel ordering [22, 23]: similar to the Simple hybrid strategy with dynamic adjustment of the ratio of the number of channels in the standard set and those in the borrowable set. Channels in a cell are sorted such that the first channel is the most preferable for local use and the last channel has highest priority for being borrowed.
- Borrowing with directional channel locking [18]: in the previous scheme, the number of available channels is low because a channel is borrowable only if it is simultaneously available in three nearby co-channel cells. Once a channel is borrowed, it can not be used in those nearby cells due to channel locking. In this scheme, the channel locking only applies to the directions affected from the borrowing i.e. the number of available channels increases. At the same time, the scheme can shift a call in a borrowed channel back to a normal channel or to a borrowed channel which makes the future borrowing more effective. Comparing with others, this scheme has the lowest blocking probability.
- Sharing with bias [27, 28]: in this scheme, refer to Figure 2-3, a cell is divided into three sectors. Calls originated in sector X can borrow channels in its two adjacent neighbour (shaded) cells. In each cell, channels are divided into two subsets as in the Simple hybrid channel borrowing strategy.

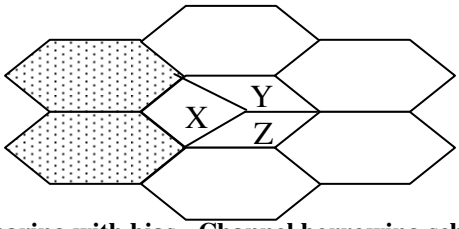


Figure 2-3: Sharing with bias - Channel borrowing scheme [29]

- Channel assignment with borrowing and reassignment [20]: channels are borrowed as they make the least effect on the blocking probability in neighbour cells. At the same time, it deploys the channel reassignment mechanism to balance the load.
- Ordered channel assignment scheme with rearrangement [30]: this scheme combines the Channel assignment with borrowing and reassignment with the Borrowing with channel ordering. All channels are numbered in the order of which it will cause the least effect on the blocking probability. Channels are selected in order until there is none in the cell available, then the cell attempts to borrow a borrowable channel in neighbour cells. The system keeps a list of available channels. When a channel is borrowed, the status of the system is updated. The system tries to shift calls on borrowed channels to normal channels. This scheme performs better than the Sharing with bias but it has more overhead signals.

Table 2-1: Comparison between fixed channel assignment schemes [29]

Scheme	Complexity	Performance
Simple FCA	Low	Better than dynamic and hybrid borrowing in heavy traffic
Static borrowing	Low-moderate	Better than FCA
Simple channel borrowing	Moderate-high	Better than FCA and static borrowing in light and moderate traffic
Hybrid channel borrowing	Moderate	Better than FCA in light and moderate traffic Better than simple channel borrowing in heavy loads

2.2.1.2 Dynamic Channel Assignment (DCA)

DCA schemes solve the inflexible operation of FCA. A channel pool keeps all channels and new calls take resource from the pool if CIR is minimal. After the call completes, the channel is returned to the pool. The algorithms are more complex. DCA are not as efficient as FCA under high load situations because excessive channel switching at the switching centre.

Different DCA schemes have difference cost functions. The DCA cost functions are to evaluate the cost of using a channel and choose the minimum cost given a satisfactory CIR condition [29]. There are many factors in the cost functions: reuse distance, frequency of the

channel, average blocking probability of the system and the future blocking probability in the neighbour cells [25].

DCA schemes can be classified into centralised and distributed schemes. In centralised schemes, the channel from the pool is selected and assigned to a call by a central controller. In the other schemes, the selection and assignment are done by the base station.

Centralised DCA schemes: the efficiency of channel assignment in these schemes are high with the trade-off of high computational load in the central controller [31-33].

- First available [34]: the first available channel satisfying conditions will be used. It has low complexity and can handle better than FCA schemes in low and moderate traffic situation.
- Local optimised dynamic assignment [18, 22]: this scheme uses only local information about the future blocking probability

The following schemes attempt to shorten the local channel reuse distance to increase the reuse of a channel in the whole system. In general, these schemes carry more traffic than the previous.

- Nearest neighbour: the cell selects a free channel in the nearest cell, which is at a reuse distance from itself [34]. It has the lowest blocking probability in light traffic, compared with other centralised DCA schemes.
- Nearest neighbour + 1: similar to *nearest neighbour* at the distance of reuse distance plus 1. The forced call termination rate and possibility of channel changing are low because the mobile is likely to keep the same channel in the neighbour cell [34]. However the blocking probability is higher than other schemes [35].
- 1-clique**:** different from the previous four schemes that it uses global channel reuse optimisation. The scheme applies graph theory to select the best possible channel. It has low blocking probability but high computational load [36].
- Selection with maximum usage on the reuse ring: a channel which is in use in most cell (maximum usage) will be selected [34].
- Mean square: the cell selects the channel that minimises the mean square of the distance among the cells using the same channel [29].

Distributed DCA schemes: the algorithm in the base station is simpler than of the centralised DCA schemes. Current schemes can be categorised into two types: cell-based or signal strength measurements. Cell-based distributed schemes use information about free channels in

the neighbour cells. Information in each cell is updated by exchanging with other cells [37-40]. These schemes "provides near-optimum channel assignment at the expense of excessive exchange of status information between base stations, especially under heavy traffic load" [29]. The other types of schemes are based on signal strength [41] measured locally by the base station. The base station does not care about situations in other cells; hence the system controls itself to maintain a good performance. The schemes support fast realtime processing but the probability of channel interference on ongoing calls in neighbour cells is higher, possible resulting in service interruption or an instable system.

Cell-based DCA schemes:

- Moving direction [39, 40]: the scheme uses the information on moving directions of the mobile terminal to reserve channels along the direction. The probability of changing channels and forced call termination is lowered. The strategy is suitable for mobile terminals moving in the same direction with approximately the same speed, e.g. people on cars on a highway.
- Locally packing distributed DCA [38]: the channel is selected based on a database called the augmented channel occupancy (ACO) matrix. The matrix is updated by exchanging information with interfering cells.
- Locally packing distributed DCA with Adjacent Channel Interference constraint [34]: similar to the Locally packing distributed DCA with the consideration of Adjacent Channel Interference. In practice, most proposed schemes are based on the co-channel interference and ignore the adjacent channel interference.

Signal-strength measurement based DCA schemes:

- Minimum signal-to-noise interference ratio [41]: the search criterion is the signal-to-noise interference ratio on the uplink.
- Sequential Channel Search [41]: both the terminal and the base station use the same order to check for the first channel with the strongest signal strength.
- Dynamic Channel Selection [42]: each mobile terminal measures the interference in all channels. The mobile chooses a base station to connect to. The decision is made upon a number of criteria, such as received signal power, channel availability and co-channel interference.
- Channel segregation [43, 44]: a cell chooses a channel with satisfactory interference level by scanning all channels. There is no fixed channel assignment to any specific

cell. The scanning order is done for each cell independently. To determine if a channel is free, the received power level is compared to a threshold. This scheme is proved to be efficient with lower blocking probability and less number of handoffs.

Below is the comparison between FCA and DCA schemes. One of the main tradeoffs is complexity over flexibility.

Table 2-2: Comparison between FCA and DCA

	FCA	**DCA**
Performance under certain traffic situations	Better under heavy traffic	Better under light to moderate traffic
Probability of forced call termination	High	Low to moderate
Cell size	Suitable for macrocell	Suitable for microcell
Flexibility	Low	High
Computational effort	Low	High
Call setup delay	Low	Moderate to high
Implementation complexity	Low	Moderate to high
Frequency planning	Complex, labour-intensive	No
Signalling load	Low	Moderate to high
Control type	Centralised	Centralised, decentralised, distributed control, depending on the scheme.

2.2.1.3 Hybrid Channel Assignment (HCA)

These schemes combine FCA and DCA. It can adapt to changing traffic and user distribution. Under high load condition, HCA is more efficient than DCA. Under low to moderate load condition, HCA works better than FCA.

In HCA, channels are divided into fixed and dynamic set. Channels in the first set are allocated to cells (as in FCA); only calls originated from these cells can use them. The other set can be used by any other call, using DCA techniques. However it is used only when the first set is completely utilised. A call will be blocked when there are not available channels in both sets. Instead of block calls, an extended HCA queues calls when no channels are available [45]. Another version applies channel reordering to switch channels of ongoing calls to achieve the best performance [46]. If channel borrowing is used with HCA, a call in a borrowed channel will be shifted to a normal channel if a normal channel is free.

In all test cases in [26], HCA performs better than FCA when the ratio of fixed to dynamic channels is 3:1 and the system load is under 50%. When the load increases, HCA is always better than FCA.

Flexible Channel Assignment (FlCA)

FlCA is similar to HCA, which divides channels into two sets. The system estimates how many channels are required for a light traffic situation in a cell, then provides enough channels. This is the fixed channel set. The other is called the flexible set, which is allocated to the cells with increasing load. These channels are used as fixed channels. The assignment is done periodically or predicatively.

How often the assignment should take place depends on operators. Too frequent assignment results in high overhead. Therefore the change is often scheduled at the beginning of a traffic variation [25]. Operators need to statistically observe the traffic to choose the peaks.

With the predictive assignment, in each cell, the system monitors an indicator (such as the traffic intensity or the blocking probability) and triggers the mechanism if the indicator falls below a threshold.

Because the system only assigns flexible channels to certain cells, it needs information to decide when and how many channels to give. The information is centrally controlled.

Fixed and Dynamic Channel Assignment

Based on the characteristics of FCA (better under heavy load) and DCA (better under low to moderate load), this scheme simply shifts from DCA to FCA when the load increases and from FCA to DCA when the load decreases. The change happens gradually to avoid blocking [39].

The main channel assignment schemes have been investigated and summarised in Figure 2-4. At a system point of view we have looked at all the current channel assignment schemes. In the next section, we will look at the admission control schemes and how they apply the channel assignment schemes at the cell point of view.

2.2.2 Admission control schemes

The admission control decides to accept a call request or to refuse it. A rejected new call is referred to as blocked and a rejected handoff call as dropped. The algorithm is based on a number of criteria, such as: the current load, the predicted load, the traffic class of the call and the call's type (a new or handoff call). There have been a number of proposed admission control schemes for cellular networks.

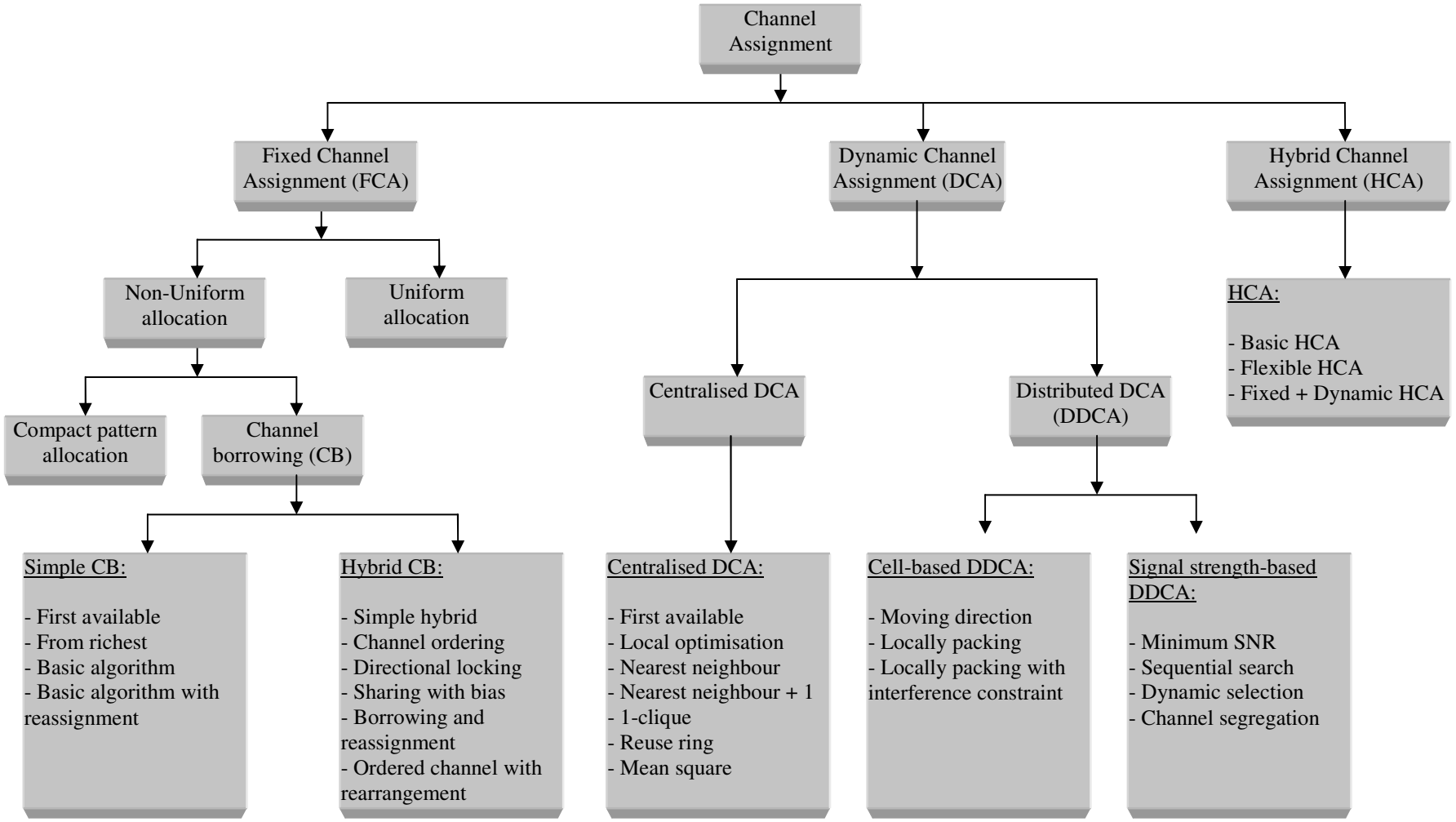


Figure 2-4: Summary of existing channel assignment schemes

Common goals

Most admission algorithms have four mutual goals.

First, they aim to utilise the resource as much as possible to generate more revenue for operators. The algorithm is expected to admit many calls as long as the effect of the admission is not severely harmful to the network. The decision must be adequately judged to avoid low quality of service due to exploiting the resource in a greedy way. Together with the admission, the algorithm needs to decide the bandwidth share for this call. An explicit example is an algorithm admitting all calls with whatever bandwidth is required. When all bandwidth is in use, to admit new call requests, the system will drop ongoing calls. In this case, the resource is over-utilised and users experience poor service (i.e. calls are forced to be terminated).

The second goal is to keep the probability of handoff dropping calls to minimum. Unexpected call termination often occurs during handoff process if handoffs are not properly handled. Improper handoffs include insufficient bandwidth reservation and late channel swapping. To users' point of view, it is more acceptable to block a new call request than to admit and drop it at a later stage [47-50]. Exclusively reserving bandwidth for handoff calls, early detecting handoff needs and reducing handoff time are possible ways to minimise the handoff call dropping probability.

The third goal is to limit the reduction of service for ongoing calls. When a call is connected, the user expects to have the same or better quality for the whole call duration. Reduction of service is a representative of an unstable network. Fluctuation in data transmission rate can result into software crashes which disturb users.

To operators, the final goal is as important as the first one: to prioritise services depending on customers request or on traffic types. Various applications have different QoS requirements which require different treatment from the system. VIP (very important person) customers or business people may have a need to be able to access the system whenever they need and be guaranteed a certain QoS requirements i.e. zero new call blocking probability, no handoff calls dropped and no service reduction. These value-added-services can bring more revenue to operators.

The second goal states that admission of handoff calls should be prioritised over new calls because people would be more upset if their ongoing calls are dropped than if they can not make a call. Before going to details of some proposed admission schemes, we are going to discuss the handoff in cellular networks.

Handoff definition and classification

Handoffs can be classified based on the simultaneous connections of a call at the handoff time. There are two types of handoff: “hard handoff” and “soft handoff”. In hard handoff, the terminal choose the handoff target base station by checking for the strongest signal and makes a “hard” decision to handed off to that station. The connection with the serving base station is broken before making the connection to the new base station. Whereas with soft handoff, the terminal monitors the signals from many base stations and builds a condition to handoff. At this temporary moment, the terminal can use traffic channels from all base stations. Connections to serving and new base stations are maintained until the condition is satisfied; then the connection to the serving base station is broken. Only 3G mobile terminals can perform soft handoff.

Another way to classify handoff is based on where the decision is made. It can be made in the base station or in the mobile terminal. The former scheme has two versions: Network-Controlled Handoff (NCHO) and Mobile-Assisted Handoff (MAHO). In NCHO, the base station is totally responsible for making decision. It sends a trigger to the terminal when handoff is required. NCHO is used in AMPS. In MAHO, the base station makes the decision based on the information sent from the terminal. The latter scheme is known as Mobile-Controlled Handoff (MCHO). The control is in the terminal. If the terminal relies on information from the base stations, it periodically communicates with them. Due to the hardware limitation of mobile terminals, most wireless systems use NCHO or MAHO.

Handoff can be differentiated to: intra-cell and inter-cell. Intra-cell handoff happens when the mobile terminal wishes to change to another channel with better quality. Inter-cell handoff is when the mobile terminal moves from one cell to another. This thesis focuses on the second type.

In hierarchical cellular networks, calls can be handed off vertically and horizontally. Vertical handoffs occur when calls are handed from a microcell to a macrocell and vice versa. Horizontal handoffs happen in the same tier.

Handoff measurements

To evaluate an admission model with handoff prioritisation, different researchers have different measurements. Some used the mean number of base stations in soft handoff or the mean number of handoffs and handoff delay [51] or even the system outage probability [52]. Other proposals used the probability of unnecessary handoff [53] or the average interferences on reverse and forward links [54]. However the most common evaluation criteria are the new call blocking probability [55] and the handoff call dropping probability [56]. We use these two probabilities as quality indicators in our simulation.

Proposed Admission Algorithms with Handoff Prioritisation

Admission algorithms have an important role in network operation. Inefficient algorithms lead to under-utilisation of the network resource or customers' complaints due to poor service. In the era of the IP expansion with multimedia applications, it is more important to maintain a call's quality for its life time than to admit more calls with poor quality. Therefore the trend of admission algorithms focuses more on handoff prioritising algorithms.

In a cell, where the number of channels is fixed, we rely on the admission control to priority handoff calls over new calls. Handoff prioritising schemes reduce the handoff call dropping probability with the expense of an increase in the new call blocking probability hence the total admitted traffic. There are many schemes with handoff prioritisation, most of which use similar ideas to channel assignment schemes.

Equal Access Sharing (or Complete Sharing)

It does not differentiate the traffic classes (realtime or non-realtime) nor does the traffic type (handoffs or new requests). All connections are equally treated. The system only checks to guarantee the request is satisfactory before admitting the call. Once admitting the call, the system never forces it to drop or reduces its transmission rate. When no bandwidth is available, all call requests are denied. The calls are served on first-come-first-served basis. This is the simplest to implement however it is impossible to prioritise calls over others. This scheme is similar to the simple DCA where none of the channels is assigned to any specific cells.

Equal Access Sharing with Priority (or Complete Sharing with Priority)

In the Equal Access Sharing scheme, when all channels are used, calls can be not admitted until the traffic is reduced. In this scheme, all calls access the resource in a first-come-first-serve basis until the resource is fully utilised. In high load situation, when there is a high priority call coming, the system will drop a low priority call and allocate the resource to the high priority call. High priority calls access the system easily, resulting in a low blocking probability. As a trade-off, there is no QoS guarantee for low priority calls because it can be dropped at any time. Although it may survive for its whole call duration, its transmission rate is maintained. Calls can be prioritised by the call nature (new calls or handoff calls) or by traffic classes. This scheme is rather annoying to users because their low priority calls can be dropped at any time in high load situation.

Equal Access Sharing with Reserve (or Complete Partition)

To overcome the unfair operation of low priority calls, the total bandwidth is divided into two partitions: all calls have equal access to the first partition but only high priority calls can use the second partition. The system may be under-utilised but it is possible to fine-tune the

blocking probability of certain calls. No calls will be dropped and their transmission rates are guaranteed. This scheme is similar to the simple FCA because different calls access different part of resource. The schemes in this stream are often referred to as "guard channel" schemes.

The schemes in [57-61] exclusively reserve a number of channels for handoff calls in each cell to ensure more successful handoffs. The other channels are shared between new calls and handoff calls. The handoff call dropping probability is reduced; however the new call blocking probability increases because less resource is available to new calls. To reduce the new call blocking probability, it is possible to queue new calls instead of blocking them [25].

The number of reserved channels is calculated based on the statistical information of the serving cell. The decision is therefore optimal locally but not necessarily globally. An enhanced version is the *distributed call admission control scheme* [62], where measurements are taken from the serving cell and surround cells.

The number of used channels is continuously monitored against a threshold value. If resource is available, a handoff call is always admitted. For new call requests, a call is admitted if the number of unused channels is greater than the guard threshold value. The value can be fixed or adjustable. Fixing the value can under-utilise the system while adjusting the value adds more traffic and computational load. The knowledge of traffic patterns is important to utilise resources and minimise computational load.

Threshold Access Sharing

This policy uses both sharing and partitioning concepts. It defines three admission levels. At the lowest threshold T_{low}, all calls are accepted equally (sharing). After the cell load reaches T_{low}, the system comes to a Conditional Access state. Calls are still admitted but they are aware that they may lose connections or have their service reduced if higher priority calls come. When the load increments to T_{high}, the system comes to the next state Priority Access. At this time, only high priority calls are admitted (Partition). If high priority calls still come, some low priority calls will be dropped or have their service reduced to spare bandwidth. The load is kept under a threshold T_{Bound} to ensure the stability of the system.

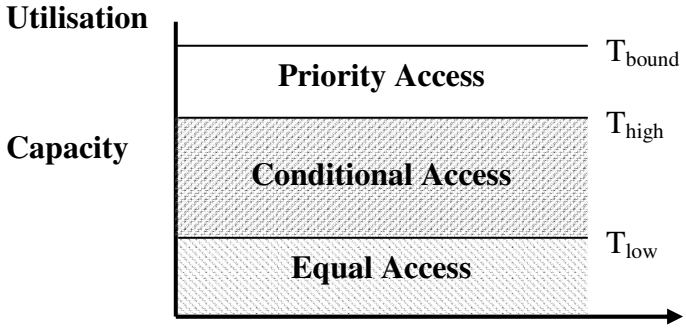


Figure 2-5: Threshold Access Sharing

Handoff queuing

The schemes do not drop handoff calls but add them to a queue for available channels. Calls in the queue are processed in priority i.e. new calls must wait until all handoff calls waiting in the queue are processed. The arrangement reduces the handoff call dropping probability but increases the new call blocking probability [59, 63].

The queue length and the processing time become the constraints in this case because handoff calls will be dropped while waiting due to weak signal strength or the queue is full [64].

The non-pre-emptive priority queuing discipline based on the mobile subscribers measurement is presented in [63]. When a handoff request arrives, it is ranked according to the distance between the mobile terminal and the base station, the speed of the terminal and the power level. Requests are handled in the order of high to low ranking.

These schemes can be used with *guard channels* to enhance performance. It is possible to queue all calls and to use another mechanism to prioritise handoff calls [60].

New call queuing

Based on the fact that new calls can tolerate more delays than handoff calls, the scheme uses a *guard channel* and queue all new calls to process handoff calls with priority [65]. When the system has handled all handoff requests, new calls in the queue are processed in turn. New calls are denied only when the queue is full. The simulation shows that the handoff call dropping is reduced, more calls are admitted and the total traffic increases.

Mobility-pattern

Using the analogy of the *Dynamic Channel Allocation - Moving direction* scheme, channels are reserved according to the knowledge of the terminal's movement. The reservation can be made adaptively [66, 67] or predicatively using mobility pattern [68-70].

The simplest adaptive scheme [71] just monitors the number of users near neighbouring cells and makes the reservation if the number exceeds a threshold. The *shadow cluster scheme* [68] uses information from a cell cluster and its shadow to estimate the bandwidth requirement. The scheme does not have the ability to define a shadow cluster in real networks.

The reservation can be done based on the knowledge of the *user's movement history* [69] as well. It assumes that each mobile will handoff to neighbouring cells with equal probability. *Mobility-Based call admission control* scheme [70] improves the movement prediction with the target cell and time by statistically analysing the mobility history. The resource is optimised while the handoff dropping probability is kept under control. A *semi reservation*

scheme is also proposed [72]. A modified version of the RSVP protocol [73] reserves resources in *all* neighbouring cells or only in a predicted destination cell.

Grouping traffic into realtime and non-realtime, other researchers [74-76] build an admission control to adaptively share the bandwidth among the classes.

We have looked at the proposed admission control schemes and analysed some of the important ones. Most of them have trade-off between the new call blocking probability and handoff call dropping probability.

In the summary, this chapter reviewed the QoS on IP-based and traditional networks with some common solutions such as IntServ, DiffServ. We have also studied the admission control schemes. In the next chapter, we will investigate in-depth the Threshold Access Sharing scheme. Its strengths and weaknesses will be critically analysed. An improved version will be introduced with supported results.

Chapter 3: IMPROVED THRESHOLD ACCESS SHARING SCHEME WITH SIMPLIFIED RATE-BASED BORROWING

Threshold Access Sharing (TAS) is one of the popular admission control schemes for cellular networks. In this chapter, we start with presentation of TAS. The operation and performance of TAS are analysed and simulated. We found that TAS has space for improvement by introducing the concept of a rate based borrowing scheme to TAS. In the second section, the rate borrowing scheme is presented and discussed. Based on our finding, an enhanced TAS scheme absorbing the good features of a simplified rate borrowing scheme is proposed. Our analysis and simulation has shown that the new scheme has improved the call blocking performance and dramatically decreased the handoff dropping probability in TAS by as much as 20%.

3.1 Threshold Access Sharing (TAS)

The Threshold Access Sharing (TAS) [77] was proposed by Moorman and Lockwood. It is proven to be one of the best access policies in admission control. The scheme can be summarised as in the following.

Traffic is divided into two classes: high priority and low priority. The first class will be treated with higher priority than the other one. High priority class includes handoff calls or data services with stringent QoS requirements. The rest of the traffic belongs to the low priority class. There are three admission levels in the capacity, namely: T_{low}, T_{high} and T_{bound}. When the utilisation reaches a level, the admission definition in that level is applied to the traffic. Its operation can be summarised as a state diagram in Figure 3-1.

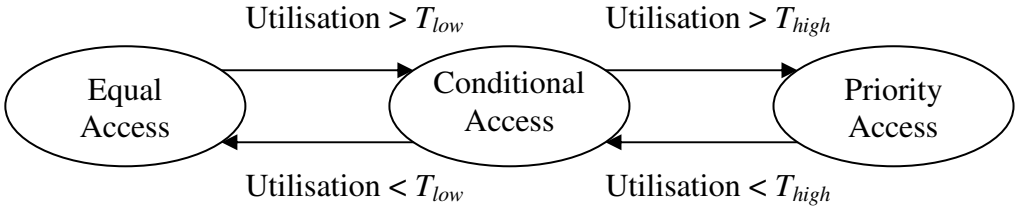


Figure 3-1: State diagram of Threshold Access Sharing scheme

If the utilisation is below T_{low}, the system is said to be in *Equal Access* state. Calls satisfying the basic requirements (bandwidth required, delay bounds, jitter etc.) will be admitted and treated in the same way. When more calls are admitted, the utilisation gets higher. Once it passes the threshold T_{low}, the system moves to the *Conditional Access* state. High priority calls are admitted while low priority calls are admitted with the condition that they may have their service reduced or discontinued if more high priority traffic arrives. The *Priority Access* is the

last state when the utilisation passes T_{high} threshold. Admission requests for new low priority calls are completely denied. Ongoing low priority calls which have been conditionally admitted will have their service reduced to accommodate resource for high priority calls. The system continuously monitors the utilisation and compares it to the thresholds to make necessary state transitions. The system keeps a T_{bound} level where all calls are rejected to avoid system breakdown.

In this scheme, the channel utilisation is high [77]. Handoff calls are treated as high priority calls and never dropped. High priority calls are never blocked. Low priority calls are affected in *Priority Access* state.

Table 3-1: Summary of Threshold Access Sharing admission

	High priority	Low priority
Equal Access	Admit	Admit
Conditional Access	Admit	Admit with condition
Priority Access	Admit	Reject

The main measurements are used to evaluate this scheme: the new call blocking probability, the handoff call dropping probability, the probability that a conditionally admitted call has its service reduced and the probability that a conditionally admitted call is forced to terminate. A flowchart summarising TAS operation is shown in Figure 3-2.

Our analysis and observation has led to the following findings:

Advantages of TAS:

Traffic is treated differently after classified into high and low priority, which guarantees customised QoS requirements. The state transitions based on the current system utilisation do not take much computational load and time. The cell only needs to maintain the current utilisation and check against the thresholds to make transitions. The scheme is verified by simulation to be the best accessing policy [78].

Disadvantages of TAS:

Traffic is classified as high priority (handoff calls) or low priority (other calls) but it did not discuss about other types of traffic such as realtime and non-realtime. The algorithm may drop conditionally admitted calls to admit higher priority calls (the trade-off of the decrease of handoff call probability and the increase of forced termination probability). Moreover, it was not clear how the service reduction occurred. When service reduction is required, the scheme did not discuss whether it attempts to reduce service of current calls before or after attempting to use bandwidth in the Priority access area. The constraint of the service reduction was not defined either. For example, will conditionally admitted calls have their service reduced until

they are dropped? It did not state whether the high priority call is admitted with whatever bandwidth requested or it will need to negotiate for a lower requirement. It is not practical just to reduce service of current calls to satisfy another with deluxe bandwidth requirement.

We found that it is possible to overcome TAS disadvantages such as the lack of traffic type prioritisation as well as the forced termination of conditionally admitted calls. We have proposed an improved Threshold Access Sharing (iTAS), in which, we replaced the idea of "dropping a current call to admit a new call" by introducing a simplified rate-based borrowing scheme. We will describe the concept of rate-based borrowing scheme in the next section.

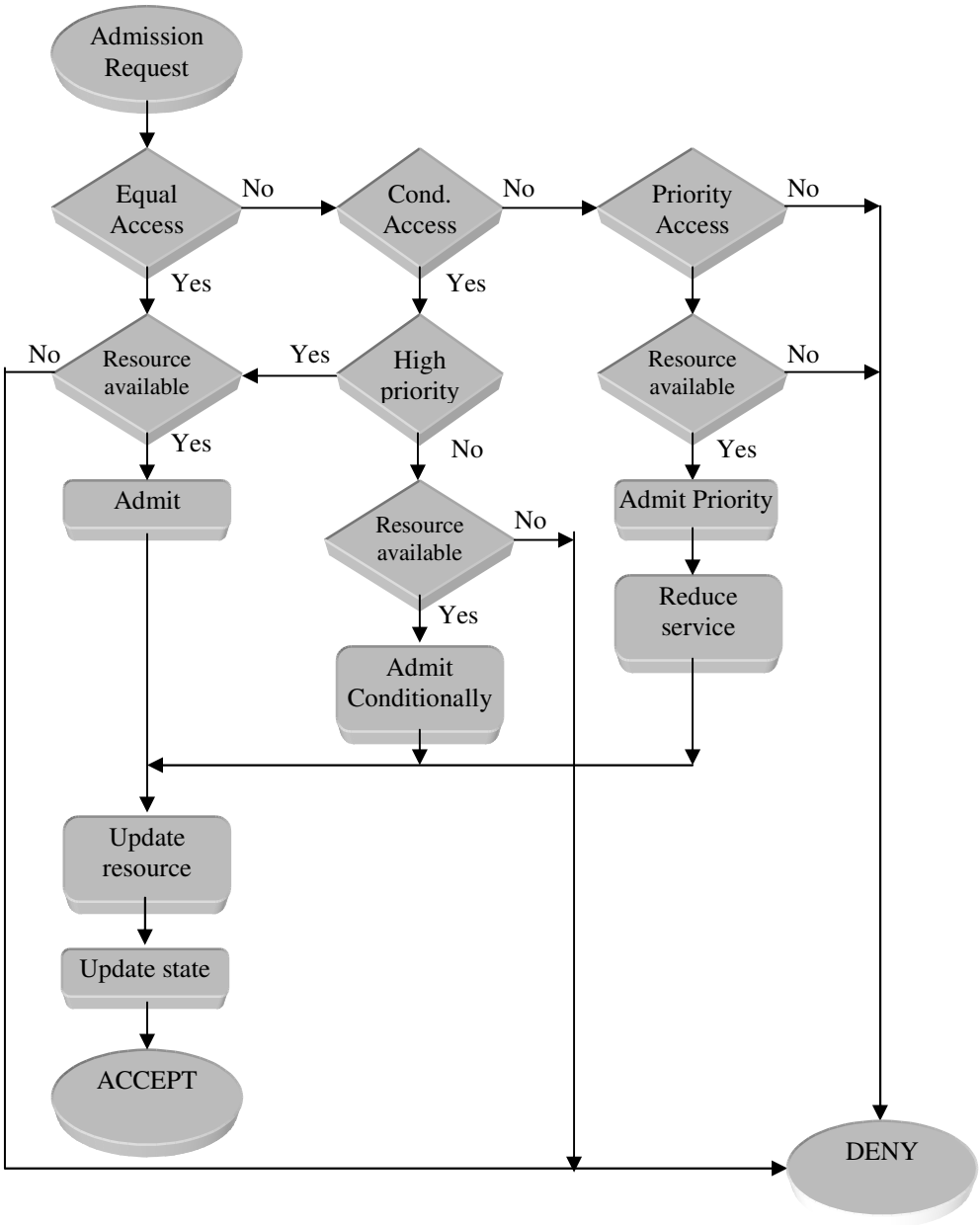


Figure 3-2: Threshold Access Sharing

3.2 Rate-based borrowing scheme

The scheme was proposed by El-Kadi and Olariu in [79] for QoS provisioning in wireless networks with multimedia traffic supporting adaptive bandwidth. In the scheme, traffic is classified as Class I (realtime) and Class II (non-realtime). Each one is treated differently. When requesting a connection, a call has to specify the traffic class, the desired bandwidth and the minimum acceptable bandwidth. The desired bandwidth is the bandwidth level at which the call wishes to operate. The minimum acceptable bandwidth is the lowest level that the call can operate with a certain QoS requirement. The borrowing mechanism works in a fair way, which means the number of bandwidth shares in borrowed from a connection is proportional to its bandwidth loss tolerance (to be discussed later). When bandwidth becomes available due to call completion or handoff, the excess bandwidth is returned to the call where it was borrowed from.

Operation Details

When a call specifies its *minimum acceptable bandwidth* (m) and its *desired bandwidth* (M), the system calculates its *Bandwidth Loss Tolerance* (BLT) as the difference between the two values:

$$BLT = M - m \tag{3-1}$$

Only bandwidth adaptive calls can tolerate bandwidth. Constant bit rate calls do not have *bandwidth loss tolerance* ($BLT = 0$) because the *desired bandwidth* and the *minimum acceptable bandwidth* are the same.

The *actual borrowable bandwidth* (ABB) of a call is defined as:

$$ABB = f \times BLT = f(M - m) \tag{3-2}$$

Where f $(0 \le f \le 1)$ is the fraction of the BLT that the call has to give up in the worst case.

The ABB of all calls is divided into λ shares, each has the value of:

$$\frac{ABB}{\lambda} = \frac{f(M-m)}{\lambda} \tag{3-3}$$

bandwidth units.

The borrowing mechanism takes bandwidth from each connection gradually, one λ at a time.

If all calls in a cell are borrowed a number of shares L, the cell is operating at level L. It means that L is an integer in the range of $[0;\lambda]$.

When a call is admitted, the cell guarantees that its bandwidth requirement does not fall below the *minimum expected* (MEX) level.

$$MEX = M - ABB = M - f(M - m) = (1 - f).M + f.m \tag{3-4}$$

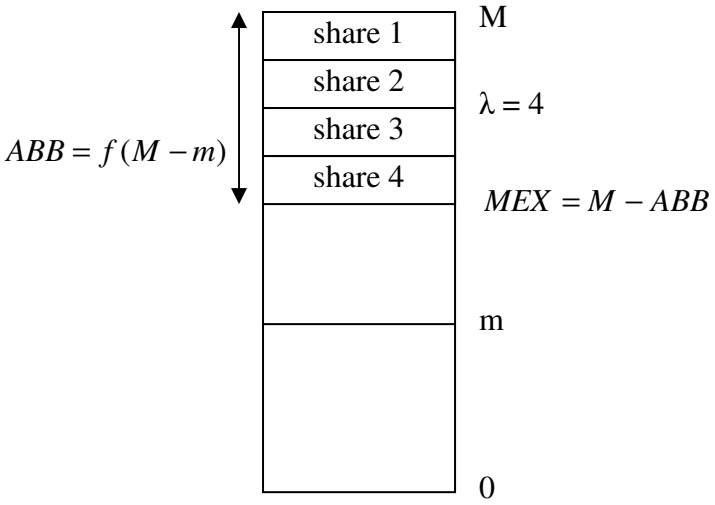


Figure 3-3: Demonstration of parameters in Rate-based borrowing scheme

In the worst case, where the system needs all borrowable bandwidth, the f value is maximal: $f = 1$. Hence $MEX = m$, i.e. the call maintains the connection at the *minimum acceptable bandwidth*.

To guarantee fairness of the scheme, the *adaptivity* (AD) parameter is introduced as the ratio of the bandwidth loss tolerance and the desired bandwidth.

$$AD = \frac{M - m}{M} \tag{3-5}$$

Higher AD means that the call is more adaptive i.e. the bandwidth is more adjustable. It also means that the call has lower forced termination probability.

Fairness

To demonstrate the fairness of the scheme, the paper [79] looks at an example. A cell operating at level L is going to accept a call C with the desired bandwidth M and minimum acceptable bandwidth m. To fit in the operating level of this cell, the call needs to reduce its operating bandwidth level to the same level (L) i.e. the call will decrease

$$L \times \frac{ABB}{\lambda} \tag{3-6}$$

bandwidth units from its desired bandwidth M.

So it will be admitted to the cell with the effective bandwidth of:

$$M - L \times \frac{ABB}{\lambda} \tag{3-7}$$

The loss ratio (*LR*) of the call is the ratio between the bandwidth reduction to fit in to the cell and the desired bandwidth *M*.

$$LR = \frac{L \times \frac{ABB}{\lambda}}{M} = \frac{L \times \frac{f(M-m)}{\lambda}}{M} = \frac{L.f}{\lambda} \times \frac{M-m}{m} = \frac{L.f}{\lambda} \times AD \tag{3-8}$$

In a cell, the number of shares λ is fixed. At a certain time, the cell at level *L* has a certain *f*. In other words, $\frac{L.f}{\lambda}$ is fixed and the loss ratio only depends on *AD*, the *adaptivity* of the call. *Adaptivity* parameter denotes how adjustable the bandwidth requirement of the call is. Different calls have different values of *AD*. Highly adjustable calls will lend more bandwidth compared to less adjustable calls. This is the fair operation of the scheme.

Admission of New Calls

Considering a new call request to a cell operating at level *L*, the cell tries to admit the call so that it will be best fitted in the cell i.e. it will work at level *L* as other calls. The effective bandwidth for the call is given in (3-7). The cell tries to accommodate the effective bandwidth given for the call. If the bandwidth does not have enough, it will estimate the excess bandwidth if it goes to the next operating level *L+1*. If there is enough bandwidth at the new level, it will advance to level *L+1* and grant bandwidth to the connection. The call is blocked otherwise (we did not allow the cell to advance to level *L+2* or higher). If the cell is at level $L = \lambda$ (i.e. all calls have given maximum borrowable bandwidth), the call is blocked.

Admission of Handoff Calls

For Class II handoff calls, the process is similar to the case of new calls. The cell first attempts to admit the call at its desired bandwidth *M* minus *L* shares as shown in (3-7). If bandwidth is enough, the call is admitted. Otherwise the cell moves to the next level. If the cell is already at the highest level, the call is denied.

The schemes reserves an amount of bandwidth for Class I handoff calls. If bandwidth is not enough, a portion of the reserved bandwidth will be used just to help the handoff call meet its minimum acceptable bandwidth requirement. If the bandwidth borrowed from other connections and bandwidth taken from the reservation can not fulfil the minimum requirement, the cell steps to the next level *L+1* to gain some more bandwidth. If it fails to

provide bandwidth in the next level, the call is dropped. Calls operating at minimum acceptable bandwidth level are not going to give in bandwidth.

Bandwidth Return

When bandwidth is available, the cell attempts to step down from level *L* to level *L-1*. The free bandwidth from the transition for each call is:

$$L \times \frac{ABB}{\lambda} - (L-1)\frac{ABB}{\lambda} = \frac{ABB}{\lambda} \qquad (3\text{-}9)$$

A flow chart of this scheme is in Figure 3-4.

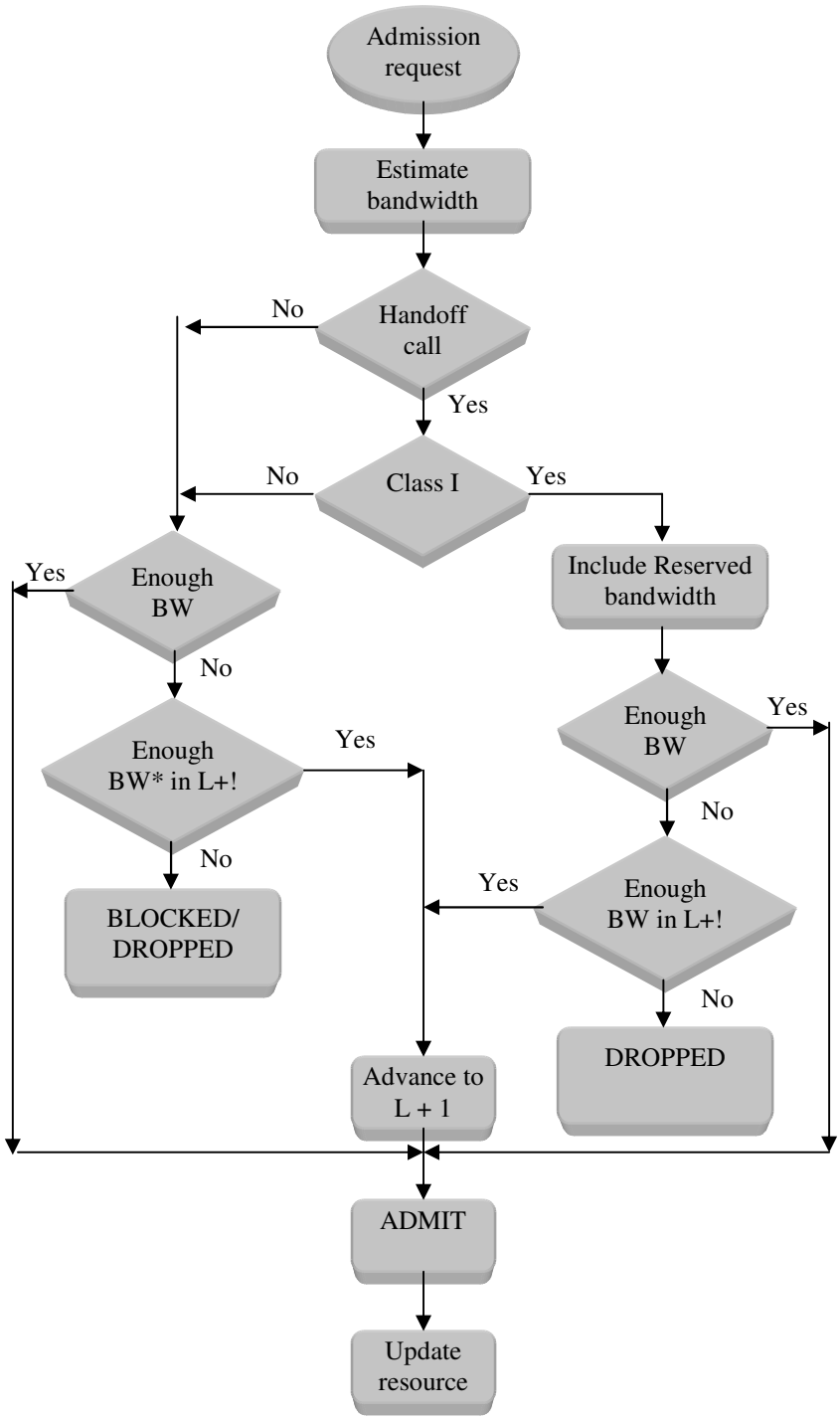


Figure 3-4: Rate-based borrowing scheme

Our analysis and observation on the rate borrowing scheme has led to the following findings:

Advantages of the Rate Borrowing Scheme:

In general the scheme is useful for bandwidth adaptive traffic, constant bit rate traffic does not benefit. Once connected, a call is never forced to terminate if it remains in the serving cell. It may however have its service reduced to a minimum level because its bandwidth may be borrowed for other connections. When the load decreases, the borrowed bandwidth will be returned. Decrement and increment of bandwidth is performed in a small increment to ensure the applications in the terminal to have enough time to adapt. The amount of bandwidth borrowed from a call is proportional to its bandwidth loss tolerance. Available bandwidth (from call handoff to another cell or call completion) will be returned to current calls. Calls operating at the minimum expected level will not lend any bandwidth.

Disadvantages of the Rate Borrowing Scheme:

Firstly, it assumes that all calls are operating in a cell at the same level L which is not necessary. Some may be at lower level (e.g. Class I handoff calls using the reserved bandwidth to make up the minimum acceptable bandwidth level). Secondly Class II handoff calls are not prioritised as Class I handoff calls are. To our's point of view, handoff calls should be treated with higher priority than new calls regardless its traffic class. Thirdly, the scheme also allows the cell to automatically adjust itself from a high level to a lower one if bandwidth is available. However it is not clear how bandwidth can be returned to the original call where it was borrowed from. It is not clear either what happens if the available bandwidth is not enough to be returned to *all* connections. Fourthly, although the scheme mentioned that the bandwidth allowance of Class I handoff calls effected at the handoff time will be replenished in a later stage; it is not clear how this is feasible. Fifthly other calls such as Class II calls were not considered in the bandwidth return process. Sixthly, level transitions in a cell and bandwidth adjustment of all connections result in computational overhead if the traffic is too transient (the transitions will happen too frequently). Seventhly the cell tries to step to the next higher level only. If the first attempt is not successful, it gives up. Therefore, the scheme can be redesigned to consider level transitions from L to $L + 2$. However this requires more overhead. We also noticed that the cell needs to maintain local parameters f and L continuously and each connection maintains parameters M, m, BLT, ABB. This takes a significant amount of bandwidth and computational load for updating. Finally parameter f has an insignificant role in the scheme. It only shows how much a cell allows its calls to lend bandwidth. This value is the percentage of the bandwidth loss tolerance that can be given away. The scheme did not discuss how to choose the best value for f to achieve the best

performance. If $f = 1$ i.e. calls are allowed to lend bandwidth until it works its minimum acceptable bandwidth level, all equations become simpler. A comparison is shown below.

Table 3-2: Complexity comparison for value *f*

	$(0 \le f \le 1)$	$f = 1$
Actual borrowable bandwidth	$ABB = f(M - m)$	$ABB = BLT = M - m$
Bandwidth allocation of each share	$\frac{ABB}{\lambda} = \frac{f(M-m)}{\lambda}$	$\frac{ABB}{\lambda} = \frac{BLT}{\lambda}$
Minimum expected bandwidth	$MEX = M - ABB$	$MEX = m$

In next section we will present our proposed solution which aims to improve the Threshold Access Sharing by integrating with a modified Rate-based borrowing scheme.

3.3 Improved Threshold Access Sharing (iTAS) With Simplified Rate-Based Borrowing

We start this section with our initial attempt to improve TAS with a reserve bandwidth pool for handoff calls. Then we will integrate a modified rate-based borrowing to iTAS. Our final solution can be briefed as following: There are four admission states, each has different admission policy. The cell advances to the next state if the utilisation passes a threshold. Some calls are admitted with the desired bandwidth and able to keep this bandwidth allocation for its whole duration. Some other calls are admitted with the condition to give away some of their bandwidth allocation if the cell receives higher priority calls. Periodically, the system tries to allocate free bandwidth back to calls affected from the borrowing. Once there are no calls affected and more bandwidth available, the system moves to a lower state.

3.3.1 Improved Threshold Access Sharing (iTAS)

In the improved version, we further classified traffic based on the traffic types and call natures. The two traffic classes are Class I and Class II as discussed above. The nature of the calls is either a new call or a handoff call. The following priority is suggested. Class I handoff calls have the highest priority, followed by Class II handoff calls, Class I new calls; finally Class II new calls have the lowest priority. The order is shown below.

1. Class I handoff calls
2. Class II handoff calls
3. Class I new call requests
4. Class II new call requests

Calls belonging to different categories will be handled differently. We have adopted the concept of bandwidth reservation in the improved version of Threshold Access Sharing scheme. At the setup phase, the call needs to specify the traffic class, the *minimum acceptable bandwidth* and the *desired bandwidth* as in the Rate-based borrowing scheme.

The bandwidth utilisation is divided into four levels as in Figure 3-5. The four thresholds are T_{low}, T_{high}, $T_{reserved}$ and T_{bound}.

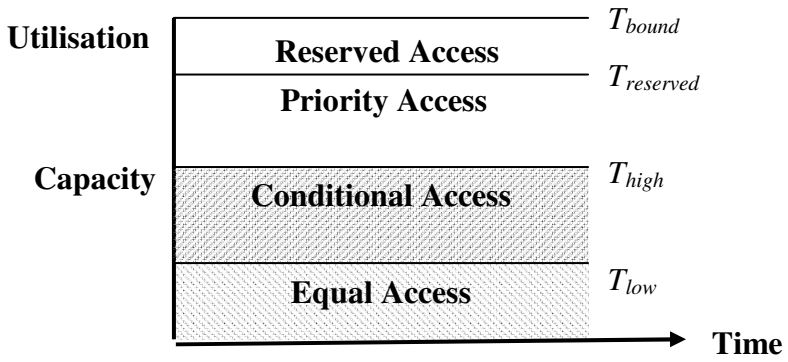


Figure 3-5: Improved Threshold Access Sharing

Below T_{low} is the Equal Access area, which accommodates all calls equally if basic conditions are satisfied. After the utilisation reaches T_{low}, the system moves to the Condition Access state. In this state, Class I calls are admitted at their desired bandwidth level. Class II calls are also admitted with the condition that their bandwidth allocation may be reduced to the minimum acceptable level. If more calls arrive, the utilisation increases above Thigh level, the system is in the Priority Access state. Resource now is used for high priority calls only. Class II new calls are rejected. Other calls are conditionally admitted. When the system progresses to the Reserved Access state all new calls are rejected and handoff calls are conditionally admitted. The system keeps the traffic below T_{bound} threshold to maintain a stable system. The operation is summarised in Table 3-3**.** The state transition diagram and the flowchart of iTAS scheme are shown in Figure 3-6 and Figure 3-7 respectively.

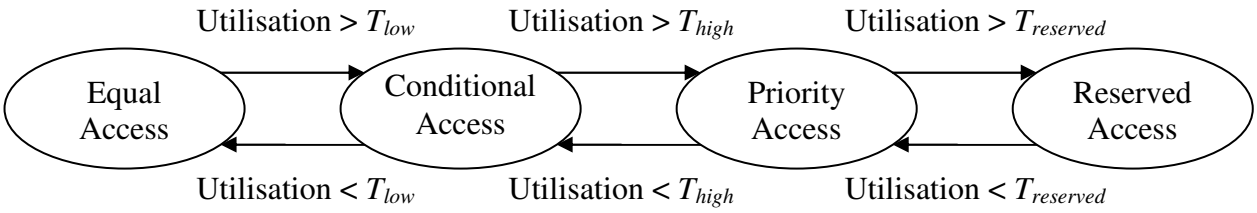


Figure 3-6: State diagram of the Improved Threshold Access Sharing

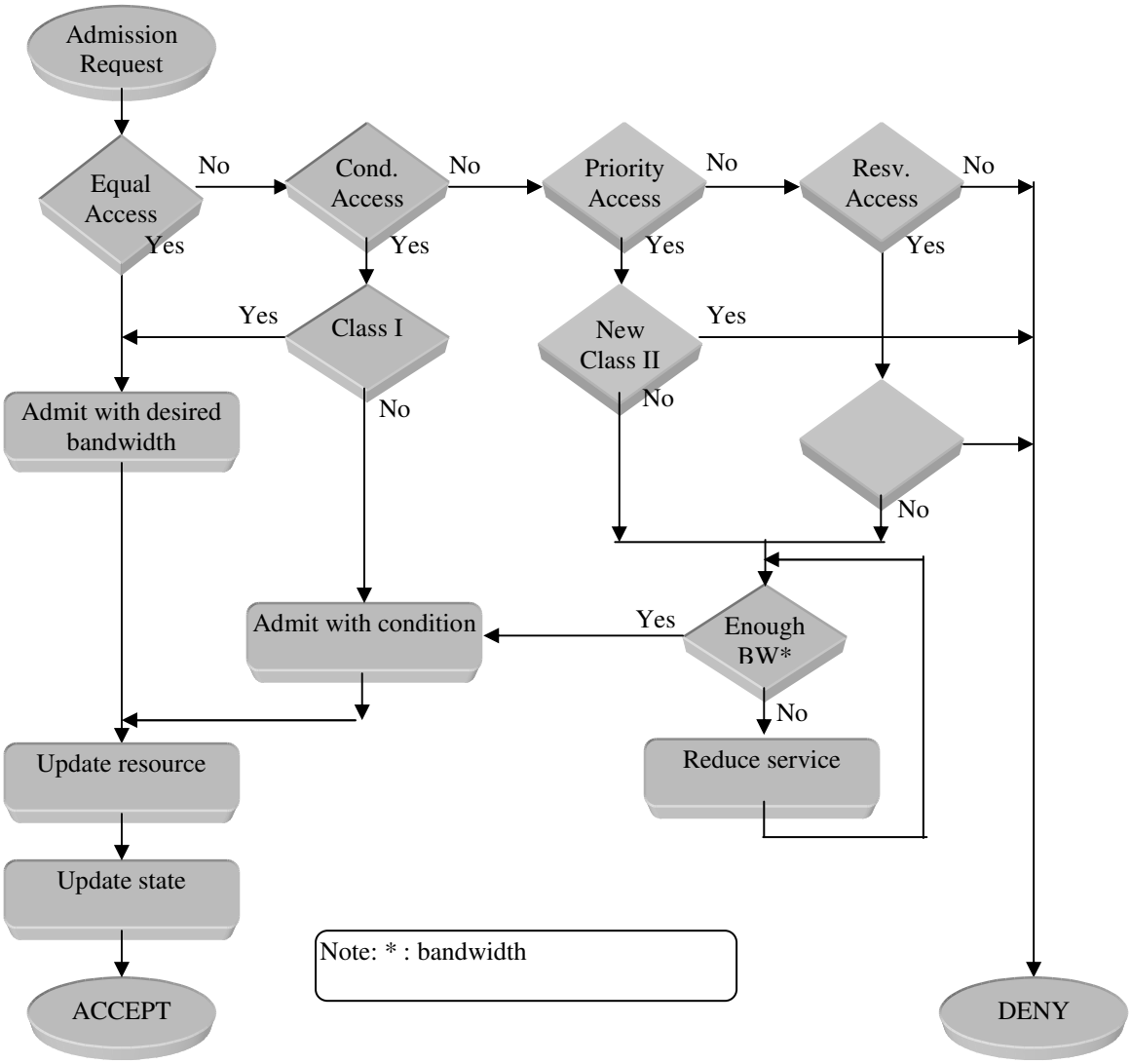


Figure 3-7: Improved Threshold Access Sharing

Table 3-3: Summary of admission decision in improved TAS

			SHARING THRESHOLDS			
Call		**Algorithm**	**Equal Access**	**Conditional Access**	**Priority Access**	**Reserved Access**
Traffic class	**Call nature**					
Class I	Handoff	*Decision*	Admitted	Admitted	Conditionally admitted	Conditionally admitted
		Bandwidth	Desired	Desired	Desired	Desired
	New	*Decision*	Admitted	Admitted	Conditionally admitted	Rejected
		Bandwidth	Desired	Desired	Desired	
Class II	Handoff	*Decision*	Admitted	Conditionally admitted	Conditionally admitted	Conditionally admitted
		Bandwidth	Desired	Desired	Desired	Desired
	New	*Decision*	Admitted	Conditionally admitted	Rejected	Rejected
		Bandwidth	Desired	Desired		

The scheme has some improvements over the original as below:

Lower forced termination probability. To ensure a continuous service for users, it is preferable that after admitted in a cell, the call will not be forced to terminate. In the improved version, once admitted, the call will not be forced to terminate if it stays in the same cell although its service may be reduced. However, it may be dropped during handoff if the target cell can not provide resource.

Service reduction. The system can reduce the bandwidth allowance of calls to the minimum acceptable level, at which calls can still survive. It only picks some calls to reduce service. The way of choosing calls is discussed later. The scheme only performs service reduction when there is not any bandwidth left i.e. it uses bandwidth in all area before attempting service reduction.

In iTAS, although we have successfully removed the forced termination case and replaced with the service reduction. We found that it is possible to increase the fairness of reducing service of current calls by borrowing bandwidth from "rich" calls with higher bandwidth loss tolerance than others. In next section, we will present a modified version of the rate-based borrowing scheme and integrate it with the improved TAS.

3.3.2 iTAS with Rate-Based Borrowing

From our observations, the rate based borrowing scheme can be modified and simplified in such a way that its main features can be kept but with far less overhead.

Parameter *f* is fixed to 1 i.e. the maximum bandwidth that a call is allowed to give away is equal to the bandwidth loss tolerance. This modification does not allow the control of borrowable bandwidth at the cell level; however, when applying with the improved Threshold Access Sharing, the micro-control can be ignored. The actual borrowable bandwidth is therefore the same with the bandwidth loss tolerance.

There are two parameters in a call: the *desired bandwidth* (M) and the *minimum acceptable bandwidth* (m) levels, specified at setup time. The cell keeps *sorting index* values of all calls. This parameter represents how severely the call is affected. The scheme borrows bandwidth from calls with the greatest *sorting index* and return bandwidth to calls with smallest *sorting index*. The way how bandwidth is returned will be discussed in details. All calls can receive bandwidth return.

The scheme attempts to borrow bandwidth from current calls until there is enough to accommodate the request. After borrowing a portion of bandwidth from all current calls, if the

bandwidth is still not enough, it tries another time to borrow another portion. This results in longer waiting time.

Operation details

The call keeps the *minimum acceptable bandwidth* (m), *desired bandwidth* (M), and *Bandwidth Loss Tolerance* (BLT) (introduced in formula (3-1)). The *effective bandwidth* (EB) is the actual bandwidth that the call is using.

Parameters *f* is set to 1, hence the *actual borrowable bandwidth* (ABB) of a call will not be used because it will take the value of the bandwidth loss tolerance. The minimum expected bandwidth can be ignored as well. The call can give an amount of bandwidth equivalent to the bandwidth loss tolerance i.e. it will give away bandwidth until its operating bandwidth allocation reaches the minimum acceptable bandwidth level.

The bandwidth loss tolerance is divided into λ shares. Each bandwidth share has

$\frac{BLT}{\lambda} = \frac{M-m}{\lambda}$ bandwidth units. The call gives away one share at a time.

The borrowing mechanism takes bandwidth from each connection gradually, one λ at a time.

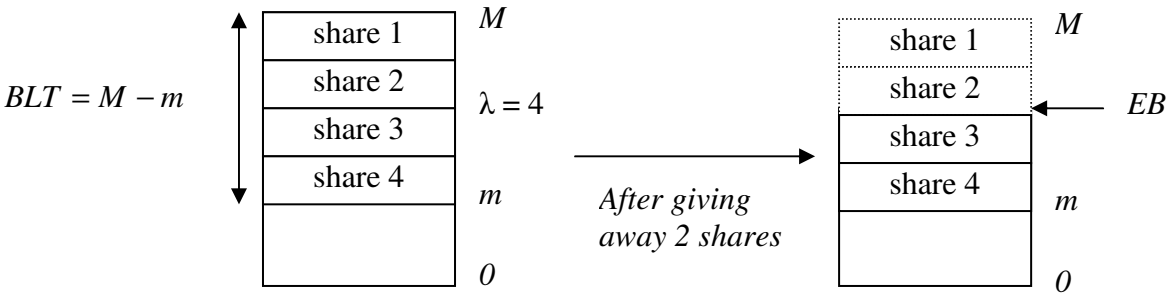


Figure 3-8: Parameters used in Simplified rate-based borrowing scheme

Parameter λ is chosen so that the ratio of $\frac{M-m}{\lambda}$ is an integer. This makes the computation easier. Instead of the adaptivity, we used the *sorting index* (*s*) to prioritise the bandwidth borrow and bandwidth return. The parameter is the ratio between the effective bandwidth and the bandwidth loss tolerance.

$$s = \frac{M-EB}{BLT} = \frac{M-EB}{M-m} \qquad (BLT \neq 0 \Leftrightarrow M \neq m) \qquad (3\text{-}10)$$

Smaller *sorting index* means the call has given away more bandwidth and it is operating at a low bandwidth level. This kind of calls are deserved to have compensation from the bandwidth return.

The cell keeps a database of all calls' *sorting indices*. When there is a need for bandwidth borrowing, the scheme picks the calls with the greatest *index* to borrow bandwidth. It then updates the database entry of those calls.

The scheme periodically checks for available bandwidth. If there is any, each call will receive one bandwidth share, starting from the one with the smallest *sorting index*. The return process stops when all free bandwidth is given back to ongoing calls or when all ongoing calls are operating at the *desired bandwidth* level. The *sorting index* database is updated. When all calls have the *desired bandwidth*, the system will attempt to move to the lower state.

The scheme does not care about the traffic class or the nature of the call when borrowing or returning bandwidth. If a call is tolerant to bandwidth loss, it will be responsible to lend bandwidth to other calls. Constant bit rate calls need to specify the same value for the *desired bandwidth* and the *minimum acceptable bandwidth* (the *bandwidth loss tolerance* is 0, and the *sorting index* can not be estimated).

A flowchart of the simplified bandwidth borrowing and returning processes is shown in Figure 3-9.

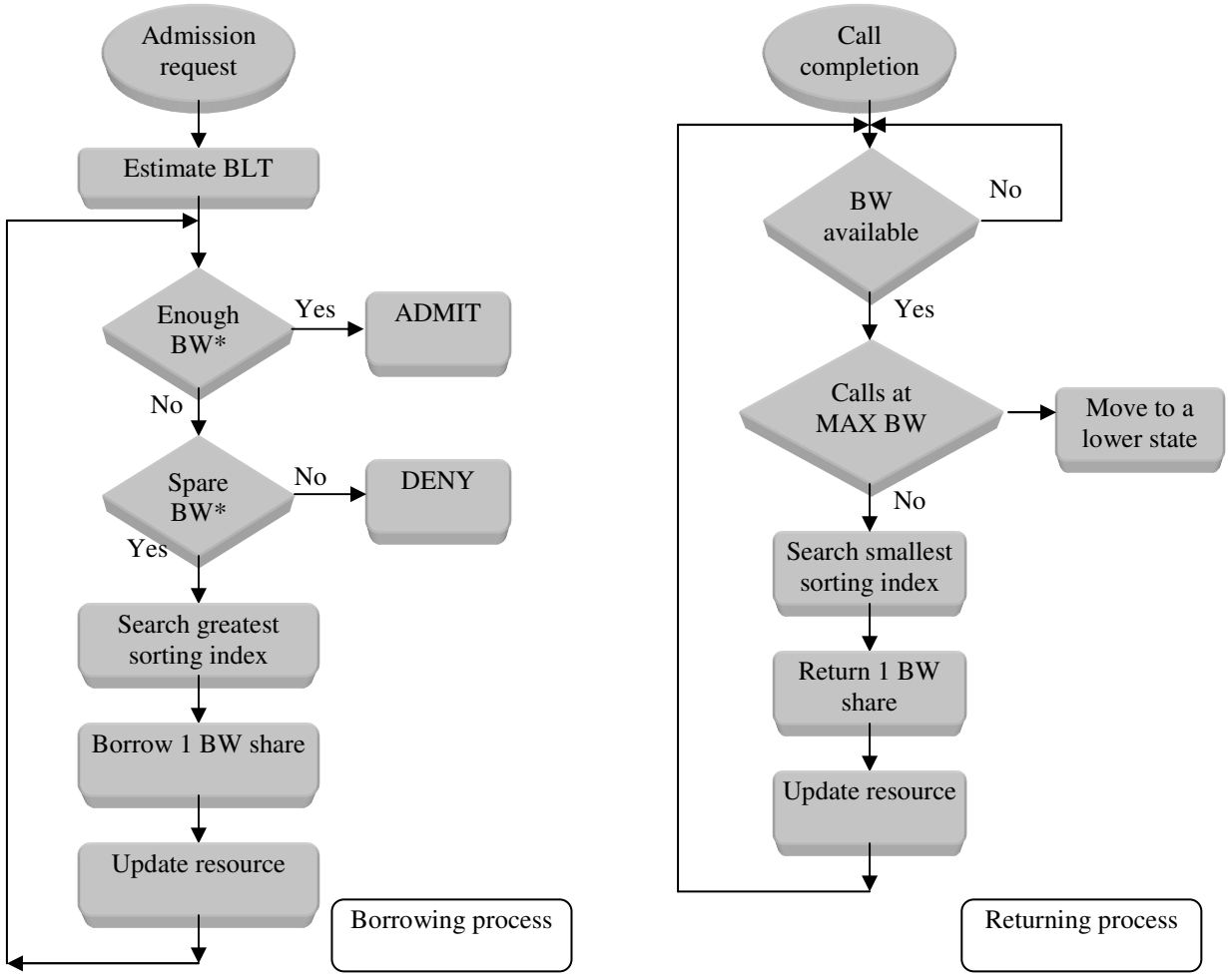


Figure 3-9: Borrowing and returning process

3.4 Simulation

Due to the complexity of the algorithm, the demonstration of the pre-eminence of the proposed scheme is shown by a simulation. We have studied the possibility of using available simulation packages such as NS-2, OMNeT++, OPNET, GloMoSim, Mathlab and even a simulation test bed from another researcher. However, none of the above is ideal for the required simulation. Therefore, we have decided to set up the simulation environment by creating our own modules using Java programming language.

The simulation implements the proposed scheme in a simple cell. The observation was taken in the range of approximately 1 to 500 erlangs of traffic.

Traffic is classified into Class I for realtime applications and non-realtime applications. Each call must specify its traffic class, the *desired bandwidth* and the *minimum acceptable bandwidth*.

There are three admission levels: full admission with the *desired bandwidth*, conditional admission with the *desired bandwidth* (but it may be reduced to the *minimum acceptable bandwidth*) and rejection.

The cell has 90 bandwidth units (BU) in total (T_{bound} = 90). Ten percents (9 BU) are reserved for handoff calls in the Reserved Access state, $T_{reserved}$ = 90-9 = 81 BU. Ten percents are allocated for the Priority Access state, T_{high} = 81 – 9 = 72 BU. Twenty percents are allocated for the Conditional Access state, T_{low} = 72 – 18 = 54 BU.

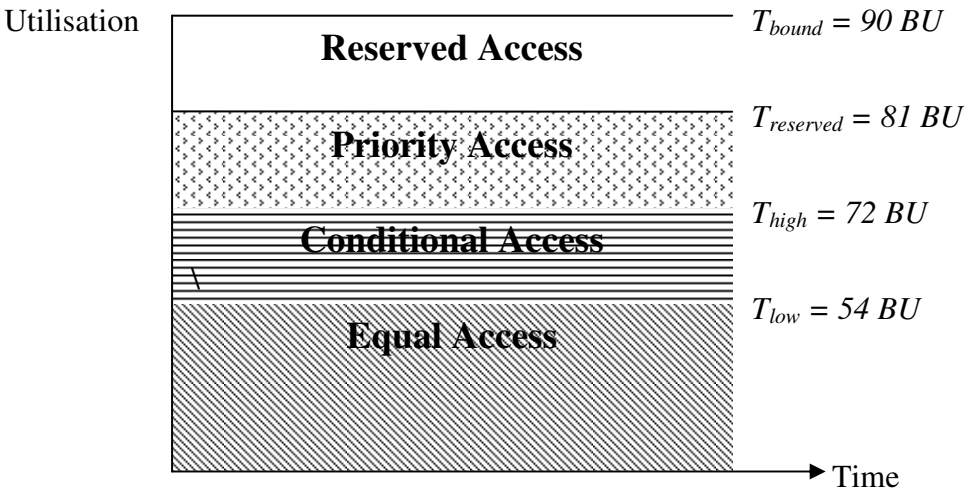


Figure 3-10: Simulation values in Improved Threshold Access Sharing

The simulation is set up to have two different, independent traffic sources. One is the new call request and one is the handoff call request. The number of handoff calls is about 20% of the new calls. The traffic is increased by increasing the arrival rates of the two sources.

Class I calls are randomly generated; the total number of Class I calls is about 20% of the total. The desired bandwidth of Class I calls is 8 BU and the minimum acceptable bandwidth is 3 BU. For Class II calls, the desired bandwidth is 6 while the minimum acceptable bandwidth is 1 BU. In the real situation, different calls have different bandwidth requirements.

So the bandwidth loss tolerance of the calls is $BLT = 8-3 = 6-1 = 5$ BU. There will be five bandwidth states dictating how many bandwidth portions are given away. The state diagram is shown in Figure 3-11.

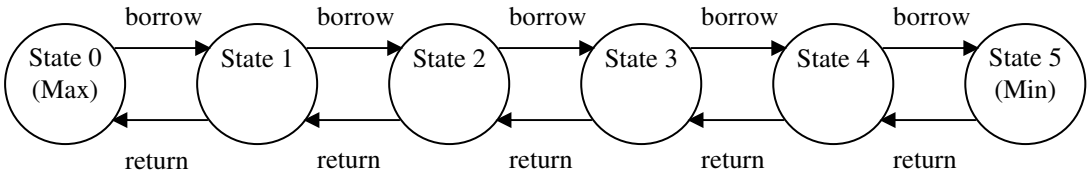


Figure 3-11: Bandwidth states diagram of a call in the simulation

At State 0, the call operates at the maximum bandwidth allocation. Every time it gives away one bandwidth portion, it moves to the next state. The final state is State 5 where the call survives from disconnection but it operates at the minimum bandwidth level. It moves in the reverse direction if it receives available bandwidth from a completed call.

Operation

A call request (either new or handoff call) tries to use bandwidth in the Equal Access state. If bandwidth in this state is not enough for the connection, the cell moves to Conditional Access state to accommodate bandwidth for the call. If the call is of Class II, it is subject to the admission condition. The process continues to higher states. The call is either blocked or admitted based on its traffic class, call type and call nature.

There are some counters in the simulation. Every time a decision is made to block or to admit a call, the counter increments.

Results and Discussion

To see the effect of the admission algorithm in the new scheme, the first result is compared to Erlang B formula. Figure 3-12 shows the better performance of iTAS over TAS in terms of handoff dropping. The new call blocking probability of the iTAS is slightly higher than that of

the TAS. The reason is the number of bandwidth available to new calls is less due to reservation. The trade-off of the higher new call blocking probability is the much lower handoff call dropping probability of iTAS. Depending on the load, the dropping probability of iTAS is about 0.2 less than that of TAS.

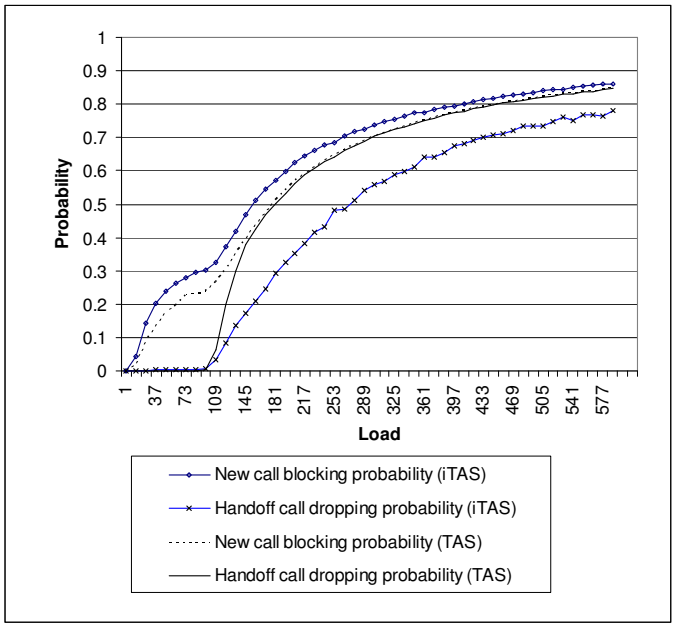


Figure 3-12: Blocking and dropping probability of TAS and iTAS

It should be noted that the Formula B values represent the simplest admission scheme, which processes call requests in the order of arrival and blocks if no resource is found. Figure 3-13 displays the observation on the blocking probability for each traffic class

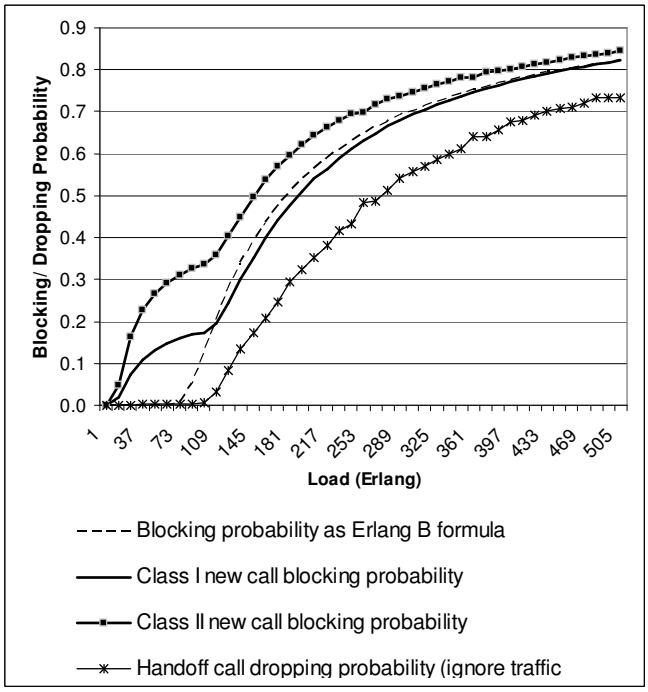


Figure 3-13: New call blocking/ Handoff call dropping probability of iTAS

It can be seen that:

1. For the load less than 100 erlangs, the new call blocking probabilities of both Class I and Class II calls are much higher than that from the Erlang B formula. In the mean time, the handoff call dropping probability is much lower. The first few values can be seen here.

Erlang	Erlang B formula	New call blocking probability	Class I new call blocking probability	Class II new call blocking probability	Handoff call dropping probability	Class I handoff call dropping probability	Class II handoff call dropping probability
1.2	2.71E-132	0	0	0	0	0	0
13.2	8.85E-44	0.042047	0.018047	0.048034	1.83E-04	7.97E-04	3.01E-05
25.2	1.02E-23	0.144405	0.073322	0.162208	9.54E-04	0.003809	2.43E-04
37.2	1.05E-13	0.204195	0.109869	0.227678	0.001773	0.006797	5.11E-04
49.2	5.47E-08	0.239866	0.132823	0.26665	0.002187	0.008353	6.43E-04

2. From 1 to approximately 100 erlangs, the hybrid scheme has better performance in terms of handoff call dropping but worse in new call blocking. This is because the actual number of channels available to new calls is smaller while the number of channels available to handoff calls is greater, in other words, it is the effect of exclusively channel reservation or handoff calls.
3. There are two critical points for Class I new call blocking probability and the formulated value. The Erlang model performs better than the hybrid scheme until the load reaches the first critical point of about 120 erlangs. From this load to the second critical point (about 350 erlangs), the new scheme has a slightly better performance. Above 350 erlangs, the simulated values are very much the same.
4. The probability that calls handed off from adjacent cells are dropped due to lack of resource is low compared to the formulated probability. The difference of the probabilities is from 0.1 to 0.2 depending on the range.
5. To the user's point of view, it is harder to make a new Class II call than in normal situation (Erlang B model). For Class I call, users have approximately the same experience (when the load is greater than 100 erlangs). However, their handoff calls from neighbour cells are admitted easier.

The following figures were taken at the same program run time. We have made some comments as well as explanation in each observation.

New Calls

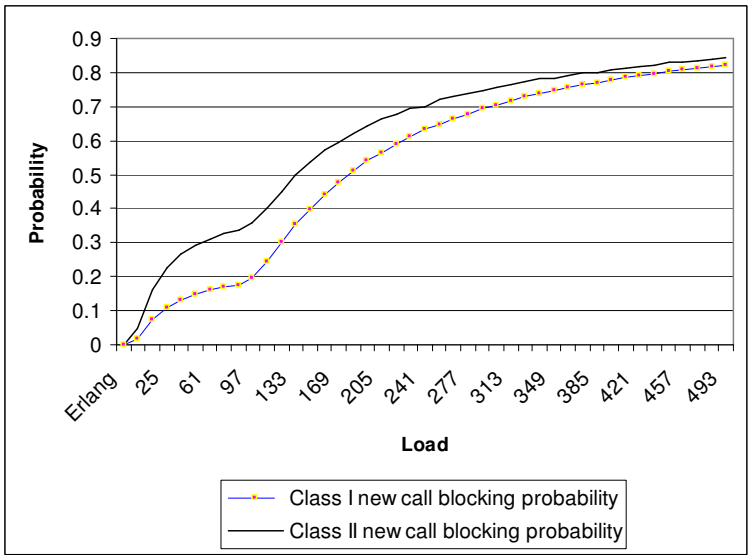


Figure 3-14: New call blocking probabilities

Figure 3-14 shows the probabilities that new calls are blocked in the cell. Note that Class I new calls do not experience blocking as bad as Class II new calls. This is because Class I new calls are still admitted in the Priority Access state while Class II new calls are blocked.

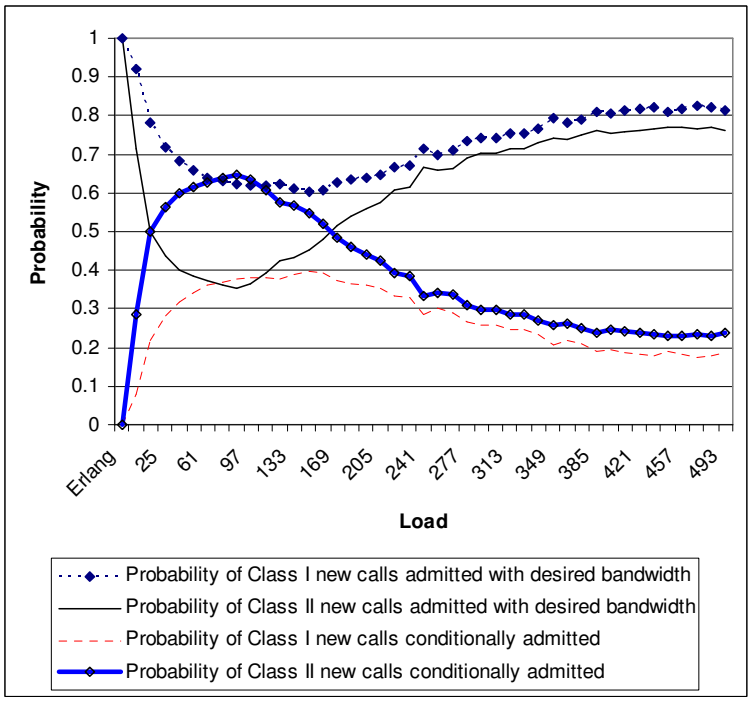


Figure 3-15: New calls admission

Figure 3-15 shows the how new calls are admitted.

- At low load (less than 60 erlangs), more Class I calls are admitted with desired bandwidth. Later when the load is higher (from 60 to 200 erlangs), more of them are conditionally admitted. When the load is very high (greater than 200 erlangs), more

Class I calls are admitted with desired bandwidth. As can be seen in Figure 3-12, at this stage, the new call blocking probability is very high. Very few new calls are admitted although they have higher chances to be admitted with desired bandwidth.

- A similar situation occurs for Class II calls. At low load (less than 25 erlangs), more of them are admitted with the desired bandwidth. From 25 to 180 erlangs, more calls are admitted with condition. But for load greater than 180 erlangs, there are not many Class II new calls are admitted (as in figure 1-8). If they are admitted, they have their desired bandwidth.
- When the load increases, the admission of Class I new calls is not affected as much as that of Class II new calls. The change of both Class I curves is not as significant as Class II curves.
- At very high load, the admission of both classes tends to be the same: more calls are admitted with desired bandwidth and less are conditionally admitted.

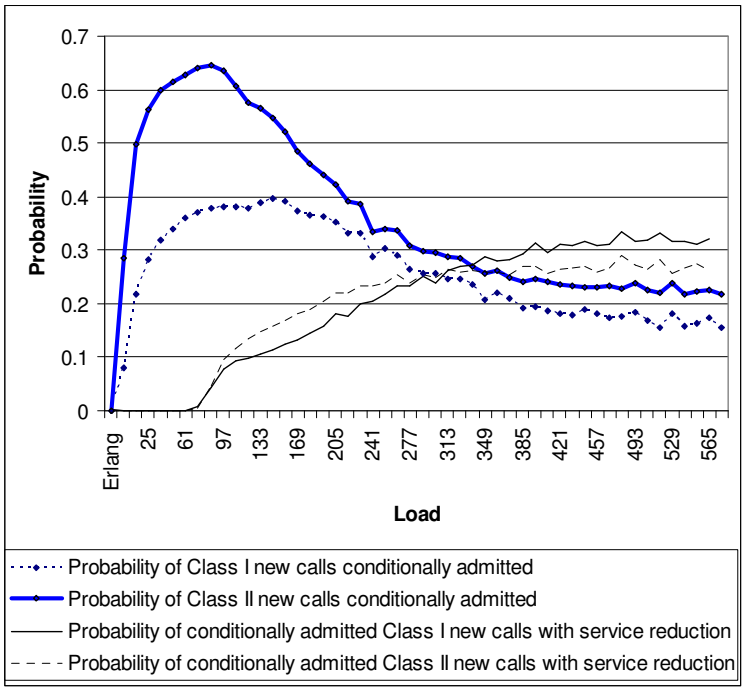


Figure 3-16: Conditionally admitted new calls with service reduction

Figure 3-16 shows how conditionally admitted calls are treated. At low load, although there are many calls admitted with the condition that their service may be reduced; the calls actually maintain their QoS for their duration. When the load increases to about 60 erlangs, the service of those calls are reduced to accommodate bandwidth for other higher priority calls.

Handoff Calls

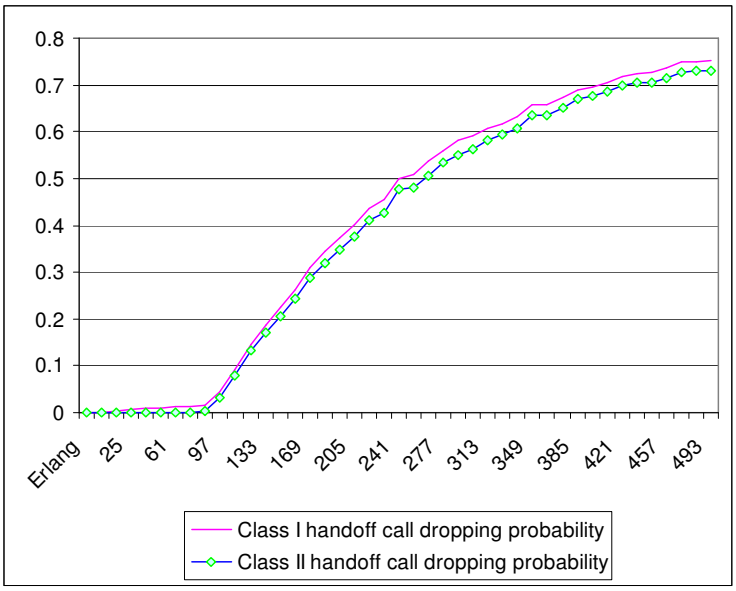


Figure 3-17: Handoff call dropping probabilities

Figure 3-17 shows that the dropping probability of Class I handoff calls and Class II handoff calls are very much the same. Class I handoff call dropping probability is slightly higher because those calls require more bandwidth and it is generally more difficult to allocate a great amount of bandwidth.

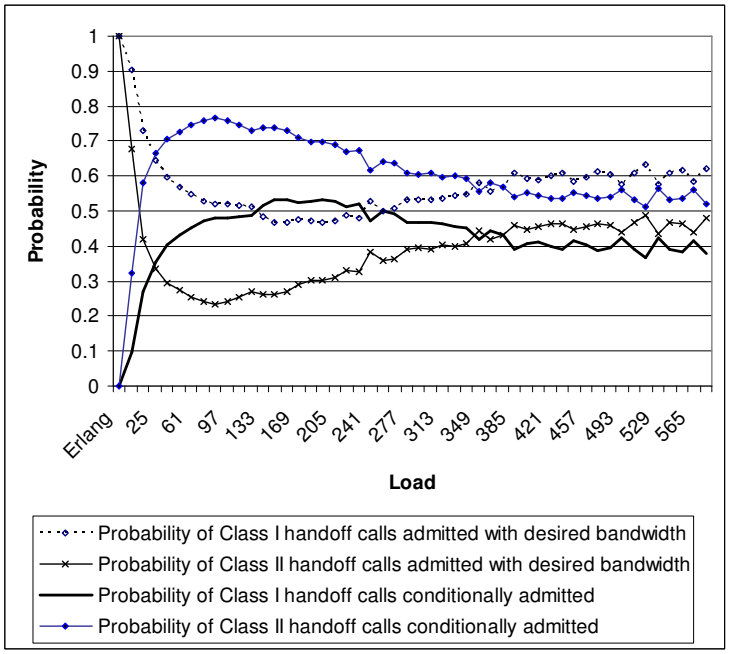


Figure 3-18: Handoff call admission

Figure 3-18 shows how handoff call requests are handled.

- At low load, most handoff calls are admitted with desired bandwidth. The admission rapidly changes to conditional admission.
- When the load is from 25 to 240 erlangs, more calls are admitted with condition. Class II handoff calls are more likely to be admitted with condition than Class I calls. Class II calls have about 70% chances to be admitted with condition while only half of Class I calls are in the same situation.
- At very high load, about half of the calls are admitted with desired bandwidth. Class I handoff calls have a higher chance to get their desired bandwidth than Class II calls.

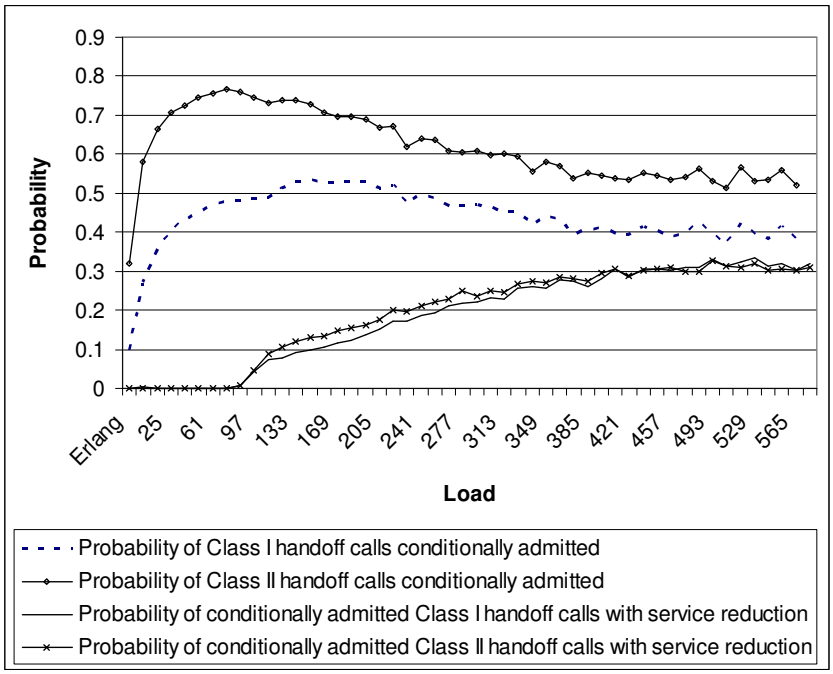


Figure 3-19: Conditionally admitted handoff calls with service reduction

Figure 3-19 shows how conditionally admitted handoff calls are treated. The situation is similar to that in new calls. At low load, more calls are admitted with the condition that their service may be reduced, but most of them maintain their QoS for their duration. When the load increases to about 97 erlangs, the service of those calls are reduced to accommodate bandwidth for other calls.

3.5 Summary

In this chapter, we started with presentation of the well known Threshold Access Sharing scheme for admission and handoff control. Our analysis and observation of TAS has shown that there is space for further improvement on the scheme, especially its handoff handling. A rate based borrowing scheme was found having many features which can be further explored and adopted in TAS. Our work started with an improved TAS (iTAS) on classification, call

prioritisation and new call and handoff call handling. Then a modified rate based borrowing scheme is integrated into iTAS. The new proposed scheme aims to lower handoff call dropping probability and to maximise the resource utilisation. Handoff calls are always admitted with their desired bandwidth level. The scheme tries to ensure that the application on the terminal will not experience a sudden change in bandwidth allocation. It works on the basis of reserving a fixed amount of bandwidth for handoff calls. To gain more bandwidth to admit those prioritised calls, the system can "borrow" bandwidth from other calls. In the design of the scheme, the complexity of implementation is also considered. Compared with original rate based borrowing scheme, the new scheme is much simpler, but with its all the main features kept and strengthened. Our simulation has shown that the new scheme has outperformed the original TAS in terms of handoff prioritisation and handling. The new scheme can be a candidate for the future wireless IP networks, in which the large variety of applications will justify the classification of services and rich multimedia applications with scalable bandwidth requirement will support the concept of rate borrowing, or in another words, the adaptive bandwidth allocation.

Chapter 4: NOVEL WEIGHT-BASED (WB) ADMISSION CONTROL IN HIERARCHICAL CELLULAR NETWORKS

In this chapter, we are going to propose a novel weight based admission control for cellular IP networks. First we will discuss the admission criteria. The scheme is described in details with help of example scenario. The scheme has considered multiple factors in the admission decision making. Our simulation has demonstrated that it has obvious good performance in handoff handling.

4.1 Weight-based admission control algorithm

4.1.1 Admission Control Criteria

Current admission control algorithms use one or two criteria to admit or reject a call request. In this model, we aim to take into account as many criteria as possible. At any certain time, the call connects to a base station of a cell, which is often referred as the serving cell. If the terminal is moving, the call will have a target cell, where it will be handed off to. We identify that there are three main areas directly affecting the admission decision: the call specifications, the serving cell and the target cell.

We are going to use this concept in a hierarchical network with two tiers: the macro-tier combined of macrocell for fast-moving terminals and the micro-tier overlapped, combined of slow-moving or stationary terminals.

4.1.1.1 Call specifications

Call characteristics

In our proposed scheme in previous chapter, at the setup phase, a call must specify its characteristics, i.e. the traffic class, the call type and the bandwidth loss tolerance.

Call type and traffic class: There are two types of calls: voice and data which are all based on digital technology. Voice calls require less bandwidth and do not have strict error rate but they are sensitive to delay. Data calls are categorised into two classes: Class I and Class II. Class I calls have real-time applications and have more priority over the other class with non-realtime applications. Voice calls are considered in Class II. This is the only criterion that does not change during the call duration.

Call nature: Calls started by users are the new calls. Calls handed off from adjacent cells are handoff calls.

Bandwidth loss tolerance: The model uses two bandwidth levels: the desired bandwidth and the minimum acceptable bandwidth. The QoS requirement is not maintained when the call bandwidth usage is reduced from the desired level to the minimum acceptable level. The difference between the desired bandwidth and the minimum acceptable bandwidth is the bandwidth loss tolerance. Calls with less bandwidth loss tolerance should have higher admission priority because they are more likely to lose the connection.

Movement prediction

It is possible to detect the movement of a call and predict its future direction. There are three possible measurements in movement prediction: the direction, the speed and the distance before the next handoff.

Direction: Simple reservation schemes do not care about the direction and make reservation in all adjacent cells e.g. six cells in a hexagonal cell system. This reservation is wasteful because only 1/6 reserved bandwidth is used. Being aware of the direction, the system can efficiently reserve bandwidth for handoff calls in the target cell (the adjacent cell at that direction). Bandwidth in other five adjacent cells can be used for other calls. This is much more efficient than the simple schemes. Geographic information can be useful in this case. For example, users in cars on a one-way highway will mostly have their calls handoff in the target cell ahead.

Speed: Speed is important to select the tier. Fast moving calls should stick with the macro-tier while slower moving calls should be controlled by the micro-tier. In our model, if calls travel faster than 20 km/h, they are considered high mobility; otherwise their speed is low mobility.

Distance: Current mobile systems can produce the exact physical location of a terminal. The model only needs to know the distance between the current locations to the cell border, where the call will handoff (at the cell border). This estimated measurement will be used with the speed to calculate the time before the call needs a handoff. The system then makes the reservation in the target cell. It is important to make bandwidth reservation at the right time. Reserving bandwidth too early is not efficient because the bandwidth can not be used for other calls once reserved. Reserving bandwidth too late may result in dropped calls.

User profile

Handoff prediction can be based on user's moving habits. Billing records can be statistically analysed extract information about the calling habits, the duration and the call types at certain time of the day. For example, a commuter on a train to work often use his/ her PDA to read news, check personal emails or stocking portfolio rather than make voice calls.

Signals and power

These measurements can be obtained directly from the mobile terminal. The terminal continuously adjust its power to maintain a strong signal strength (hence an adequately good connection) as well as to maximise their battery life. Many admission control schemes set off the handoff mechanism when the signal strength (SS) falls below a threshold. The model defines SS as strong or weak. Calls with weak signal strength have higher admission priority.

A strong signal does not guarantee a good connection if the environment has too much noise. Noise can come from the environment such as electrical surge, lightning, interference or from the equipment itself, such as thermal noise. Repeaters in analogue systems and regenerators in digital systems are responsible to boost the weak signal and reduce the noise signal to minimum. In mobile systems, if the terminal experiences a highly noisy environment, it tries to raise the signal strength to maintain the required signal-to-noise ratio (SNR).

4.1.1.2 Serving cell

In simple schemes, the load in the serving cell determines the admission decision. The load often relates to the number of ongoing calls or the bandwidth availability. If the load in the serving cell is high and the call is going to handoff, it is preferable that the handoff happens as soon as possible to free the local bandwidth. If the load in the serving cell is low, the call can stay a little bit longer.

Another the factor is the statistical data of call behaviours. Call behaviours, especially handoff occurrence, can help to make admission decision. For instance a highway traverses across a cell from South to North. Calls originated in that cell will hand off more frequently because users travel at high speed. The handoff target cells are the Northern and Southern cells. Referring to Figure 4-1, most calls in cell B will handoff to cell A or cell C.

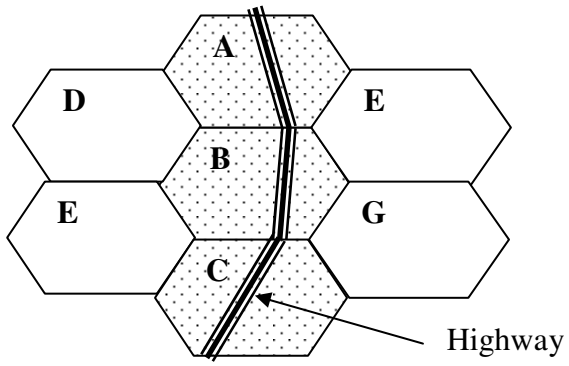


Figure 4-1: Handoff behaviours in a cell

This difference between this criterion and the user profile is the scope of the information. In the user profile, the information is specific for the user while this is more systematic, covering the information for a cell.

4.1.1.3 Target cell (handoff call admission only)

Similarly, the load of the target cell is important to admission decision. The bandwidth availability is the bottleneck. If the bandwidth is completely used, the handoff calls attempt to use reserved bandwidth. If the reserved is not available, the call is dropped. The target cell has its own bandwidth loss tolerance, which is the sum of individual calls' bandwidth loss tolerance.

4.1.2 A Weight Based Admission Control Algorithm

Table 4-1 summarises the criteria used in the algorithm. They include the call's specifications, the characteristics of the serving cell and target cell. Each criterion is assigned a weight. The weight is chosen based on the importance of the criteria to the admission decision. Each criterion has two options; each has its own weight. The "heavier" option is prioritised over the other.

For example, the signal strength and signal to noise ratio (SNR) are the most important factors. Their weights are 5 (highest). Whereas the load in the current cell does not make much significant effect on the decision and it has the weight of 1. The signal strength has two options: weak and strong. We aim to give priority to the call request with more weight, so a handoff request with weak signal strength, which obviously needs more attention, is assigned a weight of 2 while the one with stronger signal has the weight of 1.

In Table 4-1 below we select a number of factors that theoretically have some effect on the admission control. The criterion weight and option weight are chosen in the purpose of simulating the weight-based algorithm. More research is required to obtain an appropriate value.

Call's criteria

The value of the traffic class remains unchanged for the call duration while the call nature starts as a "new call" then "handoff call" until it completes. The call handoff habit information is extracted from a database in the current cell's base station. This information is updated when the call finishes.

The values of other call's criteria (BLT, SS, and SNR) change with time. A call's BLT changes depending on the allocated bandwidth. The SS and SNR values are continuously monitored by the mobile terminal.

Table 4-1: Weight-based admission algorithm

Criteria	Criterion weight		Options	Option weight	Explanation
	Handoff	New			
Call's signal strength (SS) on traffic channel	3	0	Weak	2	This applies to handoff calls only (new call requests do not have a traffic channel yet). Calls with weak signal strength should be handled with priority.
			Strong	1	
Call's signal to noise ratio (SNR) on traffic channel	3	0	Unacceptable	2	Similar to the signal strength, only applied to handoff calls. A call with unacceptable SNR has more priority.
			Acceptable	1	
Call's traffic class	3	3	Class I	2	Class I calls should be treated with more priority.
			Class II	1	
Target cell's load	2	0	Low	2	If the load in the target cell is high, the call should not be handed (because it may be dropped later). The priority is given to target cells with low load.
			High	1	
Target cell's BLT	2	0	Low	2	If the BLT in the target cell is low, the call should be admitted as soon as possible to ensure an uninterrupted connection.
			High	1	
Call's BLT	1	2	Low	2	Calls with low BLT should be admitted with priority. New calls with low BLT results less borrowing, hence less overhead.
			High	1	
Call behaviours in current cell	1	0	Frequent	2	If the historical records in the cell show frequent handoff, the handoff request should be admitted.
			Seldom	1	
User's handoff habit	1	0	Frequent	2	If the user profile shows the user frequently handoff in this location, the admission is treated easier.
			Seldom	1	
Current cell's load	0	3	Low	2	The load in the current cell is important to new calls but does not have any effect on handoff calls. Low load cell attracts more calls.
			High	1	
Call nature			Handoff	2	A new call request voids the weights of the signal criteria and the target cell to zero. This lowers the total weight. This criterion is used for bandwidth allocation only.
			New	1	

Cell's criteria

A cell observes its own load. Ongoing calls also provide their BLT on request to the base station. The total of the calls' BLT is the cell's BLT. Because individual BLT changes, the value of the cell's BLT changes as well.

The cell also has a database containing handoff information for terminals. The database structure is simple. The admission control is especially useful in high load situation, which happens in peak hours only. The database records the handoff habit of a number of mobile terminals in the cell during peak hours.

Table 4-2: Handoff information database for a cell

Mobile ID	Peak hour 1	Peak hour 2	Peak hour 3	…
IMSI 1	12	5	17	
IMSI 2	23	1	0	
IMSI 3	4	2	10	
…				

The value is incremented every time a handoff occurs. The call is said to be a "frequent" handoff if the value is higher than or equal to a threshold, otherwise it is "seldom". Periodically, the system resets all values. Values greater than the threshold (i.e. in this cell, the call is frequently handed off in this peak hour) will be set to equal the threshold. This ensures the call's handoff habit will be still considered as "frequent". Values less than the threshold are set to zeros. The threshold value depends on the time length of the recording. For example, if the database keeps information for one month (30 days), the threshold value could be $\frac{30}{3} = 10$. It means in a certain peak hour, if the call is handed off from this cell 10 times a month, it will be considered frequently handed off.

Algorithm description

When receiving a connection request, the system uses the values of the criteria to calculate the weight of the request. Each criterion is assigned a criterion weight based on its importance to the admission decision. The assigned weight is from 0 to 3, where 3 means that the criterion is very essential to make decision and 0 means the criterion does not affect the decision. In each criterion, there are two options; one should be treated with more priority than the other. The prioritised option has the higher weight, in this case is 2. The other option has lower weight of 1. The criteria's roles are different depending on the type of calls. For example a new call request does not have any target cell information; hence it does not care; whereas a handoff call request must enquire many parameters from the target cell. The summary is shown in Table 4-3.

Table 4-3: Admission weights

Criteria	Criterion weight		Option weight		Weight	
	Handoff (a)	New (b)		(c)	Handoff (a x c)	New (b x c)
Signal strength	3	0	Weak	2	6	0
			Strong	1	3	0
Signal-to-noise ratio	3	0	Unacceptable	2	6	0
			Acceptable	1	3	0
Call's traffic class	3	3	Class I	2	6	6
			Class II	1	3	3
Target cell's load	2	0	Low	2	4	0
			High	1	2	0
Target cell's BLT	2	0	Low	2	4	0
			High	1	2	0
Call's BLT	1	2	Low	2	2	4
			High	1	1	2
Call behaviours	1	0	Frequent	2	2	0
			Seldom	1	1	0
Users' habit	1	0	Frequent	2	2	0
			Seldom	1	1	0
Current cell's load	0	3	Low	2	0	6
			High	1	0	3

Considering Table 4-3, handoff calls have more criteria to consider than new calls. Therefore their weights are greater than those of new calls. In fact, the possible maximum weight is 32 corresponding to a handoff call heading to an immediate handoff. The minimum weight is 8 for a new call. Other call requests take value in the range of 8 and 32. The minimum weight of a handoff call request is 16, equal to the maximum weight of a new call request. The weights of Class I and Class II calls are different by 3. Table 4-4 displays the observation.

Table 4-4: Maximum and minimum weight

		Weight	
		Class I	Class II
Handoff calls	MAX	32	29
	MIN	19	16
New calls	MAX	16	13
	MIN	11	8

The algorithm aims to prioritise the admission based on different traffic classes as well as call natures. Handoff calls have more priority than new calls, Class I calls have more priority than Class II. Hence the priority order is:

1. Priority level 1: Class I handoff calls
2. Priority level 2: Class II handoff calls
3. Priority level 3: Class I new calls
4. Priority level 4: Class II new calls

The four admission levels are identified. The concept of improved Threshold Access Sharing is reused.

Next, we consider how to define the admission thresholds. Figure 4-2 visualises the information in Table 4-4. Each priority level has an admission area. Some areas are overlapped.

The three utilisation thresholds: T_1, T_2 and T_3, corresponding to three weight levels W_1, W_2 and W_3, respectively. The weight levels are shown in Figure 4-2. The first threshold T_1 is the average of the minimum weight of Class I handoff calls and the maximum weight of Class II handoff calls. The second threshold T_2 is the overlapped boundary of handoff calls and new calls. The last threshold T_3 is the average of the minimum weight of Class I new calls and the maximum weight of Class II new calls.

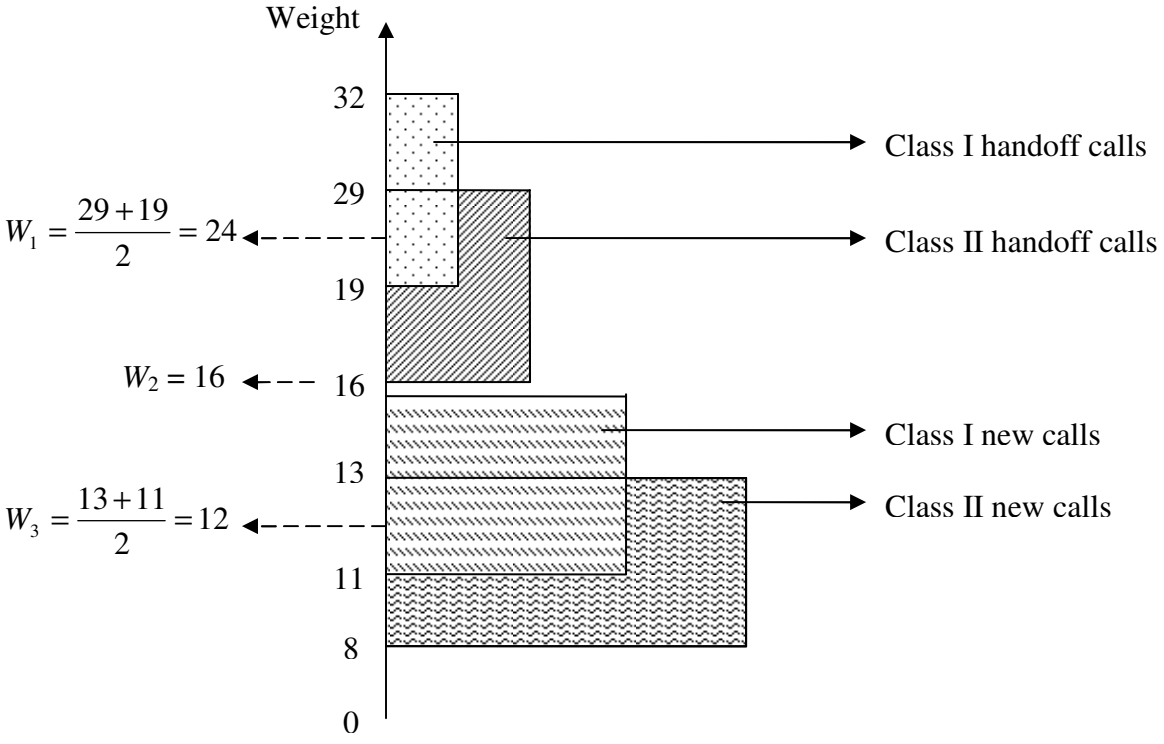


Figure 4-2: Define admission thresholds

There are four access areas directly related to the calls' weights. The admission decision is in Table 4-5. For examples, area 1 accepts only calls with weights greater than 24 i.e. Class I

handoff and some Class II handoff calls with weight from 25 to 29. The second area accepts all handoff calls (weights greater than 16). The third area admits calls with weight greater than 12, which includes all handoff calls and some new calls. The last area accepts all call regardless of their weights.

Table 4-5: Access areas and corresponding weights

Access area	Weight	Full admission
1	$> W_1$ (=24)	Admit Class I handoff and some Class II handoff calls (with weight from 25 to 29)
2	$> W_2$ (=16)	Admit all handoff calls
3	$> W_3$ (=12)	Admit all handoff calls, some Class I new calls (with weight from 13 to 16) and some Class II new calls (with weight of 13)
4	$\leq W_3$ (=12)	Admit all calls regardless weights

In the next page, Figure 4-1 shows the thresholds and the access areas.

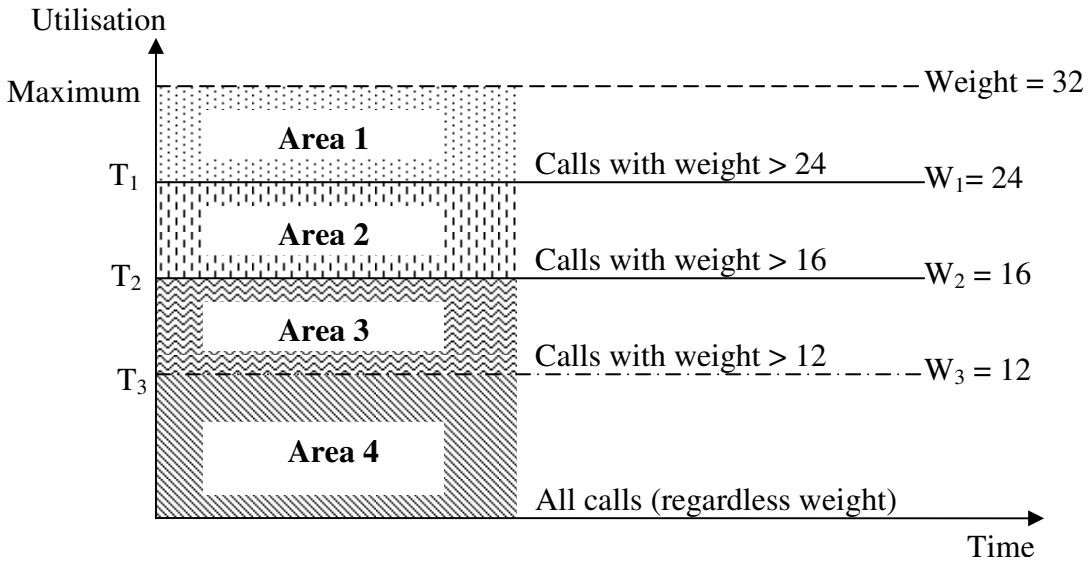


Figure 4-3: Weight-based threshold access sharing

The admission algorithm is described in the following steps:

1. The base station receives a call request.
2. It asks for necessary information and composes the weight of the request.
3. The call tries to use bandwidth in the lowest area ($i = 4$). If there is enough bandwidth, the request is admitted.
4. Otherwise, the request compares its weight to the previous threshold W_{i-1}. If the weight is greater, the call will try Area i – *1* (back to step 3). If this is the last area, the request is denied.

5. If its weight is not enough to proceed, it will compare the Bandwidth Loss Tolerance of the cell (BLT_{cell}) and its minimum acceptable bandwidth (m_{call}). If the BLT_{cell} is greater than m_{call}, the cell will reduce service of some current calls to admit the request. Otherwise the call request is denied.

There are some assumptions:

1. Calls admitted are subject to have their service reduced if necessary. A call can request an un-reducible service by giving the same value for its desired bandwidth (M) and minimum acceptable bandwidth (m). In this case: $BLT = M - m = 0$
2. The system uses the simplified rate-based borrowing scheme described in Chapter 3 to select calls with highest BLT to reduce the service first.

A flowchart of the operation is shown in Figure 4-4.

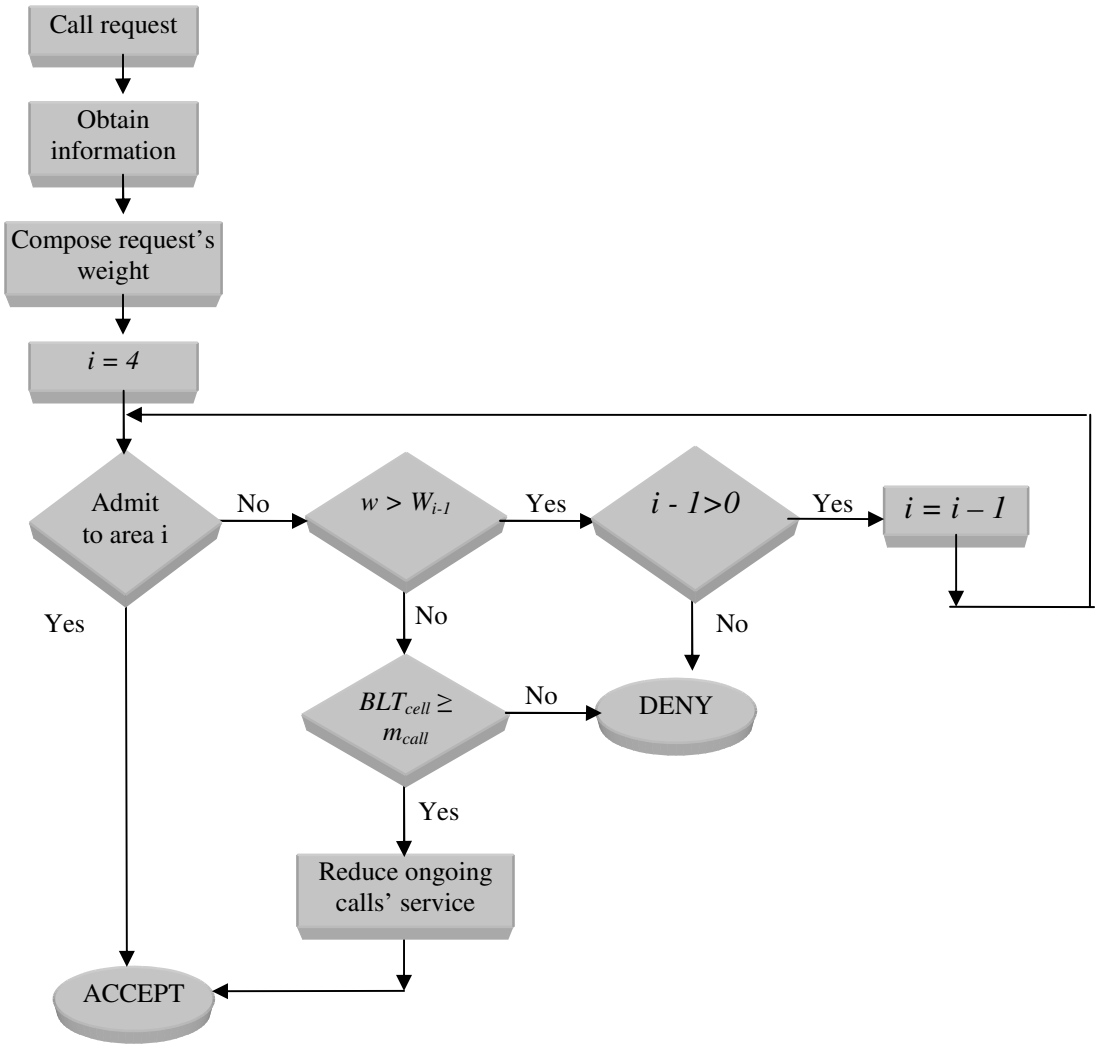


Figure 4-4: Weight-based admission algorithm flowchart

To understand how the algorithm works, the following examples can be considered. In each example, there is a call request. The weight of the request is estimated and the admission decision is made accordingly.

Example 1: There is a class I handed off from a neighbour cell; the expected weight will be in the range of 19 to 32. Its characteristics are listed in Table 4-6:

Table 4-6: Example 1 - Class I handoff call

Criteria (weight)	Classification (scalar)	Calculated
Signal strength (3)	Weak (2)	3 x 2 = 6
SNR (3)	Acceptable (1)	3 x 1 = 3
Traffic class (3)	Class I (2)	3 x 2 = 6
Load in target cell (2)	Low (2)	2 x 2 = 4
BLT of target cell (2)	High (1)	2 x 1 = 2
BLT of the call (1)	Low (2)	1 x 2 = 2
Call behaviour (1)	Seldom handoff (1)	1 x 1 = 1
User habit (1)	Seldom handoff (1)	1 x 1 = 1
	Total	**25**

The total weight is 25, which is greater than the first weight threshold W_1 (= 25). The admission decision is that this call can access all areas. The system checks the availability in Area 4 to Area 1 respectively and admits the call. If there is not enough bandwidth, the system will reduce service of some calls with high BLT and allocate the free bandwidth to the call request.

Example 2: A user tries to initiates a new call. The expected weight of the call is from 8 to 13. The characteristics are:

Table 4-7: Example 2 - Class II new calls

Criteria (weight)	Classification (scalar)	Calculated
Traffic class (3)	Class II (1)	3 x 1 = 3
BLT of the call (2)	High (1)	2 x 1 = 2
Load in current cell (3)	Low (2)	3 x 2 = 6
	Total	**11**

With the weight of 11, this new call will be admitted to Area 4 (the lowest priority) if there is enough bandwidth. If there is not enough bandwidth, a service reduction process is performed on a number of high BLT calls. If none of the calls can spare bandwidth, the request is rejected.

4.1.3 Simulation results

Simulation setup

We used the same simulation environment of the Improved Threshold Access Sharing (iTAS) to visualise the effect of the Weight-Based algorithm (WB). The number of channels is doubled to 180. The performance is observed with the load of 1 erlang to 2000 erlangs.

Table 4-8: Number of channels allocated in each area

Access area	Bandwidth allocated	Number of channels
1	13.3%	24
2	13.3%	24
3	13.3%	24
4	60.01%	108

Each call request has a number of characteristics which is set arbitrarily. Those include the signal strength, the signal to noise ratio, traffic classes, the load in the target cell, the bandwidth loss tolerance of the target cell and of the call, the user's handoff habits, the handoff pattern in the current cell and the load of the current cell. Each criteria is defined as a Boolean variable as summarised in Table 4-3. For example, weak signal strength has the weight of 2 while strong signal strength has 1. Then this weight is multiplied with the weight of the criteria, in this case is 3. Except the load of the current cell, other criteria are constant when the call request is received.

Observations

Technically this scheme is novice and incomparable to any existing schemes. The following observations are meant to be references only.

First of all, the algorithm is compared against Erlang B formula. The result in Figure 4-5 shows that the new call blocking probability is very high compared to the simple scheme (i.e. the probability of request rejected, estimated by Erlang B formula). However the handoff call dropping probability is lower than the Erlang B formula values. This somehow satisfies the goal of prioritising handoff calls.

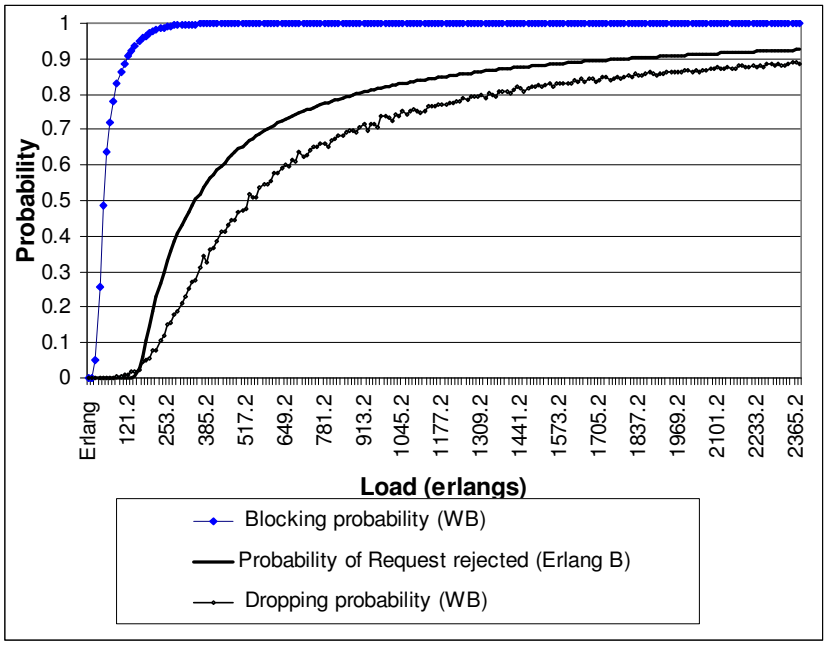


Figure 4-5: WB new call blocking and handoff call dropping probability

In Figure 4-6, the effect of the algorithm in each traffic class is considered closely. At all time, the probability of a Class I call request being rejected is always lower than that of Class II. In high load situation, the dropping probability of Class I calls is about 0.05 less than Class II and they are always lower than the rejection rate estimated by Erlang B formula. In the other hand, the blocking probability of new calls is too high that it is nearly impossible to make a new call when the load reaches 250 erlangs.

Using a weight-based approach in admission control is novel. Its performance can not be analysed against a similar weight-based algorithm. Therefore we compare it with the improved Threshold Access Sharing scheme, discuss in chapter 3.

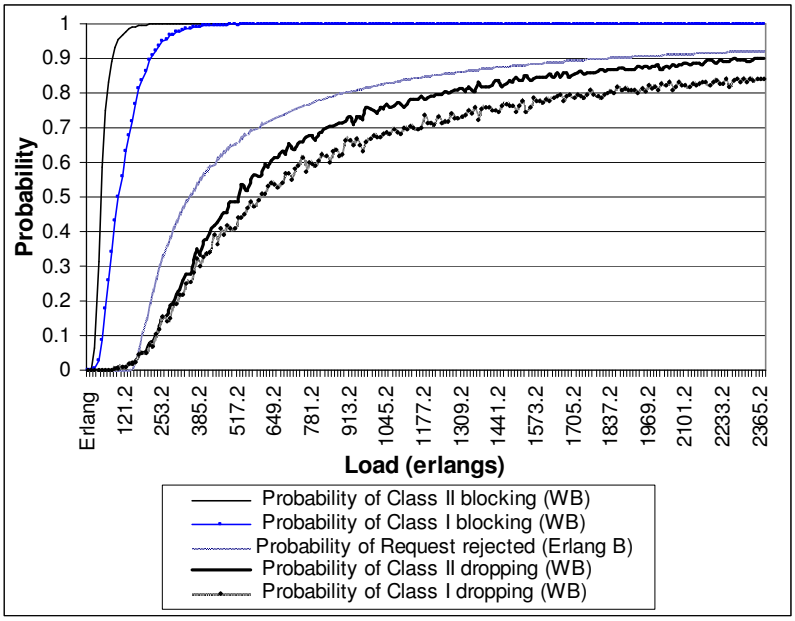


Figure 4-6: Blocking and Dropping probability in each class

The result in Figure 4-7 shows that the blocking probability of new calls in Weight-based (WB) scheme is high compared to in iTAS. At high load, the difference in the blocking probability of new calls in iTAS and in the simple Erlang B scheme is acceptable. However in low load situations, both schemes give high blocking probability compared to the Erlang B scheme.

The dropping probability of handoff calls in iTAS scheme is lower than in WB scheme. When the load is very high (above 1000 erlangs) the increase of the dropping probability in iTAS scheme is insignificant; whereas, WB dropping probability steadily increases with load.

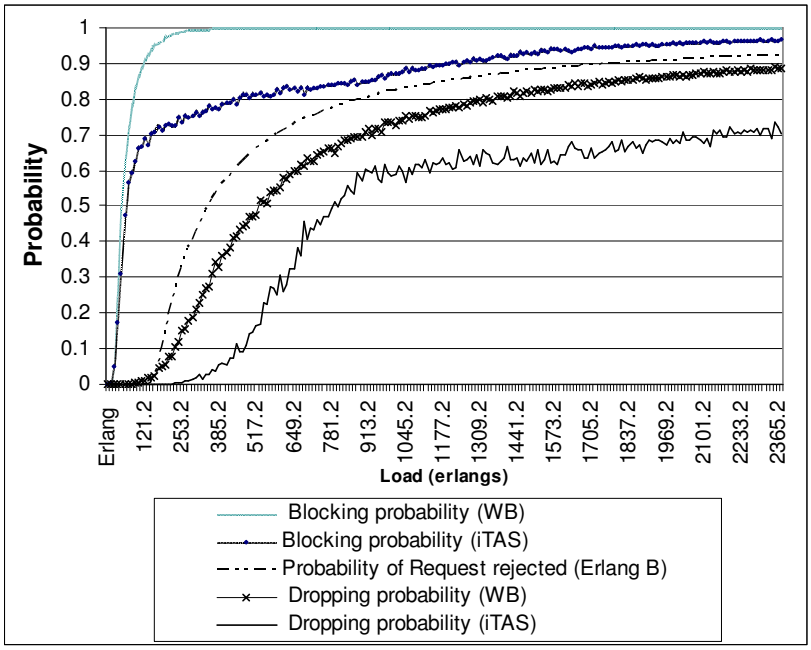


Figure 4-7: Comparison between iTAS and WB

The traffic class of call request is considered in the next part.

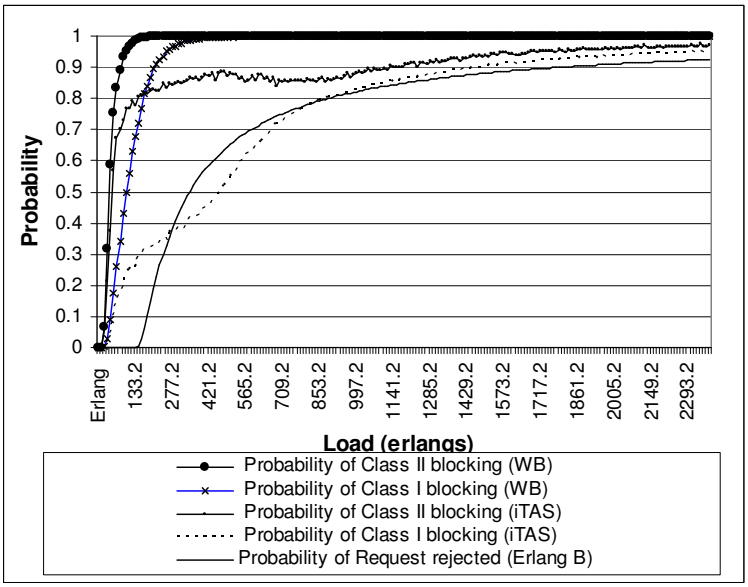


Figure 4-8: Blocking probability (iTAS vs. WB): Traffic class consideration

Figure 4-8 shows how iTAS and WB schemes behave with the consideration of traffic classes. In both cases, Class I new calls are treated with higher priority hence their blocking probability is less.

Handoff call admission is considered in Figure 4-9 below. WB scheme clearly gives Class I handoff calls more priority than Class II. On the other hand, traffic class does not play an important part in admitting handoff calls in the iTAS scheme, for which, the Class I call dropping probability is in fact slightly greater than Class II call dropping probability. However the difference is trivial.

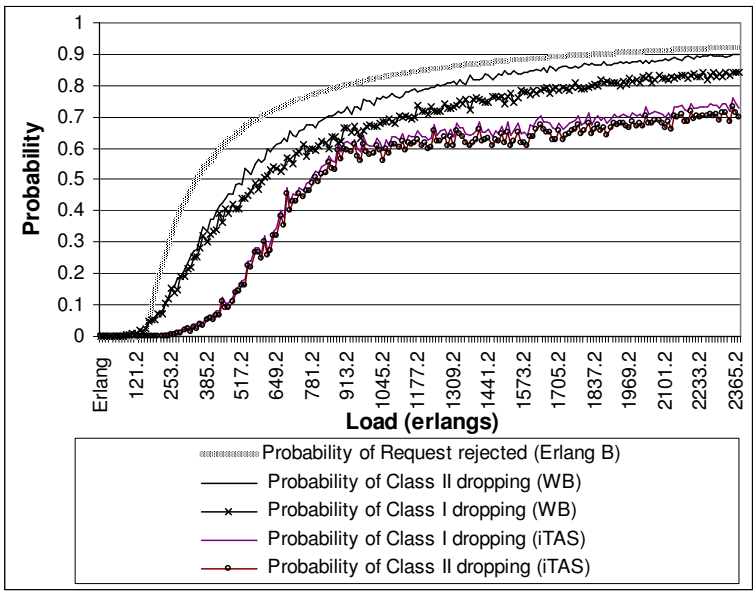


Figure 4-9: Dropping probability (iTAS vs. WB): Traffic class consideration

Another observation is the difference between conditional admissions in two schemes. In iTAS, some calls are admitted with condition that their service may be reduced if the system requires. Other calls are fully admitted i.e. their service is guaranteed throughout the call life. In WB scheme, all calls are subject to have their service reduced if necessary. It is possible to apply full admission and conditional admission in WB.

4.2 Summary

Admission control's primary concern is to maximize the efficiency of the resource utilisation and the quality of service to users. There are many factors which could affect the performance of an admission control scheme. In this chapter, we have proposed a weight based admission control scheme. In this scheme, multiple criteria are used in the admission decision making. All the possible criteria are listed, analyzed and weighted. The admission of new calls and handoff calls are based on the weights of all criteria introduced. The main achievement in WB scheme is that it gives consideration to multiple factors which have impact on the admission decision. These factors are hard to be modelled in the traditional admission models. Their impacts are not reflected in the abstract models normally used in analysis. Somehow direct comparing WB scheme with other schemes is not quite in the same ground as the latter are much simplified abstract models. But still our simulation has demonstrated that this scheme yields better performance in terms of handoff call dropping probability compared with iTAS. However its new call blocking probability has suffered due to the trade off with handoff calls.

In the weight based admission control, the nature of diversified applications in mobile IP networks has been considered. This is different from the traditional mobile phone networks in which voice is basically the only service offered. Also mobility prediction, user profile, user

call patterns, user call habits, signal strength, geological location of calls and other statistical information all can be referred in the admission process. All those information are especially useful in handoff handling. Our simulation result has demonstrated that point. The weight assigning in the scheme can be further explored as it directly affects the admission decision. Our initial proposal is based on the limited analysis and experiments.

After studying improved Threshold Access Sharing and Weight-based as the admission control schemes applicable at the base station, we are going to expand our work to an admission control model applicable to a system.

Chapter 5: ADMISSION CONTROL MODEL FOR HIERARCHICAL CELLULAR NETWORKS

The improved Threshold Access Sharing and the weight-based admission control schemes discussed in previous chapters focused on the call admission in cell level. The principles and concept can be used at typical cellular networks such as current GSM or TDMA systems. However, cellular IP networks in hierarchical structure offers more space for further improvement on the efficiency of network resource utilization and the quality of service. In this chapter we will investigate on how the current admission schemes can be applied in hierarchical networks. New modelling will be proposed to improve the handoff call admission with traffic class consideration.

5.1 Admission control in hierarchical cellular IP networks

Hierarchical cellular networks were briefly introduced in section 1.4.3. A general hierarchical cellular network has many overlapped tiers. A popular arrangement is a satellite tier with the largest cell size, a macro-tier with the second largest cell size, a micro-tier and a pico-tier with the smallest cell size. The cells are named after their tiers, e.g. satellite cells, macrocells, microcells and picocells. Hierarchical structure is aimed to provide the better performance in term of bandwidth re-use efficiency and handoff execution [5, 9, 10].

Hierarchical structure increases the overall system capacity by having many smaller cells and to keep the number of handoff events to minimum by assigning fast moving calls to larger cells. There are two phases in the admission process: tier selection and cell selection. The tier is selected based on the moving speed of the terminal. If a fast moving terminal attaches to a microcell, it will experience many handoffs during its life. If that terminal sticks to a high level cell, covering a larger area, the number of handoff events will be less. Once the tier has been identified, a corresponding cell is selected.

Our discussion in this chapter will be limited to the two-tier networks only. The results achieved can be expected to apply to networks with more than two tiers. The tier with larger cells is called the macro-tier and the one with smaller cells is called the micro-tier. Handoff in this type of networks is classified to vertical (between tiers) and horizontal (between cells).

The traditional cellular admission control has to be adapted to the characteristics of hierarchical networks. Handoff call requests and new call requests are handled in different

ways but the basic operation is the same. First, we will look at the types of handoffs in a hierarchical network. Then we will look at the generic admission control in a 2-tier network.

Types of handoffs

There are two different types of handoffs. A horizontal handoff is more popular and common in cellular networks. It happens when the serving cell (where the call is current at) and the target cell (where the call is moving to) are in one tier. A vertical handoff occurs when the serving cell and the target cell are in different tiers. Vertical handoffs are distinctive to hierarchical and they happen when the terminal changes its speed only.

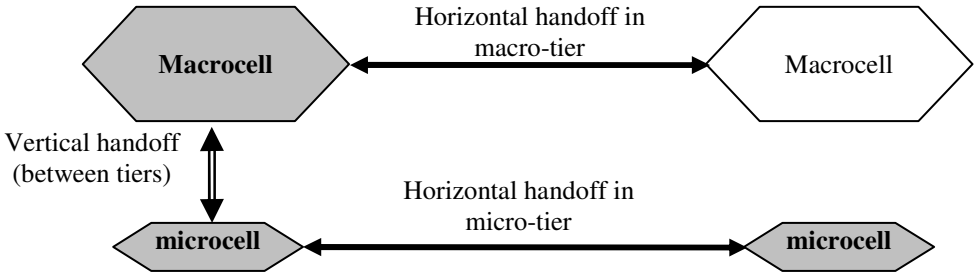


Figure 5-1: Handoff types in 2-tier hierarchical networks

New call and handoff call admission control

New call admission control occurs when an "idle" terminal attempts to make a call while handoff call admission control happens when a handoff is required. In this specific case, vertical handoffs are considered. The following steps are the procedure that the system and the terminal need to do in the admission control.

First of all, the base station checks the resource in the current cell (camped-on cell if this is a new call request; or serving cell if this is a handoff request) and the target cell (where the call is handed off to). Depending on the characteristics of the call and the cell (such as traffic class, required bandwidth), a QoS requirement is negotiated between the call and the cell. This includes estimated bandwidth allocation.

The next step involves the admission decision. For a new call, the serving base station of the terminal will make the decision. If it does not have enough bandwidth, it will borrow from existing calls. If no bandwidth is available even after borrowing, the call will be blocked. Note that the decision only depends on conditions in the current tier of the terminal.

In other hand, if the connection request is for a handoff call, the speed of the call is checked and if there is a speed change, the call is handed off vertically i.e. to different tier. There is an exception for calls in the micro-tier. If none bandwidth is available for them even after borrowing, the call "overflows" to the corresponding macrocell. This vertical handoff is only

from a microcell to a macrocell. Calls in a macrocell are not allowed to "underflow" to a micro-because they are moving at high speed and it is very likely to handoff in a short time, resulting in excess load for the overall system. The flowchart of this process is in Figure 5-2.

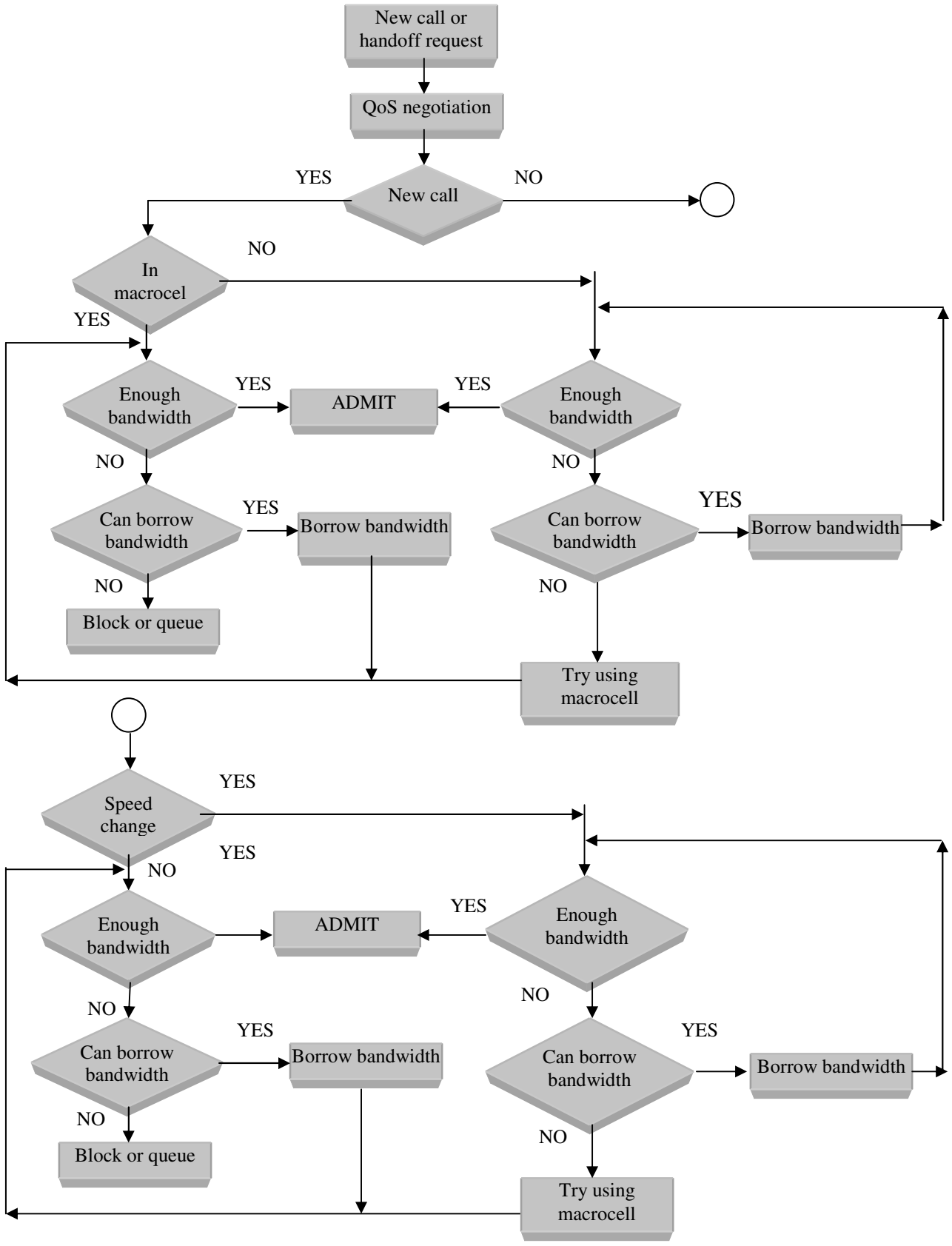


Figure 5-2: Handoff and new call admission process

The final step is to allocate bandwidth to the connection according to the negotiated QoS level. Handoff calls will have data routed to the new serving base station.

In order to balance the load in the system, it is possible to continuously or periodically monitor the speed of all users and vertically handoff their calls if the speed changes below or above a threshold. The disadvantage is the significant computational load. Therefore in this model vertical handoffs happen periodically or only when horizontal handoffs are required. In other words, when a handoff becomes necessary, the algorithm decides if it is vertical or horizontal.

5.2 Current admission control schemes for hierarchical networks

Handoff techniques

Before reviewing the current schemes, we are going to discuss some jargons used in this context. In most solutions, new calls and slow moving handoff calls are admitted into the micro-tier first. If there is not enough bandwidth, the call tries an alternative route in the macro-tier to avoid rejection (blocking and dropping). This routing process is referred to as "overflow". Two similar processes in the other direction are "repacking" and "underflow". Lo et al [80] introduced an "underflow" process to accommodate inter-macrocell handoff calls to be admitted in the micro-tier if the target macrocell can not support it. This improves the handoff call dropping probability at that instance but raises the number of handoffs in the micro-tier. Whiting and McMillan's "repacking" technique [81, 82] is to shift calls which are admitted at the macro-tier (due to insufficient bandwidth in the micro-tier) back to the micro-tier. Repacking frees bandwidth in the macro-tier so that the tier can operate as a backup bandwidth pool for all types of calls. The three processes are demonstrated in Figure 5-3.

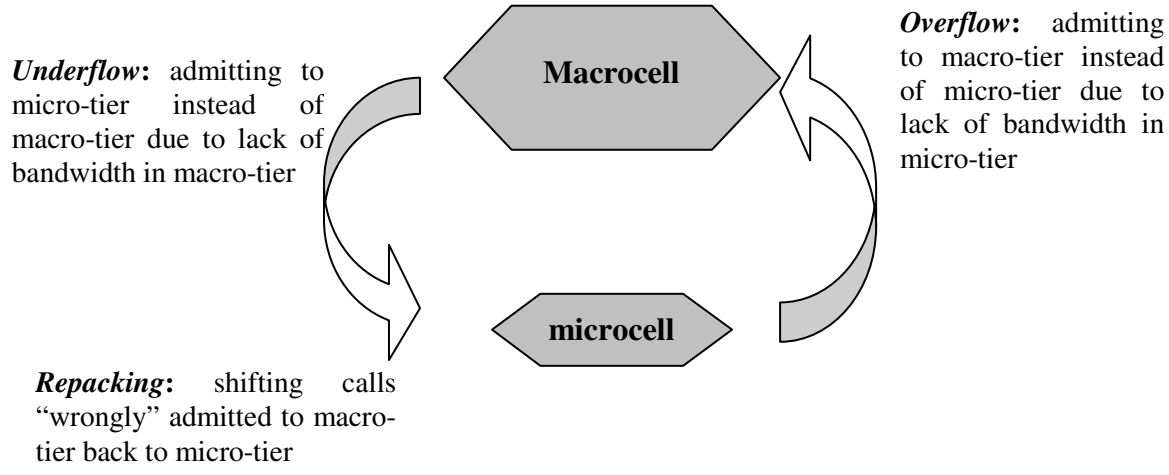


Figure 5-3: Overflow and repacking techniques

Handoff prioritisation by guard channels

In the simplest hierarchical model [83], users select tiers based on their speed and all calls are treated in the same manner. To reduce the handoff dropping rate, [84] reserves some bandwidth portion for handoff calls and allow call overflow from the micro to macro-tier. Improving the model in [63], The model in [85] uses a dynamic queuing discipline to prioritise handoff calls. Another model [86] applies call overflow in both directions to improve the utilisation with the trade-of of greater complexity. Salih and Fidanboyflu [87] uses FIFO queues with different queue times in the macro-tier to decrease handoff call dropping probability.

Vertical channel and vertical-horizontal channel strategies were introduced by Lin and Tseng [88]. Rappaport and Hu [81] propose an overflow/ no repacking scheme to reduce the blocking probabilities and dropping probability. Maheswari and Kumar's model [89] used "repacking". Another scheme [90] performs repacking whenever there is free bandwidth in the micro-tier while Valois and Veque [91] performs repacking when this is a speed change. Valois and Veque [91] also proposed a repacking on demand scheme. This occurs when the macro-tier can not admit a call and some calls in this tier can be shifted to the other tier. Because a macrocell covers a number of microcells, it is possible to select the microcell with the least traffic load to shift the call to. Otherwise the selection is purely random. [92] compared those basic schemes and found that the repacking on demand scheme gives the least handoff.

Rajput and Fapojuwo [93] proposes a macrocell size adjustment scheme. The channel capacity of macrocells can be increased by changing their size. The adjustment is done step-by-step. When macrocell size increases, the number of horizontal handoffs at the macro-tier will decrease. The simulated results show an improvement of successful handoff probability over the velocity. When macrocell size increases, the number of horizontal handoffs at the macro-tier will decrease. The simulated results show an improvement of successful handoff probability over the increase of velocity.

Anpalagan and Katzela [94] used a two-tier model with a multi-level macro-tier. Calls are selected by their speed and assigned to a level in the macro-tier. This approach works based on the assumption that an appropriate speed estimation scheme exists. Lin and Tzeng [95] group adjacent cells to form a cluster; and apply two admission control processes: one at cluster level and one at cell level. The two algorithms use guard channels to prioritise handoff calls. Others [85, 96] use complex queuing algorithm to prioritise handoff calls.

We have discussed about the general admission processes, the current admission control schemes as well as the resource reservation specially designed for hierarchical networks. To visualise the perception of this chapter, we will look at some complete models next.

Other handoff prioritisation approaches

Cimone et all [97] proposed a generic admission model which admitted all calls to the micro-tier and selectively handed fast-moving terminals to the macro-tier. Their revised models implemented overflow technique as well as admission of high bandwidth consumption calls in the micro-tier. Lagrange and Godlewski's model [98] does not allow handoff from macrocells to microcell and repacking process only happens when a horizontal handoff in the micro-tier is required. The MAHCN model [99] prioritised Class I calls by admitting them into the macro-tier immediately. Other calls are admitted to the micro-tier; however they can overflow to the other tier if necessary.

Those models have the following weaknesses:

- High frequent handoff risk for high bandwidth consumption calls [97]: Those calls often have real-time applications which are sensitive to delays. A handoff process always incurs a certain time delay. Admitting them into the micro-tier is acceptable for slow-moving terminals; however fast-moving terminals will have many handoffs during its call life. Every time a handoff happens, the call is in risk of being dropped.
- Inflexible vertical handoff [98]: The initial purpose of having two tiers is to reserve the macro-tier for fast-moving terminals and handoff from the macro-tier to the micro-tier frees resource to allocate for those terminals. Disabling this handoff ability (i.e. once the call stays in the macro-tier, it will be for its whole duration) will reduce the available resource for other admission request. However, the trade-off is less overhead in the system.
- Over-utilisation of the macro-tier [99]: Admitting all Class I calls into the macro-tier guarantees a minimum handoff requirement for them and the least overhead in the system. However stationary Class I calls will unnecessarily waste the resources in the macro-tier.

Current optimisation techniques

During handoff, the mobile receiver may lose some data when it disconnects to the serving base station and connects to the target base station. Ramjee et all [100] summarised some techniques focusing on Mobile IP. These proposed techniques are summarised with some comments below.

- Data re-direction [101-103]: All data are quickly re-directed to the new location of the mobile terminal. However, fast re-direction does not guarantee any loss in the transmission.
- Domain registration [103-107]: Cells are grouped into a domain which is controlled by the domain controller. When a terminal enters a cell, it registers to the domain, and then it can move freely inside the domain without making a handoff. This process reduces the number of handoffs in cells inside the domain. Cells at the domain border do not have any benefit. The technique can not be applied directly into cellular networks because it is specific for Mobile IP with two IP addresses.
- Transmission discontinuity [99]: Packet transmission is paused during the handoff and resumed in the new cell. This approach causes disruption in transmission. If the handoff process takes too long, the application in the terminal may time out.
- Multicasting [103]: The proposed scheme multicasts data to multiple possible locations around the mobile terminal. This approach does not take into account the target cell prediction i.e. it treats all neighbour cells as possible target cells.
- Virtual connection tree [108]: A virtual connection is a collection of connections from a fixed node in the wired network to a number of adjacent base stations (known as a cluster). When a call is admitted to the network, a virtual connection tree is established. Handing off between cells inside the virtual connection tree does not add more bandwidth to the network because the mobile simply sends packets using the connection from its current base station to the fixed node, then to the target base station. This decreases the necessary handoff signalling from the overall system.

We have seen in the admission control in hierarchical networks that tier selection is very essential to guarantee a successful operation. It is common to make the decision based on the speed of the terminal. Slow moving terminals are often attached to the micro-tier while fast moving terminals connect to the macro-tier to minimise the number of handoffs. Speed estimation is an interesting research topic. In this thesis, we are unable to cover such a broad topic; so for simplification, we assume that an appropriate speed estimation method is available and the model will use the estimated speed to make admission decision.

5.3 Efficient strategy: 2.5-tier model

We have analysed most of admission control models and schemes proposed for hierarchical networks and the techniques used to optimise the performance. It is possible to further improve the handoff call dropping rate with traffic class consideration.

Assumptions and conventions

Our model is based on some assumptions and conventions. The assumptions are from previous feasible solution.

First of all, as any other admission control for hierarchical networks, we assume that an efficient speed estimation technique exists. This ensures a precise tier selection.

The second assumption is adaptive traffic. In the traditional voice traffic, where traffic source is constant bit rate, the problem of forced-termination is one of the QoS measurements in cellular networks. For IP multimedia traffic, the problem moves toward bandwidth adaptation because the call is not dropped but has its service reduced [109]. We assume also that the system can adjust the bandwidth allocation of each call by negotiation and the running application can adapt to this flexibility.

Thirdly, we use the virtual connection tree from Acampora and Naghshineh's proposal [108] to reduce the overhead during handoff.

Fourthly, cells in the same tier are not overlapped and a cell is in coverage of one super-cell only. A super-cell is defined as a larger cell in the higher tier.

Finally, we use a two-class traffic convention as in other research. Class I calls are meant for real-time applications and Class II calls are for non-realtime applications. We also assume that a call does not change its class during its life and a terminal supports only one call at one traffic class. A moving terminal remains the same direction for its call duration.

Above are the assumptions and conventions which are general to the model. Others which are too technical will be discussed along the system description.

5.3.1 System description

The system consists of three overlaid tiers: two primary tiers (a micro-tier, and a macro-tier, combined) and a temporary tier.

A macrocell with radius R physically covers an area equivalent to N microcells, with radius r ($R = N \times r$). The cell shape is assumed to be hexagonal. The micro-tier aims to increase the number of available channels while the macro-tier prevents the number of handoffs from fast moving terminals. Only ongoing calls are allowed, location updating for idle terminals and start-up process for new terminals are not performed in those primary tiers. When a call completes or is dropped from an unsuccessful handoff, its control is passed back to the temporary tier.

Cells in the temporary tier are the largest. Those cells cover greater area than a macrocell. This tier is meant to do all the control functions for idle terminals i.e. those are not carrying a call. Idle terminals camp on this tier until they make a call. The control is passed to the appropriate tier, depending on their speed.

The speed of idle terminals is monitored closely. The system periodically requests the location of the terminals and classifies the speed to high or low.

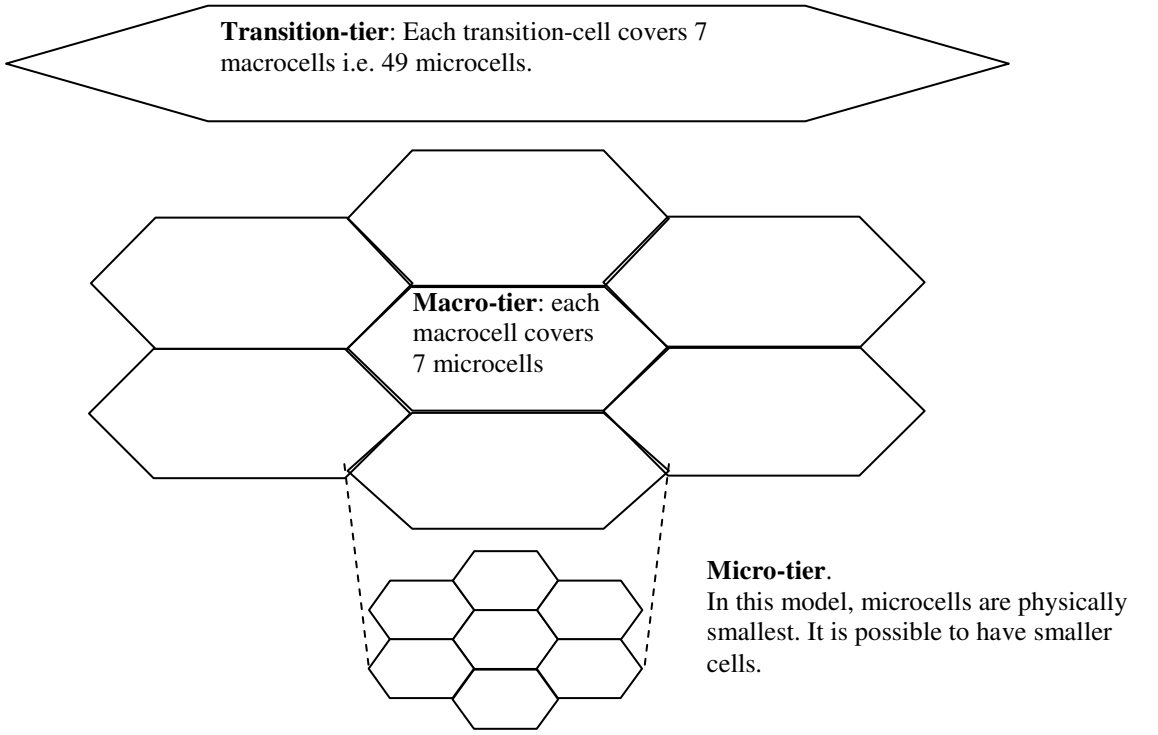


Figure 5-4: 2.5-Tier model

We use three tiers in this model but the temporary tier does not function as a full tier, which can carry calls. So we refer to this model is 2.5-tier model. Before moving to the admission algorithm of the model, we will look at how channels are allocated within the tiers.

5.3.2 Channel allocation

With the assumption that all cells are hexagonal, the system must follow the channel allocation and cell pattern of a traditional cellular network. Figure 5-5 shows how cells are arranged into patterns. Briefly speaking, any two cells using the same radio channels are physically separated at a distance sufficient enough to avoid channel interference.

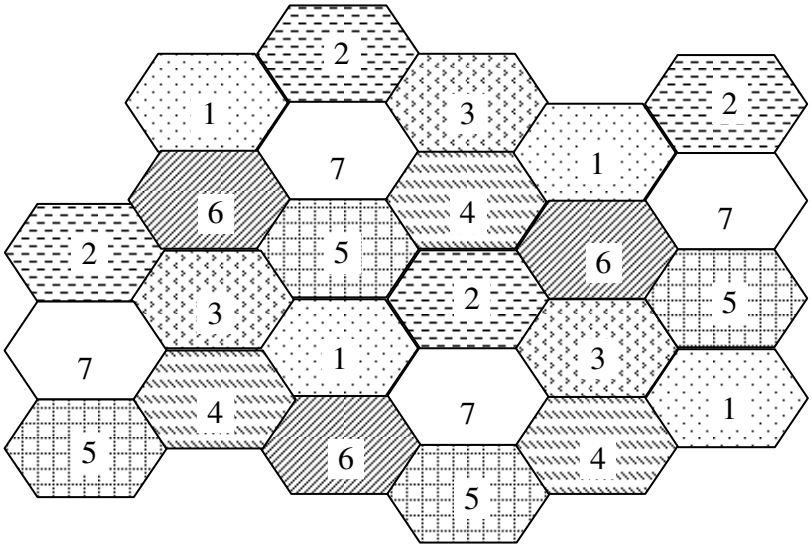


Figure 5-5: Cell arrangement in cellular networks

The total bandwidth is shared among three tiers. The micro-tier and macro-tier have an equally important role in the network while the temporary tier's role is slightly different. The channel arrangement in primary tiers can be adjusted according to the requirement of the network. For example if a cell covers a freeway, most calls will be classified as "high speed", hence the macro-tier should possess more bandwidth than the micro-tier. In other extreme, if the area is a shopping centre or an office, hardly can any people move "fast", the micro-tier should have more channels.

The temporary tier has a different arrangement. In normal cellular networks, bandwidth is shared between signalling channels and traffic channels. In our model, cells in primary tiers do not have any signalling channels except those associated with traffic channels for ongoing calls. Other signalling channels for location updating, paging etc are relocated to the temporary tiers. Therefore, the total bandwidth for this model and other models is technically the same.

Some examples of the channel arrangement between primary tiers are shown in Table 5-1.

Table 5-1: Bandwidth allocation for each tier

Tier	Percentage of total bandwidth		
	In shopping centres	In highways	In our simulation
Macro	10	90	50
Micro	90	10	50

In the next section, we will define the admission algorithm of the model.

5.3.3 Admission algorithm

We aim to prioritise handoff calls over new calls, and Class I calls over Class II calls. Among handoff calls, a horizontal handoff is more important than a vertical handoff. A horizontal handoff happens when the call moves out of the coverage of the serving cell whereas a vertical handoff occurs when the speed of the terminal changes. Based on the fact that speed change is not common, we can suppose that prioritising horizontal handoffs is more important. Moreover, the model also leaves unsuccessful vertical handoff calls at the original cell i.e. vertical handoff calls will never be dropped.

We make another assumption that a call performs a vertical handoff only when it needs a horizontal handoff. In the early hierarchical models, vertical handoffs are performed immediately when a terminal changes its speed. This approach requires tremendous overhead signals to continuously monitor the speed of all ongoing calls. We simplify this by checking the speed change at the cell border, when a horizontal handoff takes place only.

From the consideration of call characteristics and the handoff types, we define six different admission levels in total, listed in the most prioritised first:

Level 1. Class I horizontal handoff calls: Class I calls performing horizontal handoffs.
Level 2. Class II horizontal handoff calls: similar analogy
Level 3. Class I vertical handoff calls: Class I calls performing vertical handoffs.
Level 4. Class II vertical handoff calls: similar analogy
Level 5. Class I new calls: Users trying to make a Class I call
Level 6. Class II new calls: Users trying to make a Class II call

Now we are going to deeply analyse the operation of the algorithm, starting with new calls, the bandwidth negotiation process and handoff calls.

New call requests: consists of the following steps.

1. Connection request: At the initial stage, the call provides its minimum acceptable bandwidth, its desired bandwidth and its traffic class to the camped-on cell in the temporary tier.
2. Tier selection: When requesting a connection, the system checks the speed of the terminal. “Fast-moving” terminals are referred to the macro-tier, otherwise to the micro-tier.
3. Cell selection: In the appropriate tier, the cell which can provide the strongest signal strength (normally it is the one whose base station physically closest to the mobile

terminal) will process the call request. The system then hands the control of the terminal the appropriate cell.

4. QoS negotiation: Once identified which cell it will dwell in, the mobile terminal will negotiate its QoS requirement. The negotiation procedure is discussed shortly later.
5. If the negotiation is unsuccessful (i.e. there is not enough bandwidth for the call even at minimum acceptable level) and the chosen tier is the macro-tier, the call request is queued until the user abandons the call or the waiting time exceeds the maximum queuing time.
6. If the negotiation is unsuccessful and the chosen tier is the micro-tier, an overflow procedure is triggered to pass the call request to the overlaid macrocell. If this macrocell can not admit the call, it is queued until the user abandons the call or the waiting time exceeds the maximum queuing time.
7. If the negotiation is successful, the call is admitted and a virtual connection tree is established between its serving cell to its target cell. There are only two cells in a virtual connection tree in this model. The tree is rebuilt when the terminal moves to another cell. The target cell is the nearest neighbour cell which the call will handoff to. It is identified by the velocity detection mechanism.

A flowchart for new call request admission control is in Figure 5-6.

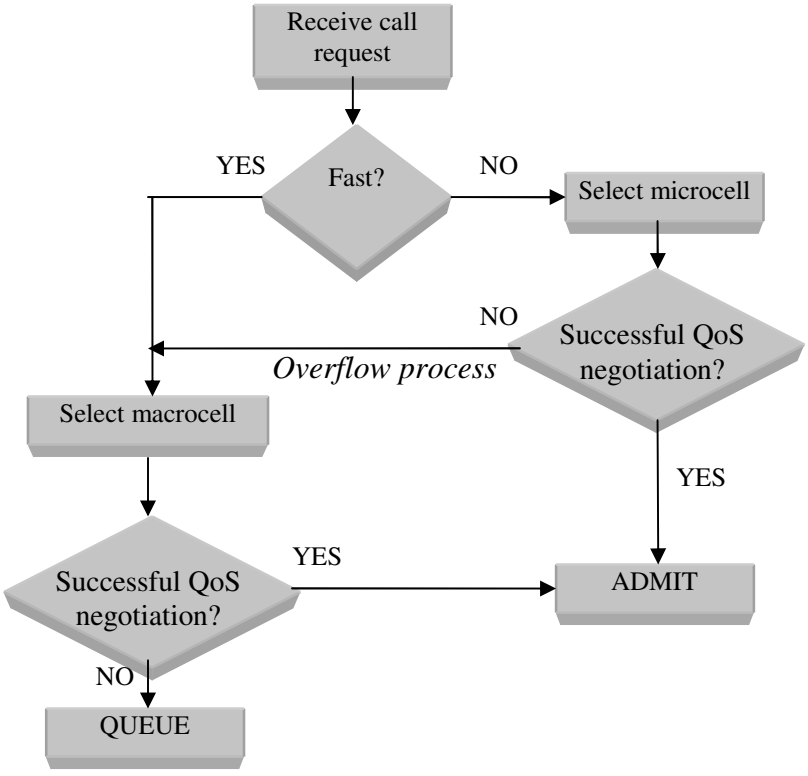


Figure 5-6: New call admission in 2.5-tier model

The bandwidth negotiation procedure consists of the following steps:

1. The call specifies its minimum acceptable bandwidth and its desired bandwidth. It expects to be admitted with its desired bandwidth and but it can still operate at the minimum acceptable level.
2. If bandwidth is sufficient, the request is accepted with desired bandwidth. Otherwise the call lowers its bandwidth requirement by one bandwidth unit until it is admitted or it is at the minimum acceptable level.
3. If it is not possible to admit the call at the minimum acceptable level, the cell reduces the service (i.e. bandwidth allocation) of some existing calls to free bandwidth. Calls with higher bandwidth loss tolerance will be subject to service reduction first. At any time the bandwidth allocation are guaranteed not to drop below the minimum acceptable bandwidth level.

The negotiation process is summarised in Figure 5-7.

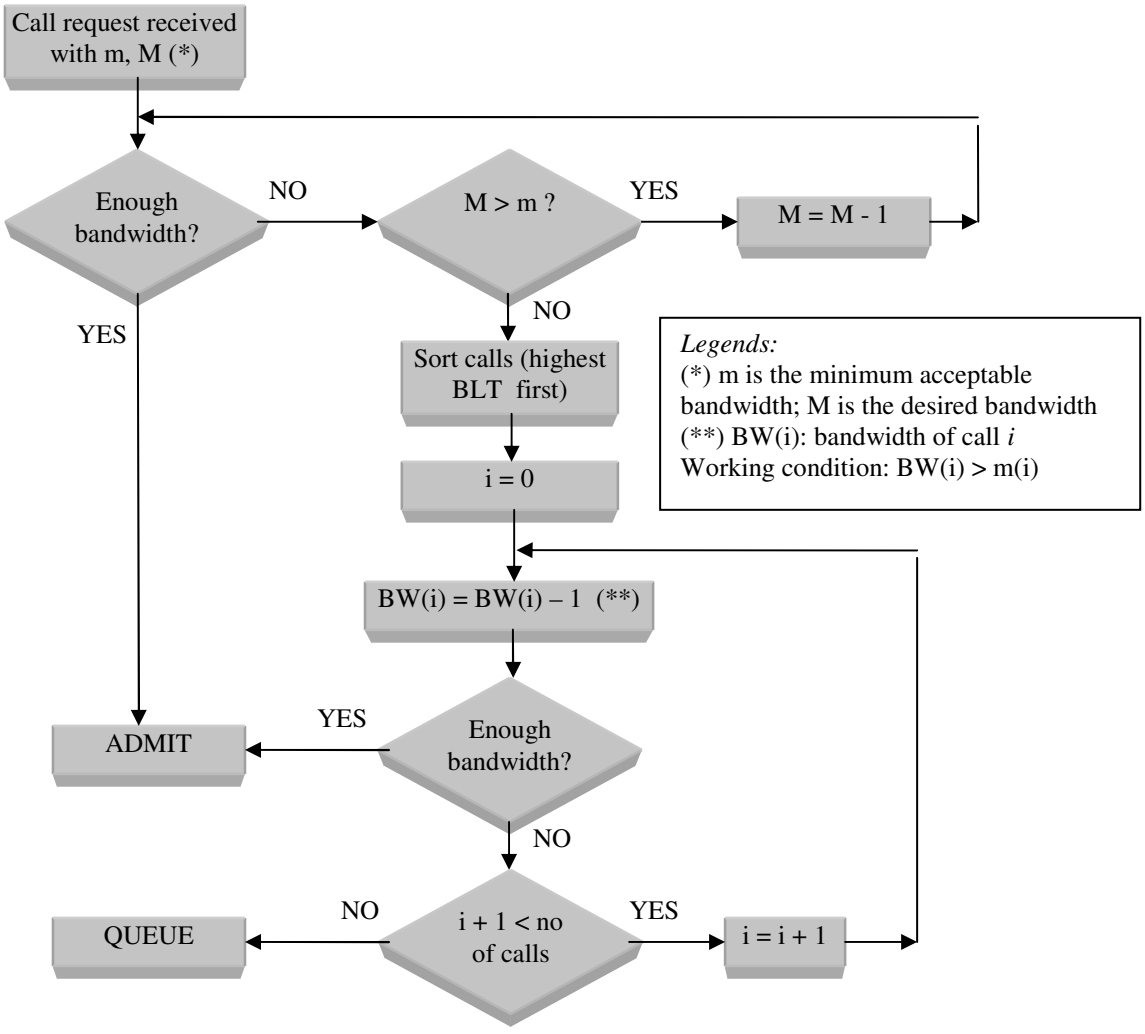


Figure 5-7: Negotiation procedure in 2.5-tier model

Handoff call requests

1. Handoff prediction: The system tries to detect a possible handoff early by estimating the moving direction of the terminal and the time instance to handoff. Only the neighbour cell in the moving direction expects the incoming call.
2. Bandwidth reservation: The predicted target cell will prepare some bandwidth for the incoming call which is the desired bandwidth of the call.
3. Horizontal handoff: In the micro-tier, if the target microcell can not admit the call, the system borrows bandwidth by reducing bandwidth allowance of current connections. If the borrowing process does not provide enough bandwidth, the call will use the overlaid macrocell, i.e. a vertical handoff is required. If vertical handoff is not successful, the connection request is placed in an admission queue. In the macro-tier, if the target macrocell can not admit the call, it is placed in an admission queue. Calls in admission queues wait for free bandwidth or are dropped because of weak signals.
4. Vertical handoff: The repacking technique monitors all terminals for speed change, and it incurs too much overhead. For simplicity, this model only considers a vertical handoff at the time when it needs a horizontal handoff i.e. at the cell border.

The summary operation of the algorithm is in Figure 5-8.

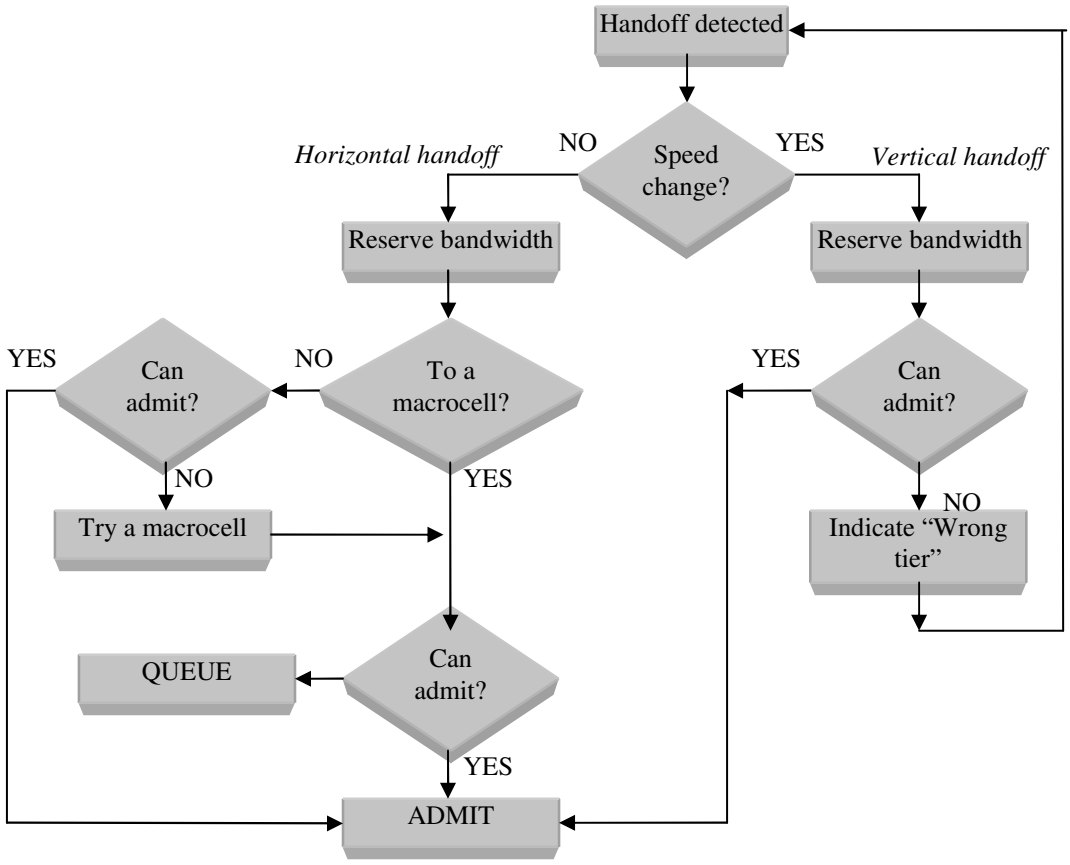


Figure 5-8: Handoff call admission in 2.5-tier model

We have described in details the model and its algorithm. The next step is to consider the allocation of bandwidth in a cell in primary tiers.

5.3.4 Bandwidth allocation in a cell

We have discusses six levels of prioritisation in section 5.3.3, page 79, corresponding to a number of admission levels. The summary can be found in Table 5-2 below.

Table 5-2: Admission levels in 2.5-tier model

Level	Call characteristics
1	Class I horizontal handoff calls
2	Class II horizontal handoff calls
3	Class I vertical handoff calls
4	Class II vertical handoff calls
5	Class I new calls
6	Class II new calls

In chapter 3, we have produced the improved Threshold Access Sharing (iTAS) admission scheme for a single cell. The scheme is proven to work better than the original Threshold Access Sharing. This model will apply the iTAS in the cell level with a slightly modification.

For six admission levels, we define four thresholds on the utilisation of the bandwidth pool. If the utilisation is under a threshold, the system uses a predefined policy, which allows calls with certain characteristics to be admitted.

The policy is defined such that horizontal handoff calls are more important than vertical handoff calls, general handoff calls are more important than new calls and Class I calls are more important than Class II.

To capture the above explanation, a summary of the admission levels and thresholds is given in Table 5-3 and Figure 5-9 below.

Table 5-3: Thresholds used in 2.5-tier model

Threshold	Admission level	Calls
$> t_1$	1	Class I horizontal handoff calls
$> t_2$	1, 2	All horizontal handoff calls
$> t_3$	1, 2, 3	Horizontal handoff and Class I vertical handoff calls
$> t_4$	1, 2, 3, 4	All handoff calls
Below t_4	1, 2, 3, 4, 5, 6	All calls

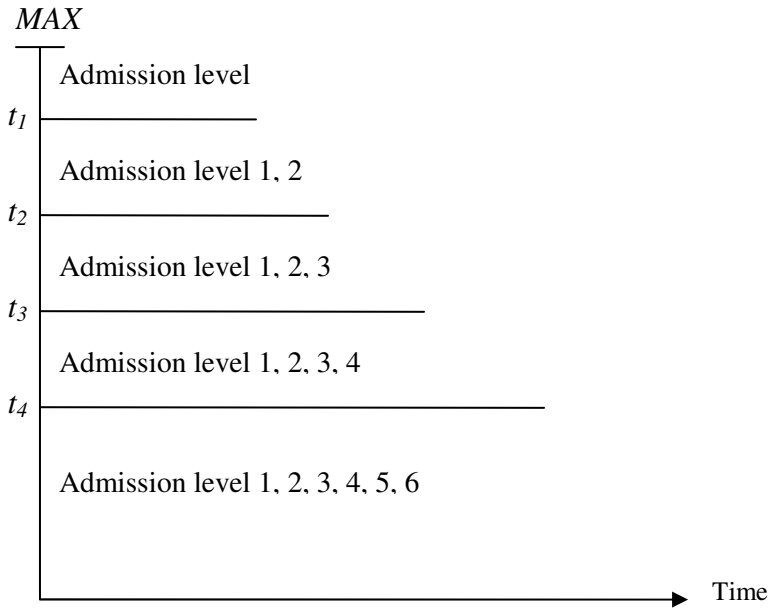


Figure 5-9: Improved Threshold Access Sharing in 2.5-tier model

The model has been described from the assumptions to the algorithm. In the next section, we will implement the idea into a simulation and study its performance.

5.4 Simulation and result discussion

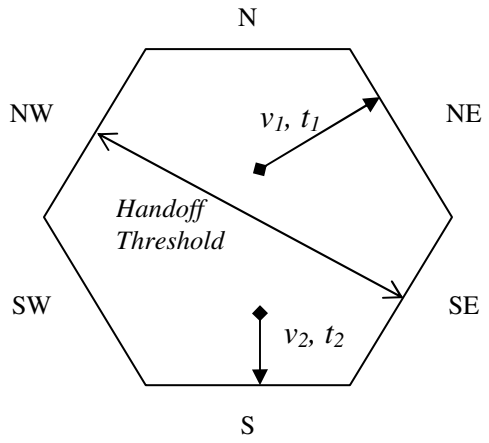
Simulation setup

There were 49 microcells in our micro-tier and 7 macrocells in the macro-tier. Each 7 microcells were overlapped by one macrocell. The total 7 macrocell was covered by a temporary cell. Because the temporary cell does not have any function in traffic transmission, we imitated its effect as a time delay in the process.

When a new call was generated, it was given a traffic class and a speed category. The traffic class did not change for the call life whereas the speed could. There were totally 20% Class I calls and 33% fasting moving calls. There were 10% calls changing their speed during handoff (to initiate vertical handoff).

A call moved to one of six neighbouring cells in the same tier in a random direction (initially set). If its speed changed, it would make a vertical handoff; otherwise it would make a horizontal handoff.

When generated, a new call could be anywhere in its cell (the position is defined by the value of a countdown timer). This timer defined when the next handoff would occur. The direction defined where the target cell was. The concept and an example are in Figure 5-10.



Call 1: moving to direction NE at v_1 speed and will take t_1 (time unit) to the next handoff.

Call 2: closer to a handoff than Call 1, moving to direction S at v_2 speed and take t_2 to the next handoff

After handing off, the timer is set to the Handoff threshold.

Figure 5-10: Call movement in simulation

The first handoff occurred when the timer reached zero. After the handoff, the timer was reset to the handoff threshold value. This represents the next handoff event. The setting was based on the assumption that the moving direction was perpendicular to the edge of the hexagon (the assumed shape of a cell).

Our simulated model was built in many phases. In the first phase we observed the effect of the overflow process in a normal 2-tier model. In the second phase, we built our 2.5-tier model. Calls were admitted to a certain admission area depending on their characteristic. In the third phase, we added the bandwidth negotiation and overflow mechanisms to improve the new call admission. In the fourth phase, we implemented the rate-based bandwidth borrowing scheme to phase 3 model to improve the handoff call admission. In phase 5, we prioritised Class I new calls by restricting Class II new calls from bandwidth negotiation. Finally in the last phase, we combined all the good features in our complete model.

There were 1,000,000 calls generated during the simulation. Many of them lasted a number of handoffs before completed. In one of the program runs, we had 2,693,108 handoff events.

Simulation results

Phase 1: Overflow for handoff calls

First of all, we looked at the normal micro-macro tier. To reduce the handoff dropping probability, we implemented the overflow process. This process allowed slow moving calls to handoff to a macrocell instead of a heavily loaded microcell. The new call requests were treated the same in both cases. As expected, the result in Figure 5-11 showed that the handoff call dropping probability was less than the second case, where overflow was used. The new call blocking probabilities (shown as overlapped on the graph) were the same and higher than the dropping values.

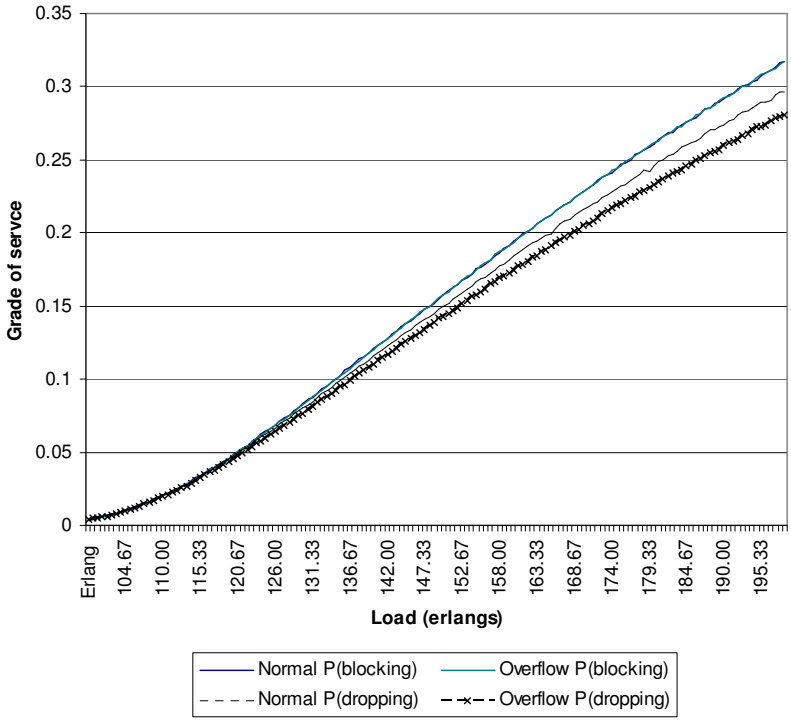


Figure 5-11: Observing the Overflow effect for handoff calls

We took further investigation in the overflow case and found that Class I calls (both new and handoff types) experienced higher failure probability. This was due to the fact that Class I calls requires more bandwidth than Class II; hence the chance of their connection being refused was higher. A similar explanation applied when the Class II handoff call dropping probability was lower than that of Class I. The results were shown in Figure 5-12.

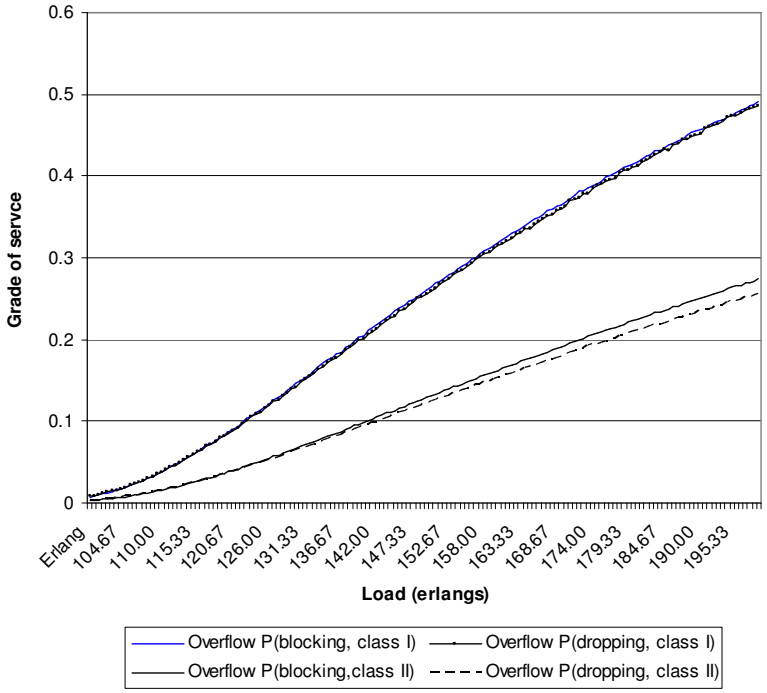


Figure 5-12: Overflow effect on handoff calls with traffic class consideration

Phase 2: Compare 2.5-tier model with normal hierarchical structure + overflow

In Figure 5-13, we implemented our proposed 2.5-tier model, prioritised handoff calls over new calls. A portion of the bandwidth was reserved for handoff calls, leaving less capacity to accommodate new calls. Overflow was applied for slow moving calls as well. Therefore the new call blocking probability was higher, compared to the normal hierarchical structure with overflow mechanism.

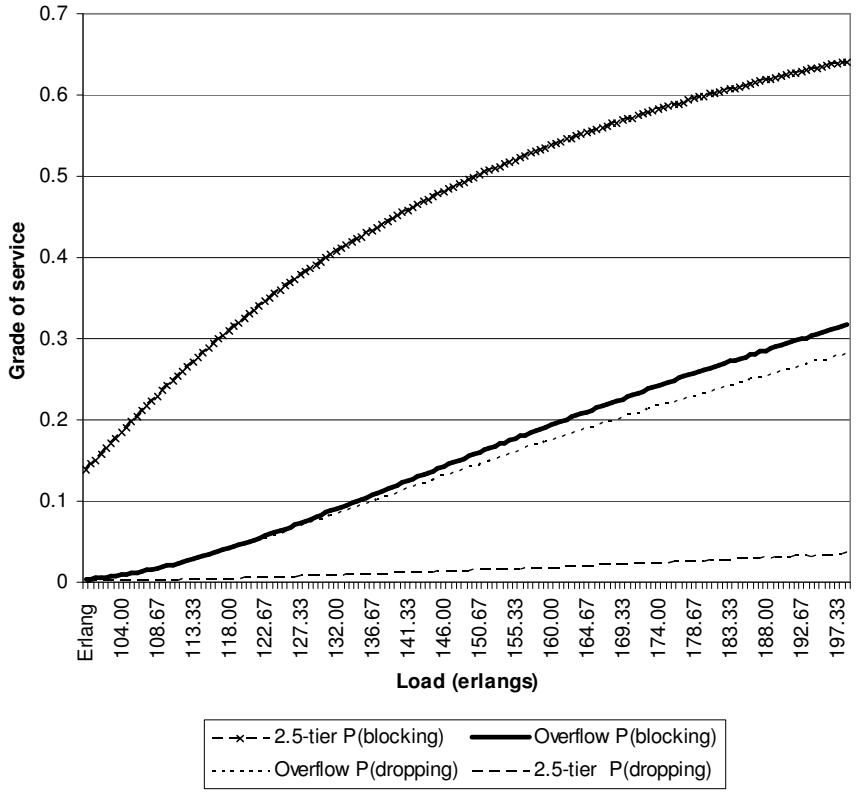


Figure 5-13: Simple 2.5-tier model compared to normal 2-tier structure

Next in Figure 5-14, we looked at the behaviours of the 2.5-tier model with more attention to traffic classes. In the 2.5-tier model, we prioritised Class I handoff calls over Class II handoff calls; however we treated new calls in the same manner, regardless their traffic classes. It resulted a high blocking rate for Class I new calls, compared to Class II. This was because Class II calls required less bandwidth and had more chances to get enough bandwidth.

The normal overflow model did not separate the traffic into classes; hence only the total blocking or dropping rate was shown as references.

In the other hand, as we had prioritised Class I handoff calls, the handoff dropping probability for Class I calls was much less than that of Class II. The graph could not shown the comparison, so we attached the first few values in Table 5-4. For example, picking the values

in the first row, the dropping probability of Class I calls in the 2.5-tier model was 50 times as low as that of the Overflow model.

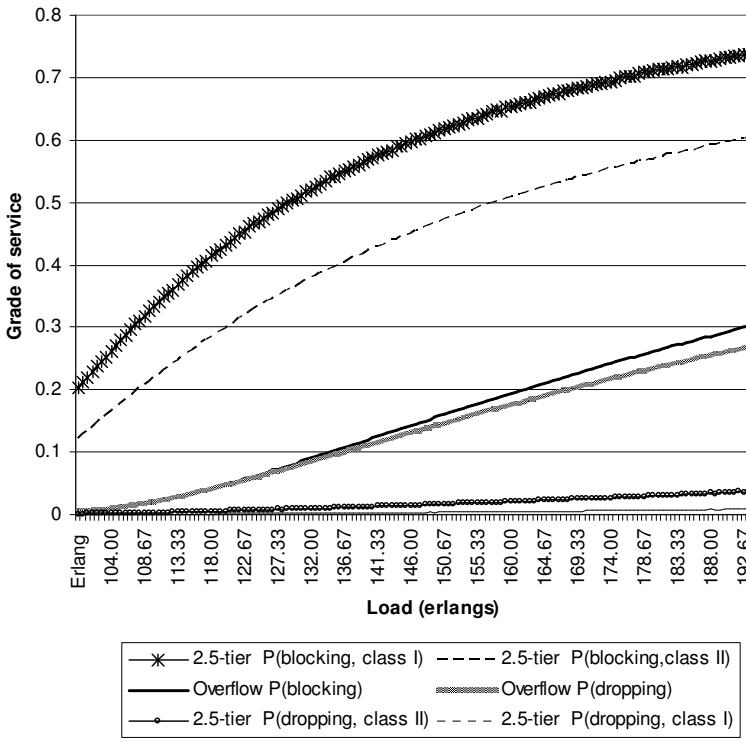


Figure 5-14: Simple 2.5-tier model compared to normal 2-tier structure with traffic class consideration

Table 5-4: Handoff call dropping probabilities comparison

	Overflow model			2.5-tier model		
	P(dropping)	P(dropping, class I)	P(dropping, class II)	P(dropping)	P(dropping, class I)	P(dropping, class II)
100	0.003968498	0.008041677	0.002965269	8.81E-04	1.62E-04	0.001045924
100.67	0.004580041	0.00905201	0.003491338	9.43E-04	2.18E-04	0.001109362
101.33	0.005255967	0.010449074	0.003990285	0.001011463	2.25E-04	0.001189938
102.00	0.005971588	0.01189999	0.004532015	0.001137608	2.60E-04	0.001334752
102.67	0.006433236	0.013073391	0.004814788	0.001241446	2.90E-04	0.001457926
103.33	0.007326561	0.01465635	0.005549947	0.001386369	2.64E-04	0.001640382

Phase 3: Add bandwidth negotiation and overflow for new calls in 2.5-tier model

Up until this phase, the trade-off to improve the handoff call dropping probability is high blocking probability for new calls. So we added a bandwidth negotiation and overflow process to give more chances to new calls. Handoff calls are not subject to bandwidth negotiation to ensure that they do not experience sudden bandwidth changes during handoff.

The result is shown in Figure 5-15. There was not any change in the admission of handoff calls so we did not include the sketches on the figure. The total new call blocking probability in the new case (Bandwidth negotiation: BW Neg.) was slightly less than that of the original 2.5-tier. The probability of new calls admitted at their minimum acceptable bandwidth level as a consequence of the negotiation process was shown as well.

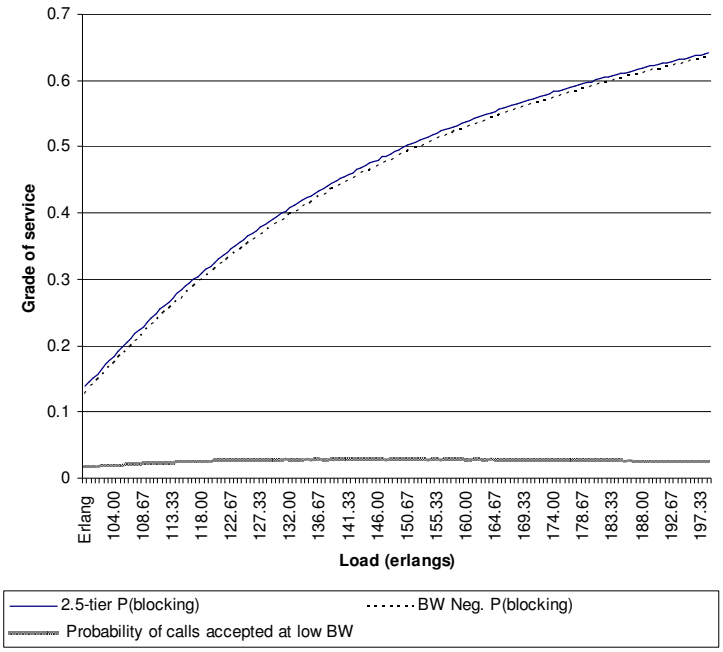


Figure 5-15: Effect of bandwidth negotiation for new calls

In Figure 5-16, we looked into the effect of the negotiation and overflow process in new calls. After applying the negotiation and overflow on new calls, the blocking rates for Class I and Class II changed and was very much the same (shown as overlapped line in the middle). The blocking probability for Class I calls reduced dramatically, whereas the blocking probability for Class II calls slightly increased. This trade-off was worth the effort of more computational load due to the negotiation and overflow mechanisms.

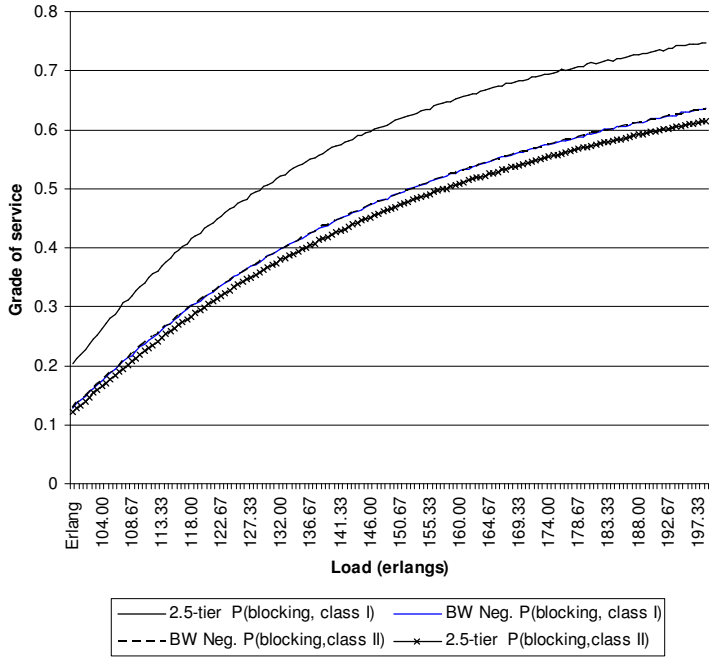


Figure 5-16: Effect of bandwidth negotiation for new calls with traffic class consideration

Phase 4: Implement bandwidth borrowing for handoff calls

To extensively exploit the characteristics of adaptive traffic in IP networks, we introduced the bandwidth borrowing feature into the current model. At this stage, this feature however applied for handoff calls only.

We ignored the values for blocking probability of new calls because the changes in this new case were only in handoff call admission. The result was in Figure 5-17. Only sketches for the previous model (2.5-tier with bandwidth negotiation for new calls) were visible. The values of handoff dropping probability in the new model were too small to display.

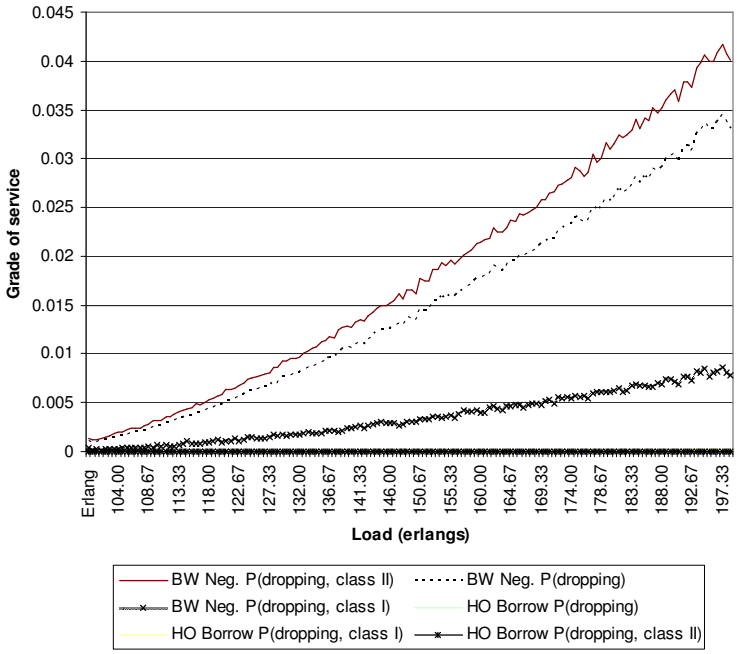


Figure 5-17: Effect of borrowing mechanism on handoff calls

We could look at the first few values in Table 5-5 to see the improvement in handoff call admission. Note that in some cases, the dropping rate of Class I calls was higher than that of Class II calls. Again, this was due to the high bandwidth requirement of Class I calls.

Table 5-5: Handoff call dropping probabilities after implementing bandwidth borrowing

Erlang	HO Borrow P(dropping)	HO Borrow P(dropping, class I)	HO Borrow P(dropping, class II)
100.00	4.31E-05	6.29E-05	3.87E-05
100.67	6.47E-07	0	7.95E-07
101.33	9.79E-07	0	1.22E-06
102.00	6.55E-07	1.58E-06	4.13E-07
102.67	6.61E-07	0	8.46E-07
103.33	3.65E-06	1.31E-05	8.60E-07
104.00	5.35E-06	2.23E-05	4.32E-07

In the current implementation, we borrowed bandwidth from ongoing connections just enough to admit the handoff call. This process ensured that only a minimum number of calls would have to give up bandwidth for other connections. In the other hand, when another handoff call

comes in, the same borrowing process is restarted. This results in computational load. Therefore we tried another approach: when a cell is heavily loaded and a handoff call comes in, the borrow mechanism will borrow from all current connections instead of borrow from a few connections, enough to admit the handoff request. This harsh way reduces the computational load because the next handoff call will not have to call the borrowing process.

This modification had more effects on handoff calls. Table 5-6compared the dropping rate from the scheme and the Harsh borrow scheme. The latter had a slightly better performance in terms of handoff call dropping. The blocking probabilities in both schemes are approximately the same, as can be seen on Figure 5-18.

Table 5-6: Comparison between selected bandwidth borrowing and harsh borrowing

Erlang	HO Borrow P(dropping)	Harsh Borrow P(dropping)
192.00	5.81E-05	5.86E-05
192.67	4.49E-05	4.46E-05
193.33	4.86E-05	5.57E-05
194.00	6.20E-05	5.08E-05
194.67	5.73E-05	3.51E-05
195.33	4.47E-05	5.90E-05

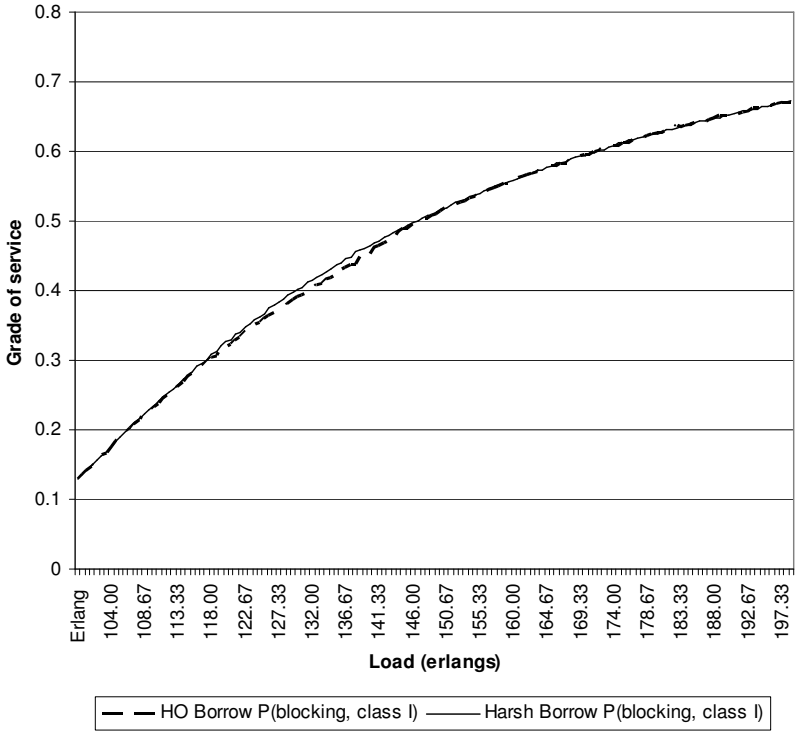


Figure 5-18: Effect of harsh borrowing and selective borrowing

This Harsh borrowing model implied some negative effects on ongoing calls (i.e. more frequent service reduction). However the improvement was not significant. Therefore, we did not follow this way, and stayed with the Handoff borrowing model (handoff calls could borrow bandwidth from existing connections).

Phase 5: Only Class I new calls can negotiate bandwidth

In phase 3, both Class I and Class II new calls could negotiate bandwidth (handoff calls do not need this because they are always given the required bandwidth); therefore Class I new calls were not prioritised over Class II new calls. In this phase, we continued the Handoff borrowing model and reserved the right to negotiate bandwidth to Class I calls only.

The result in Figure 5-19 was as expected. The blocking probability for Class I calls in the Class I borrowing model was better. As the load increased, the difference became more visible.

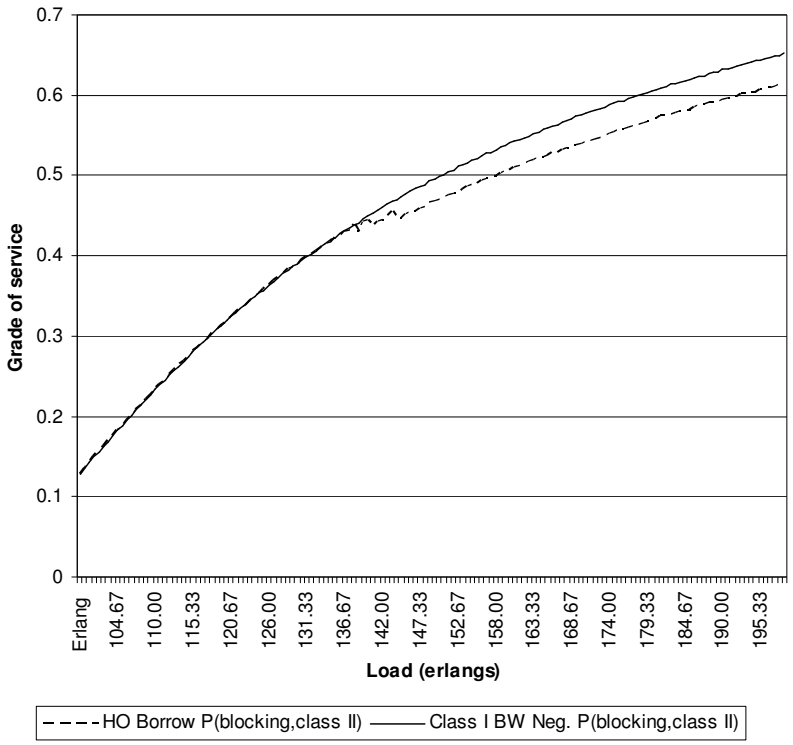


Figure 5-19: Allowing only Class I new calls to negotiate bandwidth

In the final phase, we implemented all the features with better performance.

Phase 6: Complete 2.5-tier model with handoff prioritisation

The following features were included:

- When a handoff call arrived to a target cell and there was not enough bandwidth, it would lower its bandwidth requirement and try again. If it still could not be admitted, it would attempt to borrow bandwidth from existing calls in that cell. We did not use the harsh borrowing approach, so only some of the existing calls would give up bandwidth for this handoff call. Calls with highest Bandwidth Loss Tolerance would give up their bandwidth shares first. This feature aims to prioritise the handoff calls over new calls.

- When the system received a new Class I call request and it could not accept, it would negotiate with the terminal to reduce the required bandwidth to a lower level which satisfied both parties (user and system). Class II new calls did not have this ability. This feature aims to prioritise Class I new calls over Class II new calls.
- All slow moving calls could overflow from the micro-tier to the macro-tier regardless their traffic classes. The other direction was disabled to prevent too frequent handoffs. This aims to improve the handoff call dropping and new call blocking rates.

First, we inspected the performance in the new call admission. Figure 5-20 showed the new class blocking probabilities in the final model and the previous one. In Figure 5-21 and Figure 5-22, we also looked at how calls in different traffic classes were admitted.

In all three graphs, the overall blocking probability in the final model was obviously better. However Class I new calls in the final model experienced a slightly higher blocking rate. This was due to their higher bandwidth requirement.

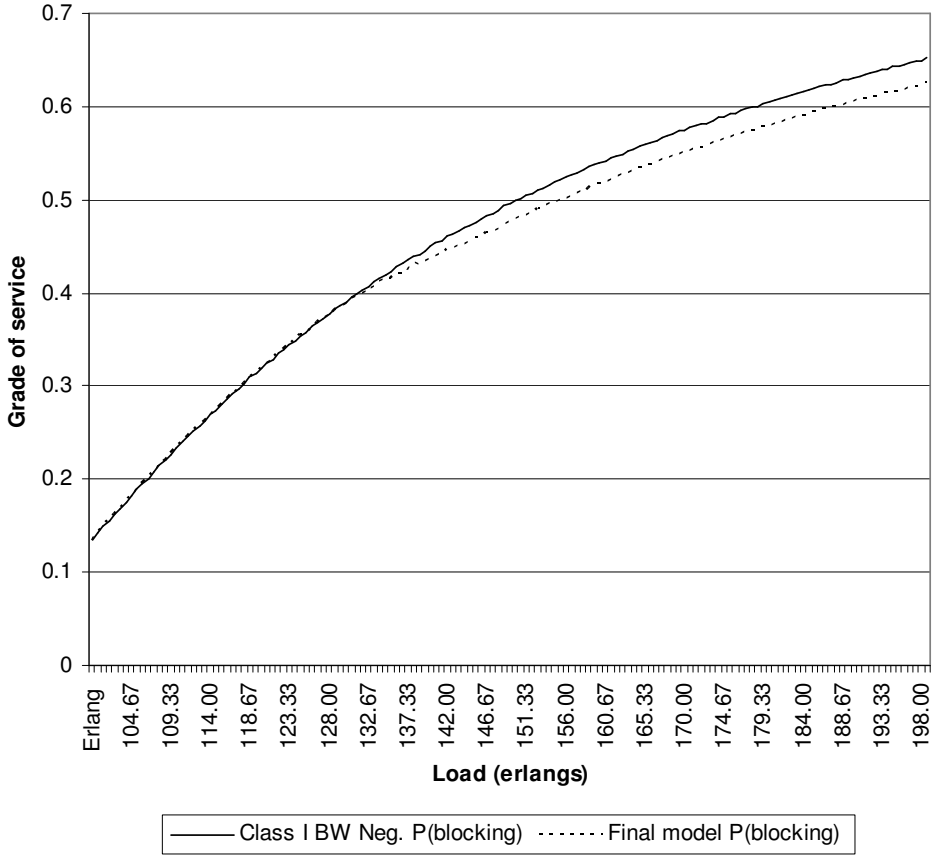


Figure 5-20: Final model compared to the previous in Phase 5

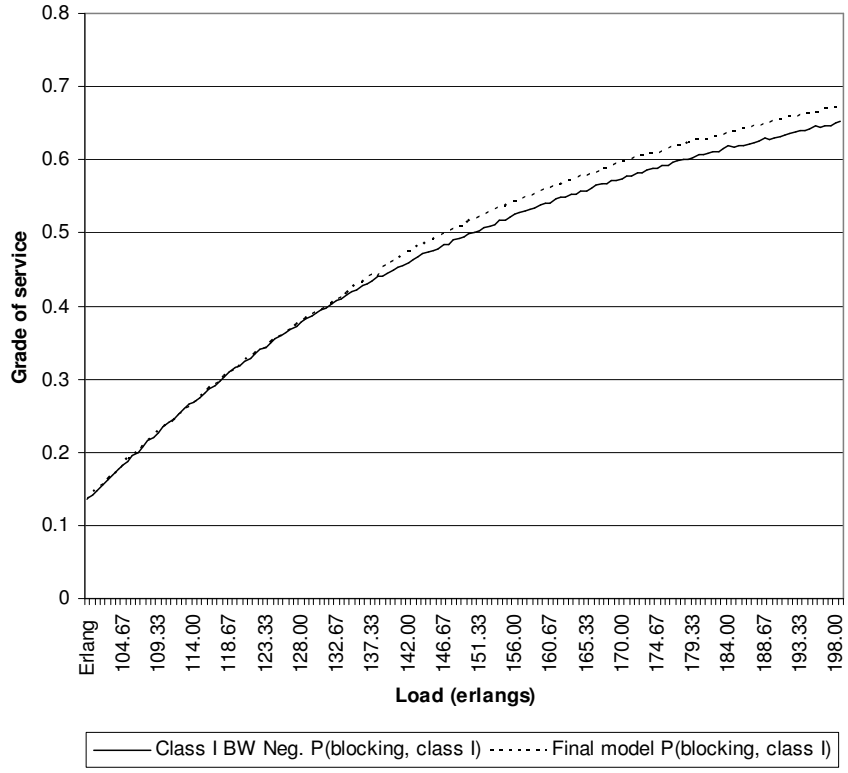


Figure 5-21: Final model (cont.) with traffic class consideration – Class I

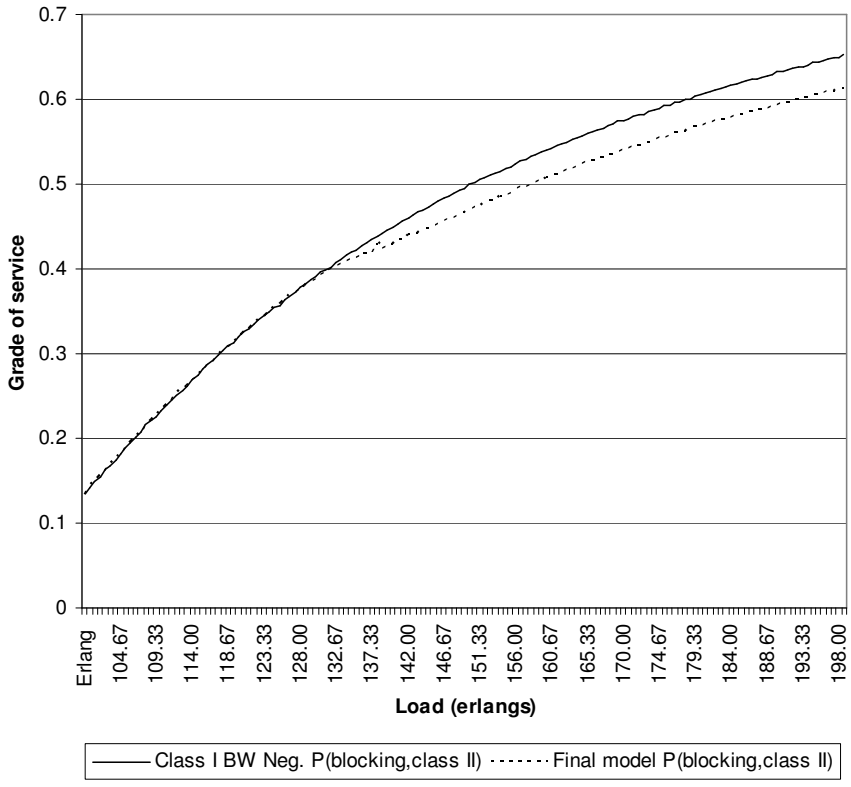


Figure 5-22: Final model (cont.) with traffic class consideration – Class II

In Figure 5-23, the handoff call dropping probabilities were compared. The occurrence was too less to achieve a smooth sketch.

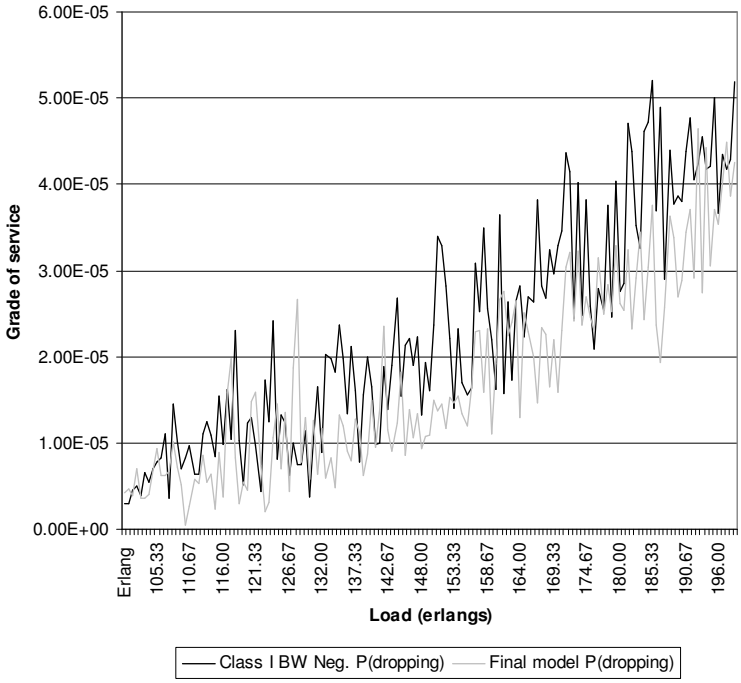


Figure 5-23: Handoff call dropping probability comparison

Going into details of the dropping probability, we found that Class I calls experienced most of the dropping. Class II calls with less bandwidth requirement did not have much experience. Table 5-7 shows the two probabilities.

Table 5-7: Handoff call dropping probability with traffic class consideration

Erlang	Final model P(dropping)	Final model P(dropping, class I)	Final model P(dropping, class II)
100.67	2.97E-06	1.08E-05	0
101.33	2.99E-06	1.12E-05	0
102.00	4.92E-06	1.60E-05	0
102.67	3.82E-06	1.49E-05	0

A further investigation was taken. We considered the number of handoff events as well as the number of drops occurring. Table 5-8 showed the number of handoff events and the number of calls dropped.

Table 5-8: Handoff call dropping probability with traffic class consideration – Details analysis

Erlang	**100.67**	**101.33**	**102.00**	**102.67**	**103.33**	**104.00**
Dropping probability	2.97E-06	2.99E-06	4.52E-06	4.92E-06	3.82E-06	6.54E-06
Total calls dropped	8	8	12	13	10	17
Total handoff events	2693108	2676947	2657657	2642363	2615101	2600164
Class I dropping probability	1.08E-05	1.12E-05	1.72E-05	1.60E-05	1.49E-05	2.59E-05
Total Class I calls dropped	8	8	12	11	10	17
Total Class I handoff events	740700	716765	698869	687701	672834	656392
Class II dropping probability	0	0	0	1.02E-06	0	0
Total Class II calls dropped	0	0	0	2	0	0
Total Class II handoff events	1952408	1960182	1958788	1954662	1942267	1943772

Conclusion:

To conclude the discussion of our proposed model, we would like to compare it to the normal micro/ macro tier structure (shown in Phase 1). First of all, the handoff call dropping probabilities were reduced to almost zero. These sketches could not be seen on Figure 5-24.

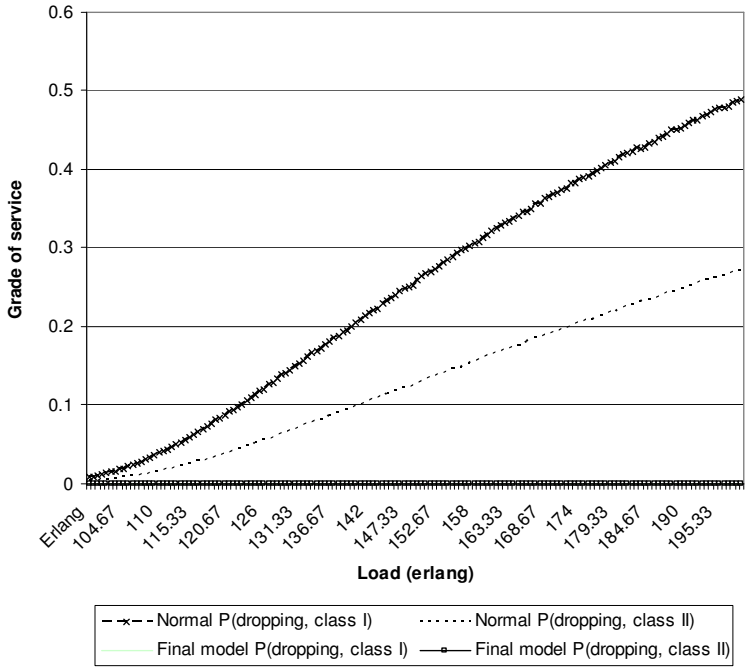


Figure 5-24: Comparison between 2.5-tier model and normal 2-tier model

As a trade-off, our model has higher new call blocking probabilities compared to the normal structure. Figure 5-25 showed the comparison. This is the best that we could achieve with the given environment settings.

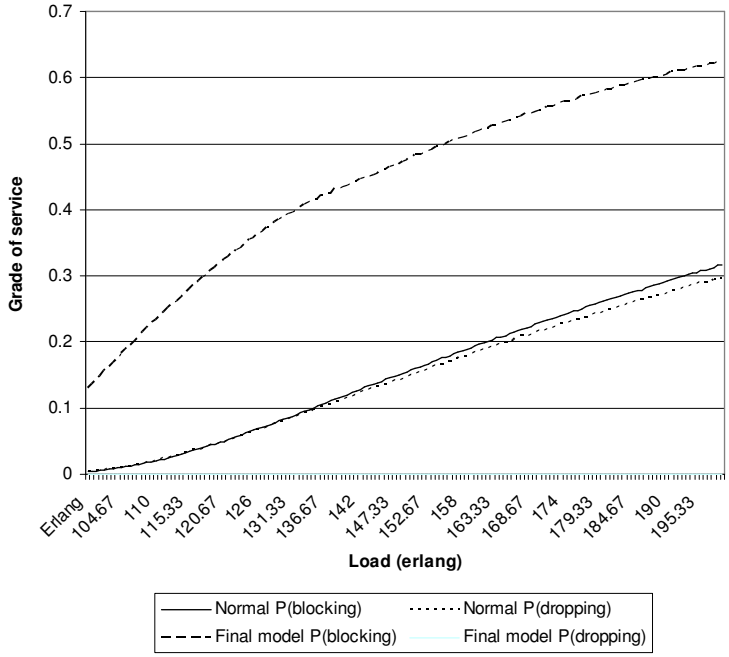


Figure 5-25: Comparison of new call blocking probability

At 200 erlang load, the handoff call dropping probability was successfully reduced from 0.30 to near zero; whereas the new call blocking probability increased from 0.31 to 0.61.

In Figure 5-26, our model showed better treatment to new Class I calls. As the load increased, it was harder for Class I new calls to be admitted; but the severity was much worse in case of the normal structure. At 200 erlang load, the blocking probability of Class I new calls in the normal structure was 0.5 and that of Class II was 0.28. In our model, at the same load, the blocking probability of Class I new calls was 0.68, compared to 0.61 for Class II.

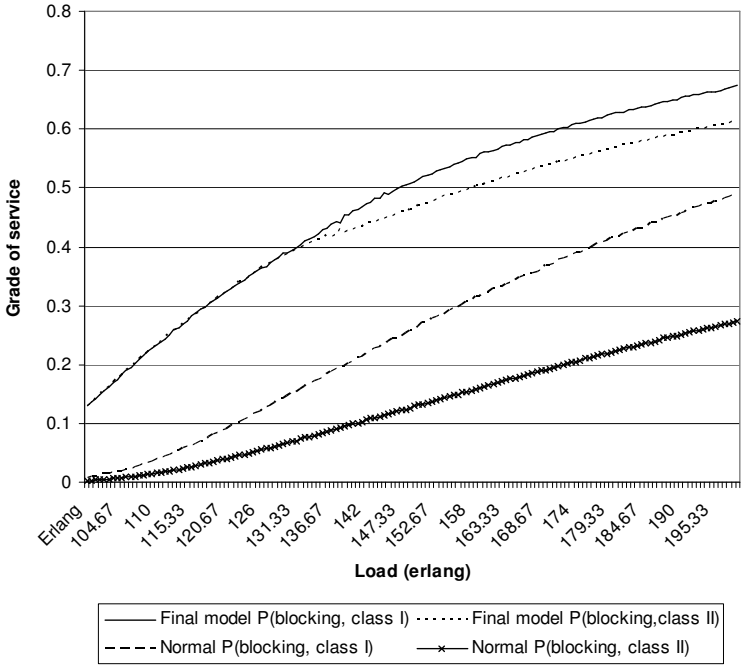


Figure 5-26: Blocking probability comparison with traffic class consideration

We have the following observations about our new model as below.

First of all, handoff calls experience little or none dropping. This is due to the threshold arrangement in the algorithm as well as the rate-based borrowing mechanism. The channel allocation can be adjusted by changing the threshold values in Table 5-3.

Secondly, we assure that the simplified rate-based borrowing is fair, in terms of taking and returning bandwidth. When performing handoff, the system can choose to borrow bandwidth from ongoing calls with the highest bandwidth loss tolerance. When bandwidth becomes available, the system returns the excess bandwidth to the ongoing calls with the lowest bandwidth loss tolerance.

Thirdly, the borrowing process does not happen during handoff to ensure that handoff calls will not experience sudden drop in bandwidth allowance. This also prevents the computational load by doing the bandwidth negotiation at the same time with handoff process.

Fourthly our model guarantees that the system will never take too much bandwidth that a call is forcefully terminated. The only case a call forcefully terminated during its life is when a handoff occurs.

Fifthly, bandwidth negotiation is feasible for new calls. If a new call can not be admitted at their desired requirement, it tries to lower the bandwidth level until it is admitted. Our model allows only Class I new calls to negotiate.

Sixthly, Class I calls are given higher priority than Class II in most cases. However their blocking and dropping probabilities are still slightly higher than Class II because they require more bandwidth; hence risk more chances of insufficient resource.

Seventhly, we found the new call blocking probabilities are higher. This is the trade-off for lowering the dropping rate.

5.5 Summary

In this chapter, an extensive review on existing admission schemes, models and optimisation techniques for hierarchical networks is presented and a new admission model for hierarchical IP networks is proposed. Our review and analysis on the existing schemes has shown that there is room for further improvement. A 2.5-tier admission control model for hierarchical IP networks is proposed. An admission scheme based on the model has implemented a slightly different version of the improved Threshold Access Sharing at cell levels. The new scheme has also adopted optimisation techniques such as overflow and rate-based borrowing aiming to achieve lower handoff call dropping probability. Bandwidth borrowing and bandwidth negotiation was deployed to ensure a fair bandwidth sharing among calls. In the scheme, calls are classified into two classes and Class I calls are prioritised over Class II calls. The admission thresholds (i.e. the parameters defining the admission area) are adaptive to traffic changes. Based on our simulation, the new admission scheme proposed on our 2.5 tier model has better performance in handoff dropping, compared with schemes on the normal 2-tier model and the improvement is significant. The scheme's new call blocking probability is comparable with existing schemes, even it is slightly increased. Some areas in the new scheme have been identified for possible future improvement. In summary, our newly proposed 2.5 tier model has introduced the concept of a virtual tier in admission modelling for hierarchical cellular IP networks. Its improvement over existing schemes has been demonstrated in our simulation. The concept can be extended to other multi-tier models of hierarchical cellular IP networks.

Chapter 6: CONCLUSION AND DISCUSSION

Quality of service (QoS) is an important issue in IP networks, which are providing more and more diversified services including the voice service and other real-time multimedia services. Wireless IP networks have become the fastest growing sectors in overall IP networks. In the case of the cellular wireless network, it has gone through 2 and 2.5 generations and entered the age of 3rd generation. Almost all services provided by the future wireless networks are expected to be IP based. Due to scarcity of the wireless bandwidth, admission control in the wireless networks became a critical component affecting the overall QoS and our study in this thesis is focused on the area.

In chapter one, we discussed about the IP networks with emphasising on wireless IP networks. The evolution of mobile phone networks is reviewed from a simple analogue cellular to a hierarchical digital network with multiple cell tiers. Call admission control in cellular networks was introduced. The research questions were raised with specification of the scope and objectives.

In chapter two, the background information of admission control is presented. We started with the review of QoS on IP-based applications and the summarisation of queuing theory, which is the fundamental theory of admission control. We have also discussed about IntServ and DiffServ, the major QoS schemes. In line with DiffServ, prioritisation and classification of calls are considered as necessary handlings in admission control schemes. The last section of the this chapter has provided an extensive review of the existing admission control schemes, which can be categorised as: Fixed Channel Allocation (FCA), Dynamic Channel Allocation (DCA) and Hybrid Channel Allocation (HCA) in the order of the simplest to the most complicated schemes. The strength and weakness of each scheme are analysed and discussed. The review has leaded us to the task presented in the following chapter.

Our work in chapter three started with discussion on the Threshold Access Sharing (TAS) scheme with call classification. Its handoff calls are prioritised by exclusively reserving a portion of bandwidth for them. The side effect of the approach is the high new call blocking probability. An adaptive admission control scheme named rate-based borrowing scheme is described. It is found to be efficient in bandwidth utilisation. Based on the schemes studied, we proposed an improved Threshold Access Sharing (iTAS) with a simplified rate-based borrowing scheme. Our simulation has shown that, compared with TAS, the scheme's performance was improved in terms of the handoff call dropping probability and new call blocking probability.

In chapter four, a new model on admission is proposed. The new idea is based on the observation that many criteria could affect an admission decision during a handoff or a new call request and most existing schemes take only one or two of them into consideration. Our new scheme is a weight-based admission control scheme with multiple criteria from the least important to the most important. Each criterion is assigned a weight and its total weight decides its admission level. This approach is novel and some initial simulation on the model has shown that the idea is actually feasible. The dropping rate for handoff calls is improved.

In chapter five, our focus is on hierarchical wireless IP networks. The hierarchical structure of the wireless IP networks enables efficient mobility handling and bandwidth utilisation. A complete admission model for hierarchical networks was proposed. The signalling load is divided among the primary tiers and a temporary tier. Each cell in the primary tiers applies a modified version of iTAS. The model has significantly reduced the handoff call dropping rate. The borrowing mechanism, making use of the characteristic of adaptive bandwidth for some applications, is used to compromise the new call blocking rate. Calls are allowed to negotiate bandwidth to ensure that the bandwidth utilisation is optimal.

Discussion of Future Work

Due to the time limit, the work and concepts developed in this thesis can be further studied. For example, in iTAS the thresholds can be calculated dynamically according to the current cell load; or different borrowing schemes can be applied. In the weight-based admission scheme, the impacts of each criterion on admission decisions should be thoroughly analysed in details before an ideal weighting is given. Bandwidth variation or bandwidth adaptation is an interesting characteristic of some IP-based applications. Therefore the study on how an application can cope with the change of its bandwidth allowance could produce useful results for adaptive admission control. In our admission control model for hierarchical networks, the allocation of the control and traffic channel could be studied more.

The cell planning in cellular networks is as vital as an efficient admission control scheme. Speed estimation is also essential to make the right tier selection in hierarchical networks. Although these are out of our research scope, we would like to stress their importance in our success.

REFERENCE

[1] O. Martikainen, presented at Next Generation Network Technologies, Rousse, Bulgaria, 2002.

[2] S. Weinstein and A. D. Gelman, "Networked Multimedia: Issues and Perspective," *IEEE Communications Magazine*, pp. 138-143, 2003.

[3] A. S. Tanenbaum, *Computer Networks*, vol. 1, 4th ed. Amsterdam: Prentice Hall PTR, 2003.

[4] R. Lloyd-Evans, *QoS in Integrated 3G Networks*, vol. 1, 1 ed. London: Artech House, 2002.

[5] W. C. Y. Lee, "Smaller cells for greater performance," *IEEE Communications Magazine*, vol. 29, pp. 19-30, 1991.

[6] I. Chih-Lin, L. J. Greenstein, and R. D. Giltin, "A microcell/ macrocell cellular architecture for low- and high-mobility wireless users," presented at Global Telecommunications Conference, 1991, 1991.

[7] X. Wu, "Supporting QoS in Overlaid wireless networks," in *Electrical and Computer Engineering*: University of California, 2001, pp. 174.

[8] M. Meo and M. A. Marsan, "Analysis of Hierarchical Cellular Networks with Multimedia Services and Different User Mobility Patterns," *Journal of Interconnection Networks*, vol. 1, 2000.

[9] S.-H. Wie, J. S. Jang, B.-C. Shin, and D. H. Cho, "Handoff Analysis of the Hierarchical Cellular System," *IEEE Transactions on Vehicular Technology*, vol. 49, 2000.

[10] M. Lott, M. Weckerle, W. Zirwas, H. Li, and E. Schulz, "Hierarchical cellular multihop networks," presented at 5th European Personal Mobile Communications Conference EPMCC 2003, Glasgow, Scotland, 2003.

[11] R. G. M. Communications, "Hierarchical cell structure of UMTS to offer global radio coverage," vol. 2005, 2005.

[12] R. Braden, D. Clark, and S. Shenker, "Integrated services in the Internet architecture: an overview," in *IETF RFC 1633*, 1994.

[13] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," in *IETF RFC 2475*, 1998.

[14] I. K. Park, "Per-Hop behaviors (PHBs)," in *QoS in packet networks*, vol. 1, 1st ed: Springer, 2005, pp. 177.

[15] A. Leon-Garcia and I. Widjaja, "Per hop behaviors," in *Communication networks - Fundamental concepts and key architecture*, vol. 1, 1st ed. Toronto: MrGraw-Hill, 2000, pp. 707.

[16] A. Leon-Garcia and I. Widjaja, "Bandwidth broker," in *Communication networks - Fundamental concepts and key architecture*, vol. 1, 1st ed. Toronto: MrGraw-Hill, 2000, pp. 708.

[17] R. Lloyd-Evans, "QoS in Integrated 3G Networks," in *Artech House mobile communications series*. London: Artech House, 2002, pp. 221.

[18] M. Zhang and T. S. Yum, "The non-uniform compact pattern allocation algorithm for cellular mobile systems," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 387-391, 1991.

[19] S. H. Oh, A. B, C. D, and E. F, "Prioritized channel assignment in a cellular radio network," *IEEE Transactions on Communications*, vol. 40, pp. 1259-1269, 1992.

[20] J. S. Engel and M. Peritsky, "Statistically optimum dynamic server assignment in systems with intergering servers," *IEEE Transactions on Vehicular Technology*, vol. 22, pp. 211-215, 1973.

[21] L. Anderson, "A simulation study of Sonle Dynamic Channel Assignment in Systems with Interfering Servers," *IEEE Transactions on Vehicular Technology*, vol. 22, pp. 210, 1973.
[22] M. Zhang, "Comparisons of channel assignment strategies in cellular mobile telephone systems," *IEEE Transactions on Vehicular Technology*, vol. 38, pp. 211-215, 1989.
[23] R. Singh, S. M. Elnoubi, and C. Gupta, "A new frequency channel assignment algorithm in high capacity mobile communications systems," *IEEE Transactions on Vehicular Technology*, vol. 31, 1982.
[24] P. Johri, "An insight into dynamic channel assignment in cellular mobile communication systems," *Euro. J. Operational Research*, vol. 74, pp. 70-77, 1994.
[25] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," *IEEE Communications Magazine*, vol. 11, pp. 934-940, 1991.
[26] T. J. Kahwa and N. Georganas, "A hybrid channel assignment scheme in large scale cellular-structured mobile communication systems," *IEEE Transactions on Communications*, vol. 26, pp. 432-438, 1978.
[27] T. S. Yum and W. Wong, "Hot spot traffic relief in cellular systems," *IEEE Journal on Selected Areas in Communications*, vol. 11, pp. 934-940, 1993.
[28] T. S. Yum and M. Schwartz, "The join-biased-queue rule and its applications to routing in computer communication networks," *IEEE Transactions on Communications*, pp. 505-511, 1981.
[29] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey," *IEEE Personal Communications*, vol. 3, pp. 10-31, 1996.
[30] S. S. Kuek, "Ordered dynamic channel assignment scheme with reassignment in highway microcell," *IEEE Transactions on Communications*, vol. 41, pp. 271-277, 1992.
[31] R. W. Nettleton, "A high capacity assignment method for cellular mobile telephone systems," presented at IEEE Vehicular Technology Conference, 1989.
[32] A. Gamst, "Some lower bounds for a class of frequency assignment problems," *IEEE Transactions on Vehicular Technology*, vol. 35, pp. 8-14, 1986.
[33] R. Beck and H. Panzer, "Strategies for handover and dynamic channel allocation in micro-cellular mobile radio communication systems," *IEEE Vehicular Technology*, vol. 1, pp. 178-185, 1989.
[34] D. C. Cox and D. O. Reudink, "Dynamic channel assignment in two dimension large-scale mobile radio systems," *Bell. Sys. Tech. J.*, vol. 51, pp. 1611-1628, 1972.
[35] D. C. Cox and D. O. Reudink, "A comparison of some channel assignment strategies in large mobile communication systems," *IEEE Transactions on Communications*, vol. 20, pp. 190-195, 1972.
[36] K. Okada and F. Kubota, "On dynamic channel assignment in cellular mobile radio systems.," presented at IEEE International Sympoum, 1991.
[37] C. L. I and P. H. Chao, "Distributed dynamic channel allocation algorithms with adjacent channel constraints," *PIMRC*, vol. B.2, pp. 169-175, 1994.
[38] C. L. I and P. H. Chao, "Local Packing - Distributed dynamic channel allocation at cellular base station," presented at IEEE Global Communications Conference, 1994.
[39] K. Okada and F. Kubota, "A proposal of a dynamic channel assignment strategy with information of moving directions," presented at IEICE Trans. Fundamentals, 1992.
[40] K. Okada and F. Kubota, "Performance of a dynamic channel assignment algorithm with information of moving direction in mobile communication systems," presented at IEICE Spring Nat'l. Conv., 1991.
[41] M. Serizawa and D. Goodman, "Instability and Deadlock of Distributed Dynamic channel allocation," presented at IEEE Vehicular Technology Conference, 1993.

[42] J. B. Punt and D. Sparreboom, "Mathematical models for the analysis of dynamic channel selection or indoor mobile wireless communications systems," *PIMRC*, vol. A.5, pp. 1081-1085, 1994.
[43] Y. Akaiwa and H. Andoh, "Channel segregation - a self-organized dynamic allocation mothod: application to TDMA/ FDMA microcellular system," *JSAC*, vol. 11, pp. 949-954, 1993.
[44] Y. Furuya and Y. Akaiwa, "Channel segregation - a distributed channel allocation scheme for mobile communication systems," *IEICE Trans.*, vol. 74, pp. 1531-1537, 1991.
[45] J. Sin and N. Georganas, "A simulation study of a hybrid channel assignment scheme for cellular land-mobile radio systems with Erlang-C service," *IEEE Transactions on Communications*, vol. COM-9, pp. 143-147, 1981.
[46] D. C. Cox and D. Reudink, "Increasing channel occupancy in large scale mobile radio systems: dynamic channel reassignment," *IEEE Transactions on Communications*, vol. 21, pp. 1302-1306, 1973.
[47] M. H. Chiu and M. A. Bassiouni, "Predictive schemes for handoff prioritisation in cellular networks based on mobile positioning," *IEEE Selected Areas Communications*, vol. 18, pp. 510-522, 2000.
[48] C. C. Chao and W. Chen, "Connection admission control for mobile multiple-class personal communication networks," *IEEE Journal Selected Areas in Communications*, vol. 15, pp. 1618-1626, 1997.
[49] R. Bolla, F. Davoli, and M. Marchese, "Bandwidth allocation and admission control in ATM networks with service separation," *IEEE Communications Magazine*, vol. 35, pp. 130-137, 1997.
[50] P. Agrawal, D. K. Anvekar, and B. Narendran, "Channel management policies for handovers in cellular networks," *Bell Labs Technical Journal*, vol. 1, pp. 96-109, 1996.
[51] R. Vijayan and J. M. Holtzman, "A model for analyzing handoff algorithims," *IEEE Transactions on Vehicular Technology*, vol. 42, pp. 351-356, 1993.
[52] O. Andrisano, M. Dell'Acqua, G. Mazzini, R.Verdone, and A. Zanella, "On the parameters optimization in handover algorithms," presented at IEEE Vehicular Technology Conference, Ottawa, Canada, 1998.
[53] M. Gudmundson, "Analysis of Handover algorithms," presented at IEEE Vehicular Technology Conference, St. Louis, MO, 1991.
[54] P. Seite, "Soft handoff in a DS-CDMA Microcellular Network," presented at IEEE Vehicular Technology Conference, 1994.
[55] S. L. Su and J. Y. Chen, "Performance analysis of Soft handoff in cellular networks," presented at IEEE PIMRC, Toronto, Canada, 1995.
[56] G. Senarath and D. Everitt, "Adaptive handoff algorithms using absolute and relative thresholds for cellular mobile communication systems," presented at IEEE Vehicular Technology Conference, 1994.
[57] B. Eklundh, "Channel utilization and blocking probability in a cellular mobile system with direct entry," *IEEE Transactions on Communications*, vol. 34, pp. 329-337, 1986.
[58] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedure," *IEEE Transactions on Vehicular Technology*, vol. VT-35, pp. 77-92, 1986.
[59] E. C. Posner and R. Guerin, "Traffic policies in cellular radio that minimize blocking of handoff calls," presented at 11th Teletraffic Congress, Kyoto, Japan, 1985.
[60] R. Guerin, "Queueing-blocking system with two arrival streams and guard channels," *IEEE Transactions on Communications*, vol. 36, pp. 153-163, 1988.

[61] M. J. Fischer and T. C. Harris, "A model for evaluating the performance of an integrated circuit and packet switched multiplexed structure," *IEEE Transactions on Communications*, vol. 24, pp. 195-202, 1976.
[62] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/ wireless networks," *IEEE Selected Areas Communications*, vol. 14, pp. 711-717, 1996.
[63] S. Tekinay and B. Jabbari, "A measurement based prioritization scheme for handovers in mobile cellular networks," *IEEE Journal Selected Areas in Communications*, vol. 10, pp. 1343-1350, 1992.
[64] C. J. Chang, T. T. Su, and Y. Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and reneging/ dropping," *IEEE/ ACM Transaction on Networking*, vol. 2, pp. 166-175, 1994.
[65] M. Sengoku, M. Kurata, and Y. Kajitani, "Applications of re-arrangement to a mobile radio communication system," *Journal of the Institute of Electronics and Communcation Engineers*, vol. J64-B, pp. 978-985.
[66] S. Boumerdassi and A. Beylot, "Adaptive channel allocation for wireless PCN," *Mobile Networks and Applications*, vol. 4, pp. 111-116, 1999.
[67] S. Choi and K. Shin, "Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks," presented at ACM SIGCOMM, Vancouver, 1998.
[68] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 1-12, 1997.
[69] S. Choi and K. G. Kin, "Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, pp. 882-897, 2002.
[70] F. Yu, "Quality of service provisioning in cellular mobile networks," in *Electrical and Computer Engineering*. Vancouver, Canada: The University of British Columbia, 2003, pp. 152.
[71] S. Bali and J. Korah, "Quality of Service in 3G Wireless Networks," Virginia Polytechnic Institute and State University.
[72] G.-S. Kuo, P.-C. Ko, and M.-L. Kuo, "A Probabilistics Resource Estimation and Semi-Reservation Scheme for Flow-Oriented Multimedia Wireless Networks," *IEEE Communications Magazine*, vol. 39, pp. 135-141, 2001.
[73] A. Mahmoodian and G. Haring, "Mobile RSVP with Dynamic Resource Sharing," presented at IEEE Wireless Commnunications and Networking Conference, 2000.
[74] A. S. Acampora and M. Naghshineh, "Control and Quality-of-Service provisioning in high-speed microcellular networks," *IEEE Personal Communications*, pp. 36-43, 1994.
[75] A. S. Acampora and M. Naghshineh, "QoS provisioning in microcellular networks supporting multimedia traffic," *IEEE INFOCOM*, pp. 1075-1085, 1995.
[76] T. Kwon, Y. Choi, C. Bisdikian, and M. Naghshineh, "Call admission control for adaptive multimedia in wireless/ mobile networks," presented at 1st ACM International Workshop on Wireless and Mobile Multimedia, Dallas, Texas, 1998.
[77] J. R. Moorman and J. W. Lockwood, "Wireless call admission control using threshold access sharing," presented at Global Telecommunications Conference, 2001.
[78] J. R. Moorman and J. W. Lockwood, "Real-time prioritized call admission control in a base station scheduler," presented at WOWMOM, 2000.
[79] M. El-Kadi, S. Olariu, and H. Abdel-Wahab, "A Rate-Based Borrowing Scheme for QoS Provisioning in Multimedia Wireless Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, pp. 156-166, 2002.

[80] K.-R. Lo, C. J. Chang, C. Chang, and C. B. Shung, "A combined channel assignment strategy in a hierarchical cellular systems," presented at IEEE ICUPC, San Diego, CA, 1997.
[81] K. Maheswari and A. Kumar, "Performance analysis of microcellization for supporting two mobility classes in cellular wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 49, pp. 321-333, 2000.
[82] R. Beraldi, S. Marano, and C. Mastroianni, "A reversible heirarchical scheme for microcellular systems with overlaying macrocells," presented at IEEE Infocom, 1996.
[83] E. Ekici and C. Ersoy, "Multi-tier cellular network dimensioning," *Wireless Networks*, vol. 7, pp. 401-411, 2001.
[84] X. Wu, "Supporting Quality of Service (QoS) in overlaid wireless networks," University of California Davis, 2001.
[85] C. Mihailescu, X. Langrange, and D. Zeghlache, "Analysis of a two-layer cellular mobile communication system," presented at IEEE Vehicular Technology Conference, Arizona, 1997.
[86] T. Salih and K. M. Fidanboyflu, "Performance analysis and modeling of two-tier cellular networks with queuing handoff calls," presented at IEEE International Symposium on Computers and Communications, 2003.
[87] K. J. Lin and Y. C. Tseng, "Channel sharing strategies in two0tier cellular PCS systems," *Computer Communications*, vol. 25, pp. 1333-1342, 2002.
[88] S. S. Rappaport and L. R. Hu, "Microcellular communication systems with hierarchical macrocell overlays: traffic performance models and analysis," presented at IEEE Conference, 1994.
[89] P. A. Whiting and D. W. McMillan, "Modeling for repacking in cellular radio," presented at 7th UK Teletrafic Symposium, 1990.
[90] B. Jabbari and W. F. Fuhrmann, "Teletraffic modeling and analysis of flexible heirarchical cellular networks with speed-sensitive handover strategy," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1539-1548, 1997.
[91] F. Valois and V. Veque, "QoS-oriented channel assignment strategy for hierarchical cellular networks," *IEEE PIMRC*, vol. 2, pp. 1599-1603, 2000.
[92] H. M. Tsai, A. C. Pang, Y. C. Lin, and Y. B. Lin, "Channel assignment for hierarchical cellular networks," presented at International Conference on Parallel Processing, 2003.
[93] I. Rajput and A. O. Fapojuwo, "Performance of two-tier cellular networks with macrocell size adjustment," *IEE Electronics letters*, vol. 39, pp. 1857-1859, 2003.
[94] A. S. Anpalagan and I. Katzela, "Overlaid cellular system design with cell selection criteria for mobile wireless users," presented at 1999 IEEE Canadian Conference on Electrical and Computer Engineering, Alberta, Canada, 1999.
[95] H.-C. Lin and S.-S. Tzeng, "Double-threshold admission control in cluster-based micro-picocellular wireless networks," *IEEE Vehicular Technology*, vol. 3, pp. 1440-1444, 2000.
[96] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," *IEEE Communications Magazine*, vol. 29, pp. 42-46, 1991.
[97] G. Cimone, D. D. Weerakoon, and A. H. Aghvami, "Performance evaluation of a two layer hierarchical cellular system with variable mobility user using multiple class applications," presented at IEEE Vehicular Technology Conference, 1999.
[98] X. Lagrange and P. Godlewski, "Performance of a hierarchical cellular network with mobility-dependent handover strategies," presented at IEEE Vehicular Technology Conference, 1999.
[99] Y. Pan, M. Lee, J. B. Kim, and T. Suda, "Smooth handoff scheme for stream media with bandwidth disparity in wireless cells," *IEEE Communications*, pp. 9-16, 2003.

[100] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S.-Y. Wang, and T. L. Porta, "HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ ACM Transaction on Networking*, vol. 10, pp. 396-410, 2002.
[101] A. T. Campbell, G. Gomez, S. Kim, Z. Turanyi, C.-Y. Wan, and A. Valko, "Design, implementation and evaluation of Cellular IP," *IEEE Personal Communications*, vol. 8, pp. 42-49, 2000.
[102] A. Helmy, "A multicast-based protocol for IP Mobility IPv6," presented at ACM SIGCOMM Second International workshop on Networked Group Communication, Palo Alto, 2000.
[103] Y. Cheng and W. Zhuang, "DiffServ resource allocation for fast handoff in wireless mobile Internet," *IEEE Communications Magazine*, vol. 40, pp. 130-136, 2002.
[104] K. Brown and S. Singh, "M-TCP: TCP for mobile cellular networks," presented at ACM SIGCOMM Computer Communication Reviews, 1997.
[105] T. Goff, J. Moronski, and D. Phatak, "Freeze-TCP: A true end-to-end enhancement mechanism for mobile environments," presented at INFOCOM, 2000.
[106] A. Bakre and B. R. Badrinath, "I-TCP: Indirect TCP for mobile hosts," presented at International Conference on Distributed Computing Systems, 1995.
[107] H. Balakrishnan, S. Seshan, and R. H. Katz, "Improving reliable transport and handoff performance in cellular wireless networks," *ACM Wireless Networks*, vol. 1, pp. 469-481, 1995.
[108] A. S. Acampora and M. Naghshineh, "An architecture and methodology for mobile-excuted handoff in cellular ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 1365-1374, 1994.
[109] Y. Xiao, C. L. P. Chen, and Y. Wang, "Quality of Service and call admission control for adaptive multimedia services in wireless/ mobile networks," *IEEE NAECON*, vol. 4, pp. 214-220, 2000.