LTE Optimization and Resource Management in Wireless Heterogeneous Networks

submitted to the
Faculty of Physics and Electrical Engineering,
University of Bremen

for obtainment of the academic degree

Doktor-Ingenieur (Dr.-Ing.)

Dissertation

by

Umar Toseef
from Sheikhupura, Pakistan

First assessor:	Prof. Dr. rer. nat. habil. Carmelita Görg
Second assessor:	Prof. Dr.-Ing. Christian Wietfeld
Submission date:	February 5, 2013

I assure that this work has been done solely by me without any further help from others except for the official support of the Communication Networks group of the University of Bremen and Technical University of Hamburg. The literature used is listed completely in the bibliography.

Bremen, 30th of January 2013

(Umar Toseef)

Dedication

This work is dedicated to my parents for their warmest affection, compelling motivation and to their sacrifices they made for the betterment of my life. And a special dedication goes to my wonderful wife who has always supported me in every endeavor, whose selflessness has been my inspiration and who has enriched my life with her pure love and deep understanding.

Preface

I have accumulated many debts on the path toward completion of this thesis. My supervisor Prof. Dr. Carmelita Görg rightfully deserves the foremost acknowledgements. She has been inspiring and patient from the beginning of this work. Her valuable guidance and continuous encouragement provided the essential means to carry out this work. She has been a support for my intellectual and geographical journeys which helped me to be in touch with state-of-the-art work activities in my field.

I am also very much appreciative of Prof. Dr.-Ing. Andreas Timm-Giel for his enthusiastic support of my vision for this research work. I am greatly indebted to him for his sound advices and gentle prodding which was not only limited to my thesis but extended to my other scientific work.

Besides, it is with pleasure and deep gratitude that I acknowledge the good advice and support of my colleagues within the ComNets department of the University of Bremen and of the Hamburg University of Technology. Dr. Yasir Zaki, Dr. Xi Li, and Dr. Thushara Weerawardane supplied their continuous motivation for this work, participated in many technical discussions and exchanged valuable research ideas. Dr. Phuong Nga Tran deserves a special thanks for her great support in the analytical portion of this work. I am also grateful to Asanga Udugama, Dr. Koojana Kuladinithi and Ming Li for being great project partners as well as for their support at both academic and personal level. Finally, I extend the warmest thanks to all other colleagues especially to Dr. Andreas Könsgen, Amanpreet Singh, Markus Becker, Liang Zhao, Dr. Mohammad Muttakin Siddique, Chunlei An, Safdar Marwat, and Thomas Pötsch for their great help to me.

Umar Toseef

Abstract

Mobile communication technology is evolving with a great pace to offer richer user experience and make an operator's business more profitable at the same time. The development of the Long Term Evolution (LTE) mobile system by 3GPP is one of the milestones in this direction. 3GPP specifications for LTE mobile systems serve as the high level standards leaving room for improvements by researchers. This work highlights a few of such areas in the LTE radio access network where the proposed innovative mechanisms can substantially improve overall system performance. This includes a novel air interface scheduler design which can coordinate with the core network entities to avoid imminent network congestion. Another proposed air interface scheduling algorithm exhibits an adaptive behavior and reacts to network load conditions in optimizing the scheduler operations. Similarly, packet queue management for buffers of the LTE air interface scheduler is an important subject which has significant impact on user perceived QoE and inter-site handover operations. The thesis discusses all these topics in great detail and proposes practical solutions which are proven to be effective with the help of simulation based analysis.

The advent of mobile devices with multiple radio interfaces has increased the opportunity for users to stay connected through any available network type. This makes operators realize that the integration of 3GPP networks (e.g., LTE, HSPA etc.) and non-3GPP networks (e.g., WLAN, WiMAX etc.) is inevitable. This integration would enable operators to offload the select user traffic from 3GPP networks to the integrated WLAN networks with overlapped coverage. However, it comes with the responsibility of the operators to actively manage the bandwidth resources of the two network types in order to get most out of this integration. The thesis addresses this issue in immense detail. For this purpose, a comprehensive system architecture is developed as an overlay of the 3GPP defined SAE architecture. The proposed architecture serves as a framework for implementing network bandwidth resource management mechanisms. In addition, this work also proposes several resource management mechanisms which can operate in conjunction with the purported overlay system architecture. The performance of these mechanisms is evaluated using a heterogeneous network simulator, developed by the author in this work.

Another contribution of this thesis is the development of an analytical solution for the optimal network resource allocation problem. The proposed solution is based on ‘Linear Programming’ which is a popular mathematical optimization technique. With the help of simulation studies, the analytical solution is shown to outperform other discussed resource management mechanisms in improving user QoE and network capacity. In order to make resource allocation operations less processing-intensive and more practical for real world products, alternative heuristic based algorithms are also proposed in this work which can achieve near-optimal performance.

The concepts, mechanisms, and the investigations presented in this work are of great value to operators to carry out optimization of overall LTE network operations in general and that of LTE radio network in particular. In addition, the concept of user multihoming in heterogeneous networks along with the proposed system architecture to support efficient resource management operations provide an excellent framework for operators in performing traffic offloading. A number of developed resource management mechanisms and their proven effectiveness, in achieving user QoE enhancement and network capacity improvement, serve as a motivation for operators to further exploit the hidden potential of integrated heterogeneous networks.

Kurzfassung

Mobile Kommunikationstechnik entwickelt sich mit großer Geschwindigkeit, um eine besseres Nutzungserlebnis bereitzustellen und gleichzeitig das Geschäft eines Netzbetreibers profitabler zu machen. Die Entwicklung das mobilen Long Term Evolution (LTE)-Systems durch 3GPP ist einer der Meilensteine in dieser Richtung. 3GPP-Spezifikationen für mobile LTE-Systeme dienen als Standards auf einer hohen Ebene, die Platz für Verbesserungen durch Forscher lassen. Diese Arbeit beleuchtet einige solcher Gebiete im LTE-Funkzugangsnetz, wo die vorgeschlagenen innovativen Mechanismen das gesamte System-Leistungsverhalten wesentlich verbessern kann. Dies schließt einen neuartigen Entwurf der Luftschnittstelle ein, die sich mit den Einheiten des Kernnetzes koordinieren kann, um eine bevorstehende Überlastung des Netzes zu vermeiden. Ein anderer vorgeschlagener Scheduling-Algorithmus für die Luftschnittstelle weist ein adaptives Verhalten auf und reagiert auf Lastbedingungen durch Optimierung der Reaktion des Schedulers. In ähnlicher Weise ist das Management der Paketwarteschlangen für die Puffer der LTE-Luftschnittstelle ein wichtiges Thema, das wesentliche Auswirkungen auf die vom Benutzer wahrgenommene QoE und Handover-Vorgänge zwischen einzelnen Standorten hat. Die Arbeit diskutiert all diese Themen ausführlich und schlägt praktische Lösungen vor, deren Effektivität mit Hilfe von simulationsbasierten Analysen bewiesen wird.

Die Einführung mobiler Geräte mit mehreren Funkschnittstellen hat Benutzern zusätzliche Möglichkeiten gegeben, mit Hilfe jedes verfügbaren Netztypes verbunden zu bleiben. Dies führt dazu, dass Betreiber die Integration von 3GPP-Netzen (z.B. LTE, HSPA usw.) und Nicht-3GPP-Netzen (z.B. WLAN, WiMAX) als unvermeidbar erkennen. Diese Integration würde es Betreibern ermöglichen, ausgewählten Benutzerverkehr von 3GPP-Netzen auf integrierte WLAN-Netze mit überlappender Abdeckung umzuschichten. Allerdings ergibt sich aus der Verantwortung des Betreibers, die Bandbreiten-Ressourcen der zwei Netztypen aktiv zu steuern, um den größten Nutzen aus dieser Integration zu erhalten. Die Arbeit behandelt diese Aspekte sehr ausführlich. Zu diesen Zweck wird eine umfassende Systemarchitektur als Überlagerung der durch 3GPP definierten SAE-Architektur entwickelt. Die vorgeschlagene Architektur dient als Framework zur Implementierung von Mechanismen zum Ressourcen-Management von Netzbandbreite. Zusätzlich schlägt diese Arbeit auch verschiedene Mechanismen zum Ressourcen-Management vor, die in Verbindung mit der vorgesehenen überlagerten Systemarchitektur arbeiten können. Das Leistungsverhalten dieser Mechanismen wird mit Hilfe eines vom Autor in dieser Arbeit entwickelten heterogenen Netzsimulators bewertet.

Ein weiterer Beitrag dieser Arbeit ist die Entwicklung einer analytischen Lösung für das Problem der optimalen Zuweisung von Netzressourcen. Die vorgeschlagene Lösung basiert auf Linearer Programmierung, einem verbreiteten mathematischen Optimierungsverfahren. Mit Hilfe der simulativen Untersuchungen wird gezeigt, dass die analytische Lösung andere diskutierte Mechanismen zum Ressourcenmanagement bei der Verbesserung der QoE und der Netzkapazität übertrifft. Um Techniken für die Ressourcenzuweisung weniger verarbeitungsintensiv und praxisnäher für reale Produkte zu gestalten, werden auch alternative heuristische Verfahren in dieser Arbeit vorgeschlagen, die ein nahezu optimales Leistungsverhalten erzielen können.

Die Konzepte, Mechanismen und Untersuchungen, die in dieser Arbeit gezeigt werden, sind von großem Wert für Betreiber, um Optimierungen des gesamten LTE-Netzbetriebes durchzuführen, insbesondere des LTE-Funknetzes. Zusätzlich stellt das Konzept des Benutzer-Multihomings in heterogenen Netzen zusammen mit der vorgestellten Systemarchitektur zur Unterstützung effizienten Resourcemanagements ein hervorragendes Framework für Betreiber zur Durchführung von Verkehrsumschichtung dar. Eine Anzahl entwickelter Mechanismen für das Ressourcenmanagement und deren bewiesene Effektivität beim Erreichen von Verbesserungen der Benutzer-QoE und der Netzkapazität dienen als Motivation für Betreiber, das versteckte Potenzial integrierter heterogener Netze weiter auszuschöpfen.

Contents

List of Figures

List of Tables

List of Abbreviations

2G	The Second Generation Mobile Communication System
3G	The Third Generation Mobile Communication System
3GPP	3rd Generation Partnership Project
ACK	Acknowledgement
AMC	Adaptive Modulation and Coding
AP	Access Point (WLAN)
ARP	Allocation and Retention Priority
ARP	Address Resolution Protocol
ARQ	Automatic Repeat Request
BE	Best Effort
BER	Bit Error Rate
BET	Blind Equal Throughput
BLER	Block Error Rate
BSC	Base Station Controller
BTS	Base Transceiver Station
CDF	Cumulative Distribution Function
CN	Core Network
CQI	Channel Quality Indicator
CTS	Clear To Send
DCF	Distributed Coordination Function
DE	Decision Making Entity
Diffserv	Differentiated services
DIFS	DCF Inter-Frame Space
DL	Downlink
DSCP	Differentiated Services Code Point
DVB	Digital Video Broadcasting
EDGE	Enhanced Data for GSM Evolution
EE	Execution and Enforcement Entity
EF	Expedited Forwarding
EMA	Exponential Moving Average
eNodeB	enhanced NodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
ETSI	European Telecommunication Standards Institute
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
EXP	Exponential Distribution
FD	Frequency Domain
FDD	Frequency Division Duplex
FDM	Frequency Domain Multiplexing
FDMA	Frequency Division Multiple Access

FM	Flow Management
FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GGSN	Gateway GPRS Support Node
GMSK	Gaussian Minimum Shift Keying
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
GTP	GPRS Tunneling Protocol
GTP-U	GPRS Tunneling Protocol User Plane
HARQ	Hybrid Automatic Repeat Request
HCF	Hybrid Coordination Function
HD	High Definition
HLR	Home Location Register
HO	Handover
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSS	Home Subscriber Server
HSUPA	High Speed Uplink Packet Access
HTTP	Hypertext Transfer Protocol
IE	Information Management Entity
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
ISP	Internet Service Provider
ITU	International Telecommunication Union
ITU-T	Telecommunication Standardization Sector of ITU
LAN	Local Area Network
LP	Linear Programming
LTE	Long Term Evolution
MAC	Medium Access Channel
MaxT	Maximum Throughput
MCoA	Multiple Care-of Address
MCS	Modulation and Coding Scheme
ME	Mobile Equipment
MIMO	Multi Input Multi Output
MIP	Mobile IP
MME	Mobility Management Entity
MOS	Mean Opinion Score
MS	Mobile Station
MSC	Mobile Switching Center
non-GBR	non-Guaranteed Bit Rate
NSS	Network and Switching Subsystem
OFDMA	Orthogonal Frequency Domain Multiple Access
PCF	Point Coordination Function
PCRF	Policy and Charging Rules Function
PDA	Personal Digital Assistant
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDN-GW	Packet Data Network Gateway
PDU	Protocol Data Unit
PF	Proportional Fair
PHB	Per-Hop Behaviors
PHY	Physical Layer

PLR	Packet Loss Rate
PRB	Physical Resource Block
PRBs	Physical Resource Blocks
QCI	QoS Class Identifier
QoE	Quality of Experience
QoS	Quality of Service
RD	Random Direction
RED	Random Early Detection
RFC	Request for Comments
RLC	Radio Link Control
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RRC	Radio Resource Control
RTS	Ready To Send
RWP	Random Way Point
SAE	System Architecture Evolution
SC-FDMA	Single Carrier Frequency Domain Multiple Access
S-GW	Serving Gateway
SIFS	Short Inter-Frame Space
SIM	Subscriber Identity Module
SIR	Signal-to-Interference Ratio
TBS	Transport Block Size
TCP	Transmission Control Protocol
TD	Time Domain
TDMA	Time Division Multiple Access
TNL	Transport Network Link
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunication System
USIM	User Service Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VoIP	Voice over Internet Protocol
WCDMA	Wideband Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WWAN	Wireless Wide-Area Networks

List of Symbols

Symbol	Meaning
A_{xx}	Constant real number value
$C^{\text{n-GBR}}$	Available bandwidth capacity for non-GBR bearer traffic at last-mile S1 link
$C_{S1_{UL}}$	Total uplink bandwidth capacity of last-mile S1 link
d_i	IP packet packet size of user *i*
$E[D]$	Average delay experienced by a successfully transmitted packet
$\widehat{E[D]}$	Extended value of $E[D]$ in a WLAN network of users with different PHY data rate
e_i	Fractional throughput share of a user *i* in overall WLAN access point throughput
$E[slot]$	Average length of slot time
$\widehat{E[slot]}$	Extended value of $E[slot]$ in a WLAN network of users with different PHY data rate
E_r^{video}	Measured egress data rate of video traffic
E_r^{voice}	Measured egress data rate of voice traffic
$E[X]$	Average number of slot times used for a successful packet transmission
F_i	Binary variable; it represents whether a user *i* is active in WLAN
g	Packet queue occupancy
G	Mean IP packet size for traffic flows of active WLAN users
h	Bearer throughput reported by IE entity
L	A set of access network links
M_i^{AF}	Priority metric of ’Adaptive Fair’ scheduler for a user *i*
M_i^{BET}	Priority metric of ‘Blind Equal Throughput’ scheduler for a user *i*
M_i^{MaxT}	Priority metric of ’MaxT’ scheduler for a user *i*
M_i^{PF}	Priority metric of ‘Proportional Fair’ scheduler for a user *i*
MSS	Maximum TCP segment size
n	Number of active users in WLAN
N_{PDCP}	Number of PDCP PDUs
N_{PDCP_L}	Number of large sized PDCP PDUs which cause packet fragmentation

Symbol	Meaning
	at the IP layer
N_{PDCP_S}	Number of small sized PDCP PDUs which cause no packet fragmentation at the IP layer
N_{RLC}	Number of RLC PDUs
p	Probability of a collision seen by a packet being transmitted on the WLAN access medium
p_a	Packet drop probability
PLR	Packet loss rate
P_s	Probability that an occurring packet transmission is successful
P_{suc}	Probability that transmission occurring on channel is successful
P_{tr}	Probability of having at least one transmission in the considered time slot
q	Probability of successful packet transmission
$Q_{k,i}$	QoS weight which represents the relative priority of a bearer i of service class k
r_i	Achievable data rate for a user i when he is the only active user associated to an access point
$\widehat{R}_i$	Instantaneous achievable data rate based on the channel quality for bearer i
$\overline{R_i}(t)$	The average throughput value of user i
$R_{j,l}$	Data rate carried over the access link l to user j
R_j^{lte}	Data rate in carried over the LTE access link to user j
R_j^{wlan}	Data rate in carried over the WLAN access link to user j
RTT	TCP segment round trip time
T_{ACK}	Duration of ACK control frame in IEEE 802.11 networks
T_{backoff}	Time spent by a user in back-off phase
T_c	Average time the channel is sensed busy due to a collision
T_{CTS}	Duration of CTS control frame in IEEE 802.11 networks
$\widehat{T}_c$	Extended value of T_c in a WLAN network of users with different PHY data rate
T_{data}	Time required to transmit WLAN MAC frame including PHY headers
T_{DIFS}	Duration of DIFS frame space in IEEE 802.11 networks
$T_{E[P]}$	Time to transmit a data packet of mean size $E[P]$
$\widehat{T}_{E[P]}$	Time to transmit a data packet of mean size $E[P]$ in the presence of users with different PHY data rate
T_H	Time to transmit protocol header data of WLAN MAC & PHY
T_{RTS}	Duration of RTS control frame in IEEE 802.11 networks

Symbol	Meaning
T_s	Time required to transmit one packet excluding any collision and back-off delays
T_{SIFS}	Duration of SIFS frame space in IEEE 802.11 networks
$\widehat{T_s}$	Extended value of T_s in a WLAN network of users with different PHY data rate
$\widetilde{T_s}$	T_s including back-off time
U	A set of multihomed users
$w_j(t)$	Weight factor for cell j
$W_{\max}$	Maximum value of contention window in IEEE 802.11 networks
$W_{\min}$	Minimum value of contention window in IEEE 802.11 networks
Y	Mean uplink user throughput in WLAN network
$\check{Y}$	Maximum possible value of Y
Z	Total number of active and inactive users in WLAN network
α	Data rate dependent part of the LTE link cost
β	Data rate independent part of the LTE link cost
$\check{\gamma}$	Maximum packet queuing delay for PDCP buffer
$\gamma_{\text{de-jitter buffer}}$	Size of play-out or de-jitter buffer
$\gamma_{\text{tcp reorder buffer}}$	Size of TCP reordering buffer
δ	Propagation delay of electromagnetic waves
Δ_{SM}	Safety margin value
ε	Traffic amount sent by DE_n entity to LTE access link of a user
$\widehat{\varepsilon}$	Adjusted amount of traffic sent by DE_n entity to LTE access link of a user
ζ_c	Throughput to be allocated to a cell c as seen at Uu interface
$\widetilde{\zeta_c}$	Throughput to be allocated to a cell c as seen at TNL
θ_i	Length of time slot assigned to a user i for exclusively transmission
$\widehat{\kappa}$	Sum of κ_c of cells having surplus throughput
κ_c^{BET}	The TD priority term for 'Blind Equal Throughput' scheduler
κ_c^{PF}	The TD priority term for 'Proportional Fair' scheduler
λ_j	Minimum aggregated data rate demand of a traffic flow destined to user j
Λ_j	Maximum aggregated data rate demand of a traffic flow destined to user j
μ	Target occupancy for the PDCP buffer
$\widehat{\mu}$	Difference of current and target PDCP buffer occupancy
$\widehat{\nu}$	Overall surplus throughput of all cells in eNodeB;
ν_c	Surplus from the allocated throughput to a cell c
$\widetilde{\rho}$	Estimation of the effective cell throughput at TNL
ρ	Measured cell throughput value at the RLC layer

Symbol	Meaning
ρ^{OH}	Throughput of TNL protocol overhead bits
σ	Duration of ‘slot time’ as defined in IEEE 802.11a standard
τ	Probability that a node transmits in a randomly chosen slot time
ϕ_j	Path cost of WLAN access link in [sec/kbps] for user j
χ	Binary variable; it represents the product of any two binary
ψ	PDCP buffer occupancy reported by IE entity
Ω_l	Available resources on access network l
$\omega^{\mathrm{ch}}_{\mathrm{user}i}$	Downlink throughput of the user i using ’channel aware’ service decipline
$\omega^{\mathrm{ch}}_{\mathrm{AP}}$	Downlink throughput of WLAN access point using ‘channel aware’ service decipline
$\omega^{\mathrm{rr}}_{\mathrm{user}i}$	Downlink throughput of the user i using ‘round-robin’ service decipline
$\omega^{\mathrm{rr}}_{\mathrm{AP}}$	Downlink throughput of WLAN access point using ‘round-robin’ service decipline
$\omega^{\mathrm{trr}}_{\mathrm{user}i}$	Mean uplink throughput of the user i using ‘time round-robin’ service decipline

1 Introduction

This chapter provides a brief introduction of the Long Term Evolution (LTE) and heterogeneous networks, which are the main topics of discussion in this thesis. It also highlights the motivation for this work and lists the main technical contributions made by this thesis. In addition, it offers an overview of the thesis structure along with the brief description of each chapter.

1.1 LTE and Heterogeneous Networks

The Long Term Evolution (LTE) of the Universal Mobile Telecommunication System (UMTS) is one of the latest milestones achieved in advancing series of mobile telecommunication systems by the Third Mobile Generation Partnership Project (3GPP). LTE is well positioned today, and is already meeting the requirements of future mobile networks. LTE employs orthogonal frequency division multiplexing (OFDM) as its radio access technology, together with advanced antenna technologies like multiple-input and multiple-output (MIMO), spatial multiplexing, and beam-forming. The particular choice of OFDM technology not only helps LTE fulfill the requirement for spectrum flexibility but also enables cost-efficient solutions for very wide carriers with high peak rates. By making use of state-of-the-art communication technologies, LTE achieves 3 to 4 time higher spectral efficiency as compared to HSPA (Release 6) networks. This makes LTE an excellent choice for the network operators because an efficient utilization of scare radio spectrum resources brings twofold benefit. First, it enhances user Quality of Experience (QoE) by satisfying application Quality of Service (QoS) requirements. Second, it increases network capacity by serving more users within the available radio spectrum bandwidth.

In addition to LTE, 3GPP has also defined an IP-based, flat core network architecture. The architecture is based on an evolution of the existing 2G/3G core network, with a particular focus on simplified operations, cost-efficient deployment and the capability to support uptake of mass-market multimedia services. This architecture, called Evolved Packet Core (EPC), eliminates the need for circuit-switching by providing IP-based solutions for all types of voice, video, and data services. Owing to the fact that each service type has its own QoS demands, the LTE-EPC has adopted an effective class-based QoS concept. This provides a foundation for operators to offer service differentiation, depending on the type of application or subscription. This work further exploits the potentials of LTE access technology and proposes a few sophisticated mechanisms to enhance the overall system performance.

It is not only the mobile telecommunication systems which have evolved to offer LTE; the technology of handheld mobile devices has also made significant advancements in the recent years. This has made mobile broadband subscriptions to increase rapidly worldwide. Every year, hundreds of millions of users are subscribing for mobile broadband services. This is because a number of broadband applications have been redesigned to substantially enhance user experience by taking advantage of mobility support and large data rates of new access technologies. Such applications include social-networking (e.g., Facebook, Google+, Twitter etc.), multi-player gaming, content sharing (e.g., Youtube, Cloud Storage etc.), WebTV, video telephony, search engines etc. The traffic data generated by rapidly increasing broadband subscribers due to use of the aforementioned applications is manifold higher in volume compared to pure voice traffic. The existing 3GPP mobile communication networks (e.g., HSPA and LTE) are already facing difficulties to meet this high demand for wireless data. This has made users and operators to rely onto Wireless Local Area Networks (WLAN) based on IEEE 802.11 set of standards. The modern WLANs are capable of offering very high data rates but provide a small coverage area and limited mobility support. Therefore, they are more suitable to areas with highly dense demand for high data rate wireless access with limited mobility support. On the other hand, 3GPP networks are designed to provide ubiquitous coverage through mobility support and therefore well suited to areas with moderately dense demand for wireless access with high mobility. In this way, WLAN and 3GPP networks can complement each other in making high-speed Internet access a reality for a large population. This work discusses how the integration of these two technology types can be realized, what benefits are possible for the users and operators from this integration, and what are the challenges involved in the resource management of these heterogeneous networks. This work also proposes several mechanisms for efficient resource management of heterogeneous networks and evaluates their performance with the help of simulation based studies.

1.2 Technical Contributions

The LTE air interface scheduler bears a significant importance in the LTE system. It intelligently schedules the radio resources to deliver the required QoS to the active radio bearers. The scheduling algorithms employed for this purpose have a substantial impact on the performance of the individual base station and on the overall LTE radio access network. The scheduler design must take different considerations into account like service type, application QoS demands, and throughput fairness among same user types etc. However, its operation remains indifferent to the core network state. This work proposes an enhanced design of the air interface scheduler which actively coordinates with core network in order to efficiently allocate scarce radio resources. This coordination enables air interface scheduler to foresee congestion situations in the core network and take appropriate measures during the scheduling process to circumvent it. This keeps the network in a stable state, enhances radio network coverage, and improves user QoE.

The LTE air interface scheduling is a complex process whose optimization must involve certain compromises. For example, if a scheduling algorithm optimizes the system capacity, it fails to offer throughput fairness among the users and vice versa. Therefore, it remains a hard choice for network operators to choose the right scheduling algorithm for a certain base station. This work relieves network operators by proposing an adaptive scheduling algorithm which dynamically changes its behavior based on different network load conditions. This ensures an optimized air interface scheduling operation in all situations without requiring human intervention. In addition, this work also addresses the packet queue management issues related to the LTE air interface scheduler. With the help of the proposed mechanisms, not only the user QoE is improved for both uplink and downlink communication but also the inter-site handover process for mobile users is ameliorated. The aforementioned enhancements of the LTE access interface have also been published in the proceedings of several reputed scientific conferences, e.g., [TWG^{+}13], [U. 12b], [U. 11b], [U. 11a], and [U. 12a].

WLAN access technology has been widely deployed in urban areas which allow mobile devices to access the Internet as long as they remain in the limited coverage of a WLAN access point. During the other times, these mobile devices automatically connect to the 3GPP wide-area networks for the Internet access. Though this strategy achieves the data offloading and helps alleviate congestions in 3GPP networks, it allows a limited multi-access functionality. In order to fully exploit multi-access functionality, WLAN access points must be integrated with the 3GPP networks. With this integration not only a seamless mobility is achieved between the two access technology types but it also opens opportunities for network operators to optimize their network operations and enhance the user QoE. 3GPP has already realized this potential performance gain and has published the standards to allow integration of non-3GPP access technologies (i.e., WLAN, WiMAX etc.) to the existing the 3GPP access technologies. This feature of the System Architecture Evolution (SAE) was introduced in 3GPP release 8 standards.

Following the aforementioned 3GPP standards for integration, one can develop heterogeneous wireless access networks where mobile devices are provided with seamless mobility between 3GPP and non-3GPP networks, allowing a continuity of existing sessions. However, this standard still limits the multihoming capability of the users, i.e., they cannot access and use the two network types simultaneously. This work extends the 3GPP proposals to realize multihoming support for mobile devices in wireless heterogeneous networks. In addition, this work also discusses the problem of network resource allocation in integrated heterogeneous wireless access networks. More specifically, this problem involves the network decision of how much data rate should be served on each access link of a multihomed user. In order to address this problem, a comprehensive system architecture is proposed to actively manage the traffic flows of the users in the heterogeneous wireless network environments. This architecture overlays the 3GPP defined SAE architecture and provides all necessary support to execute sophisticated procedures related to efficient network traffic flow management of multihomed users. Based on this architecture several network resource management mechanisms have been proposed in this work, a few of them have also been published in [TZGTG12a], [TZGTG12b], and [TZZ$^+$12]. However, due to the utmost importance of network resource allocation in integrated heterogeneous network, the investigations are extended in this area by developing analytical models of the air interface of WLAN and LTE access technologies. These models pave the way to employ mathematical optimization techniques like ‘Linear Programming’ in network resource allocation problems. The performance of these mechanisms for optimized resource allocation is evaluated by their implementation and then integration into the developed simulation model. These mechanisms are shown to offer a superior user QoE and extended network capacity. Owing to the fact that Linear Programming based solutions are processing-intensive, alternative heuristic based techniques for network resource allocations are also developed within this work. A few details of this work have been published in [TZTGG12] and [TGTG12].

The extensions of the SAE architecture proposed in this work to realize user multihoming in heterogeneous networks have been validated through an implementation of a network simulator. For this purpose, the basic OPNET simulation models of LTE [Zak12] and WLAN [OPN13] have been extensively evolved by, e.g., incorporating a new WLAN channel model and the IETF’s proposed extension of Mobile IPv6, as well as multi-interface mobile device models. The simulation model also has a full implementation of the proposed overlay system architecture required for resource management in heterogeneous networks. In addition, this work also contributes to the implementation of several popular user QoE evaluation mechanisms to the OPNET simulation software. The resulted heterogeneous network simulator has been used to carry out a variety of simulation based studies which served as proof-of-concept for the mechanisms proposed in this work. The developed simulation model of integrated LTE and WLAN networks is also a valuable contribution of this work to the scientific research community. For example, it has been used to contribute a number of findings and mechanisms in the 'Open Connectivity Service (OConS)' work package of the SAIL European project [Sp13]. Furthermore, the simulator is also in active use by fellow researchers and institutions to extend the investigations in this field, e.g., [X. 12], [X. 13], [M. 10b], [M. 10a], [HWG$^+$12], and [ZZU$^+$11].

In addition to the above mentioned contributions, the author has also been involved in a number of other research activities which are not discussed in this thesis for the sake of brevity. Considering the fact that these activities belong to the research field which is also shared by this work, it is worth mentioning them here to further intrigue the interest of reader in this work. For example, the details about the test-bed implementation of user mobility mechanisms in heterogeneous networks along with the support of basic flow management can be found in [U. 07b]. Aforementioned work also involved the development of a mechanism which processes the link layer performance metrics to assist in making timely vertical handovers [U. 07a]. Similarly, another study on performance evaluation of PMIPv6 in real a test-bed environment can be accessed in [UIT$^+$09]. The motivation behind this study was the fact that Proxy Mobile IPv6 (PMIPv6) has gained a lot of attention due to its adoption in the 3GPP SAE architecture and its feasibility in the mobility management of low-end user devices.

An interesting work on employing 'Game Theory' in user-centric network selection is available in [M. 10b]. The investigations in this work have been extended to introduce a new concept of telecommunication network paradigm where users are not bound to long term contracts with operators. Instead, the user service requests are auctioned to competing operators through a third party platform [M. 10a].

No discussion about network performance, access interface selection, and resource management can be concluded without discussing the mechanisms of user QoE evaluation. A work which explores state-of-the-art mechanisms of service quality evaluation and extends them to develop a user satisfaction function for use in network selection, has been carried out in [TKGTG11] and [KT11]. Another piece of work in this area has been published in [X. 12] and [X. 13], where the user-centric bandwidth resource management has been investigated. Moreover, based on the aforementioned work, a comprehensive guide was prepared on the development of simulation models for heterogeneous networks which was accepted as a chapter in the book "Simulation in Computer Network Design and Modeling: Use and Analysis" [TK12].

It is said, "*Necessity is the mother of invention*". This proverb appeared to hold when doing research on heterogeneous networks. As a result several inventions were made on the course of developing efficient mechanisms for user multihoming, robust mobility management, and network resource management. Seven of these inventions have already been reported to the European Patent Office (EPO) which are the in process to be recognized as patents, e.g., [TFG$^+$11], [TGPU09b], [TGP$^+$09], [TGPU09a], [TGF$^+$09], [TGF$^+$10], and [TGU$^+$10].

A cost effective network design is the key requirement to keep operators in business. The link bandwidth of backhaul networks is an expensive commodity whose optimal use guarantees the best cost-efficiency of the access network. Link dimensioning is that particular task which determines the appropriate bandwidths for the backhaul (or transport) network with the objective of maximizing the utilization of the allocated transport resources while ensuring the QoS requirements of individual services. The author has contributed to an extensive research on LTE transport network dimensioning carried out with collaboration of a leading industry partner 'Nokia Siemens Networks, Germany'. The details of this work can be found in [LTW$^+$10b], [LTW$^+$10a], [LBD$^+$11], and [LTB$^+$11]. An interesting extension of this work can be accessed in [LLT$^+$12] where the dimensioning is performed for a transport network which is shared by LTE and HSPA networks.

In addition to dimensioning of transport network, the operators are also interested in defining minimum requirements of transport network QoS parameters which can still meet target QoE of the end users. This involves research investigations in quantifying the impact of transport network impairments on the end user QoE. A detailed simulation based study in this area has been performed by the author as published in [TLL$^+$11] and [LTL$^+$11].

1.3 Thesis Overview

The thesis work is organized as follows: Chapter 2 provides an introduction to wireless mobile communication history. After a brief introduction of first generation mobile systems, an overview of the system architecture of the most popular second generation mobile system, Global System for Mobile Communication (GSM), is given. The discussion is extended to the Universal Mobile Telecommunication System (UMTS) which is third generation mobile system. Then a comprehensive discussion is carried out about the LTE mobile communication system which is often referred to as the 3.5 generation mobile system. The discussion encompasses the LTE standardization, motivations and targets, key features, QoS management as well as overall system architecture including both the radio access and core network of LTE. The topic is concluded with a short overview of beyond-LTE technologies, i.e., LTE-Advanced. In addition to this, the most widely used non-3GPP wireless access systems (i.e., IEEE 802.11 networks) are also extensively discussed in this chapter with a special focus on the IEEE 802.11a extension. The chapter is concluded with a description of possible approaches to integrate 3GPP and non-3GPP wireless access networks.

Chapter 3 begins with a discussion which highlights the importance of simulation techniques in the development of communication networks. Then a general introduction is given about the OPNET network simulator, a tool used to build up the simulation platform for the integrated LTE and WLAN networks in this work. Afterwards, a step by step approach is adopted to explain the implementation of important network entities in integrated heterogeneous network simulator using the OPNET tool. Another section of the chapter has been dedicated to discuss various user traffic models which are used within the scope of this work. Finally, the statistical evaluation methods used in simulation based studies of this thesis are explained.

Chapter 4 presents various novel techniques to enhance the LTE radio access network interface. This includes a special LTE air interface scheduler design which can coordinate with the core network in order the circumvent uplink congestion situations. Another LTE air interface scheduling algorithm discussed in this chapter is capable of dynamically changing its behavior in response to network load conditions so that an optimal network operation is realized over time. In addition, the problem of packet queue management for the LTE air interface is also addressed in this chapter. This involves the feasibility discussion of the most popular queue management schemes in the context of the LTE air interface scheduler. The performance evaluation of each of these techniques is performed using the OPNET based LTE network simulator.

Chapter 5 targets the design of a flow management system architecture which can be used by network operators to manage the bandwidth resources of multihomed users in an environment of integrated heterogeneous networks. This discussion encompasses the description of functional entities of the system architecture, inter-entities communication, as well as its incorporation into the 3GPP defined SAE architecture. Moreover, several techniques and mechanisms are also developed in order to fully exploit the potential of user multihoming in integrated heterogeneous wireless access networks. The effectiveness of the proposed mechanisms is evaluated with the help of simulation studies. Their performance is also compared against the default 3GPP proposed behavior of mobile users in integrated LTE and WLAN networks.

Chapter 6 presents analytical solutions for optimized network resource allocation to multihomed users. For this purpose, the optimization technique 'Linear Programming' is used whose introduction is given at the start of the chapter. The proposed solution involves the analytical modeling of user network access links which then leads to the formulation of the resource allocation problem in 'Mixed Integer Linear Programming'. Afterwards, the performance of the proposed analytical solution is assessed using simulation based studies. A computational complexity analysis reveals that the proposed analytical solution is infeasible for real world products because of its processing-intensive nature. This problem is addressed by proposing alternative solutions which are based on heuristic methods and provide near-optimal performance without requiring large computational resources.

Chapter 7 gives the overall conclusion of the work, highlights all the main points and major achievements. Finally, an outlook concerning future work is given.

2 Mobile and Wireless Communication Systems

Mobile and wireless systems and services have seen a remarkable development in the last decades and have become an everyday commodity. Today, various types of wireless communication systems are being deployed which are often distinguished by their coverage and services. For example, an around the globe coverage can be provided using Satellite Communication Systems. A wide-area coverage for pedestrian and vehicular users can be achieved by using the terrestrial cellular and micro-cellular networks often categorized under Wireless Wide-area Networks (WWAN). Wireless Local Area Networks (WLAN) offer high speed access to communication networks supporting user mobility within a limited coverage, e.g., in a campus, office building or in a café. Finally, Wireless Personal Area Networks (WPAN) provide inter-connectivity to the devices centered around an individual person's workspace. Though all of the above mentioned wireless communication systems are of importance, this work focuses on the most deployed two network types (i.e., WWANs & WLANs) and their inter-connectivity.

Mobile communication technologies developed for WWAN are often divided into generations. For example, analog mobile radio systems of the 1980s are the 1st generation (1G), the first digital mobile systems are the 2nd generation (2G), and the first mobile systems handling broadband data are the 3rd generation (3G). The Long Term Evolution (LTE) is often labeled as 3.9G and LTE-Advanced is referred to as the fourth generation (4G). The first and second generation of mobile communication technologies were developed locally in different regions of the world without focusing much on the interoperability. From the second generation, the task of developing mobile technologies has changed from being a regional concern to becoming a global task involving thousands of participants tackled through standards-building organizations such as the Third Generation Partnership Project (3GPP).

As far as wireless local area networks are concerned, many systems based on the proprietary technologies for air interfaces and communication protocols already existed when the first standard was introduced in Europe by ETSI (European Telecommunications Standards Institute) in 1996. This standard, named the 'high performance radio local area network' or HIPERLAN, promised a data rate of 23.5Mbps operating in the 5.2GHz spectrum band. Later revisions of this standard were capable of offering much higher data rates going up to 155Mbps. In parallel to this, the IEEE 802.11 standardization group was established in 1997 which produced the first WLAN standard to provide 1 and 2Mbps aggregate rates. In 1998, IEEE 802.11b working group enhanced the air interface to support data rates up to 11Mbps. During the same year, IEEE 802.11a introduced a new standard based on orthogonal frequency division (OFDM) to provide data rates up to 54Mbps operating at 5GHz. Despite the better performance figures of HIPERLAN no products were available in the market while many companies soon offered simple to implement 802.11 compliant equipment. Due to the lack of available commercial implementation further development of the HIPERLAN standard was stalled and much of the work on HIPERLAN version 2 was included in the physical layer specification of IEEE 802.11a.

As a brief outlook about WWANs and WLANs has been provided, this chapter now further describes the background for the development of the LTE system from WWANs and IEEE 802.11 based WLAN. First, an overview of the technologies and mobile systems leading up to 3G will be given. Next, the system architecture and performance specifications of LTE will be described. Then, IEEE 802.11 standards for WLANs will be discussed. Finally, the chapter will be concluded with a discussion on interworking of WWANs and WLANs.

2.1 First Generation Mobile Systems

The first generation of mobile communication systems to see a large scale commercial growth was introduced in the 1980s. Many countries developed and deployed their individual first generation mobile systems based on Frequency Division Multiple Access (FDMA) and analog Frequency Modulation (FM) technology. For example, the Nippon Telephone and Telegraph (NTT) system was the first operational analog mobile communication system. In 1981, the Nordic Mobile Telephone (NMT) system was introduced in Scandinavia, and in 1983, Advanced Mobile Phone System (AMPS) was started in United States as a trial. Other first generation analog mobile systems include TACS, ETACS, C-450, RTMS, and Radiocom 2000 in Europe and JTACS/NTACS in Japan. These systems were designed only for voice application and were incompatible with one another so that which roaming between countries was not possible.

2.2 GSM

The second generation of mobile systems developed across the world was based on digital communication technologies. These include the Global System for Mobile communications (GSM) standards in Europe, IS-54 & IS-95 standards in USA, and the Personal Digital Cellular (PDC) standard in Japan. With the help of digital technology the second generation mobile systems offered an opportunity to increase the system capacity, to give an improved and consistent quality of service, and to develop light weight and attractive handsets. Among all second generation mobile systems, GSM has been a real success in terms of its widespread deployment and a well-defined system architecture that served as a basis for the development of other systems both in 2G and 3G.

It was due to the deployment of GSM that pan-Europe roaming became a possibility in 1992. Being a digital system, GSM also came with the capabilities to provide data services over the mobile communication networks. Though GSM was originally intended to operate in the 900MHz band, a number of variants have also been developed to operate in other frequency bands to meet the regional deployment requirements outside Europe. Such measures helped GSM become the most widely accepted standard supporting 4.4 billion subscribers in more than 230 countries in 2011 [Str13].

GSM uses the Gaussian Minimum Shift Keying (GMSK) modulation method providing a typical over-the-air bit rate of 270kbps. Moreover, as its access method, GSM employs a combination of Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA). The FDMA part involves the division of available spectrum into frequency carriers of 200kHz. Each of these carrier frequencies is then divided in time, using a TDMA scheme, into eight time slots. One time slot is used by the mobile phone for transmission and one for reception.

There are two basic types of services offered through the GSM system, i.e., telephony or tele-services and data or bearer services. Tele-services are mainly voice services including voice calls, facsimile, short text message (i.e., SMS) etc. Data services enable a GSM phone to receive and send data, e.g., to access the Internet. Although the supported data rate of GSM is just 9.6kbps other enhancements (discussed later in this section) can be used to provide much higher data rates.

2.2.1 System Architecture

Figure 2.1 gives an overview of the hierarchical system architecture of the GSM system. The architecture is composed of three subsystems, the Radio Subsystem

(RSS), the Network and Switching Subsystem (NSS), and the Operation Subsystem (OSS) [Sch03].

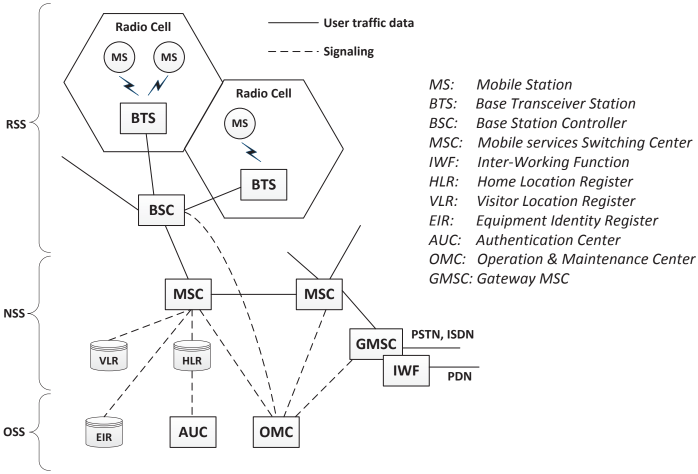

Radio Cell
MS
MS
BTS
Radio Cell
MS
BTS
BSC
MSC
MSC
GMSC
IWF
PSTN, ISDN
PDN
VLR
HLR
EIR
AUC
OMC
RSS
NSS
OSS
User traffic data
Signaling
MS: Mobile Station
BTS: Base Transceiver Station
BSC: Base Station Controller
MSC: Mobile services Switching Center
IWF: Inter-Working Function
HLR: Home Location Register
VLR: Visitor Location Register
EIR: Equipment Identity Register
AUC: Authentication Center
OMC: Operation & Maintenance Center
GMSC: Gateway MSC

Figure 2.1: Functional architecture of the GSM system [Sch03].

a) Radio Subsystem (RSS)

The RSS comprises radio specific entities, e.g., the Mobile Station (MS) and the Base Station Subsystem (BSS). The mobile station consists of hardware and software necessary to access a GSM system as well as a Subscriber Identity Module (SIM). The SIM stores the user data required for authentication and charging mechanisms. The BSS is mainly responsible for maintaining a radio connection to the mobile station. A GSM network may have many BSSs, each consisting of a Base Station Controller (BSC) and a Base Transceiver Station (BTS). A BTS houses the radio transceivers and handles the radio link protocols. It can serve either a single cell or several cells using sectorized antennas. The size of a GSM cell may range from 100m to 35km. One or more BTSs can be managed by a BSC which handles the user handovers from one BTS to another, reserves radio frequencies, and performs paging of the mobile station.

b) Network and Switching Subsystem (NSS)

The NSS contains a variety of different elements and is also termed core network of the GSM system. It is responsible for several key operations and mechanisms, e.g., handovers between different BSCs, worldwide localization of users, charging, accounting, and roaming of users between different network operators etc. The main component in NSS is a Mobile Services Switching Center (MSC) which is a high speed ISDN switch. An MSC often manages several BSCs in a geographical region and is responsible for setting up connections to other MSCs and to the BSCs. In addition, a gateway MSC also provides connection to external networks like PSTN and ISDN.

The Home Location Register (HLR) and Visitor Location Register (VLR) help the MSC provide call routing and roaming capabilities. The administrative information of each subscriber and its last known location can found in the HLR which helps route calls to the relevant base station. There is logically one HLR per GSM network. The VLR contains the temporary information from the HLR required to provide service to a subscriber currently located in a geographical area controlled by the associated MSC of the VLR.

c) Operation Subsystem (OSS)

The OSS is connected to all entities of the NSS as well as to the BSCs. It is used to monitor the overall system, control the traffic load of the BSS, and perform the maintenance activities. The OMC (Operation and Maintenance Center) entity of OSS accesses other network entities via SS7 signaling and typically performs the traffic monitoring, obtains the status reports from network entities as well as participates in subscriber and security management tasks. A unit of OSS named authentication center (AUC) takes care of subscriber authentication and ciphering of call data on the radio channels. The OSS also prevents calls from stolen, unauthorized, or defective mobile stations using the information contained in the Equipment Identity Register (EIR).

2.2.2 Data Service Enhancements

At the time when GSM was developed, the standard data rate of 9.6kbps available for data services used to be considered adequate. However, with the rapid growth of the Internet services like web browsing, email exchange and file download etc., this data rate became insufficient to meet these application demands. To improve the data transmission capabilities of GSM, two enhancements were developed. The first enhancement called High Speed Circuit Switched Data (HSCSD) combines several traffic channels (each providing 9.6kbps) to increase the overall user data rate. HSCSD is capable of providing up to 57.6kbps data rate. The second enhancement termed General Packet Radio Service (GPRS) is fully packet-oriented which provides more powerful and flexible data transmission. Though GPRS can offer data rates as high as 171.2kbps, typical data rates are 53.6kbps in downlink and 26.8kbps in uplink. GPRS was then further evolved to the EDGE technology. The name EDGE stands for Enhanced Data for GSM Evolution and it supports data transmission speeds up to 384kbps. Often referred as 2.5G system, EDGE uses the 8 PSK modulation scheme to offer a significantly higher data rate than GPRS.

2.3 UMTS

The standardization activities for the 3rd generation of mobile systems started in ETSI in 1996. The Wideband Code Division Multiple Access (WCDMA) proposals from Europe and Japan were merged in early 1998 and came out as the 3G standard for the European market termed Universal Mobile Telecommunication Service (UMTS). A few months later standards-developing organizations from all regions of the world founded the Third Generation Partnership Project (3GPP) in order to solve the problem of maintaining parallel development of aligned specifications in multiple regions. Soon after, 3GPP introduced the initial release of UMTS standards in 1996 which is referred to as release 99 or Rel-3. After Rel-3, the work on Rel-4 and Rel-5 was started by 3GPP in the year 2000. Rel-4 was concluded in March 2001 with the introduced features like, QoS in the fixed network including several execution environments (e.g., MExE, mobile execution environment) and new service architectures. Rel-5 specified a new core network to support IP-based multimedia services (IMS) as well as a high speed downlink packet access (HSDPA) service. Rel-6 focused on Multiple Input Multiple Output (MIMO) antennas, enhanced Multimedia Service (MMS), interworking with wireless LAN (WLAN), High Speed Uplink Packet Access (HSUPA), and many other management features. Rel-7 which was released in 2007, introduced High Speed Packet Access Evolution (HSPA+) service, improvements to QoS for realtime applications, and reduced the packet latencies.

UMTS uses Code Division Multiple Access (CDMA) as the multiple access technology which offers numerous advantages over the schemes used in 2G systems that were predominantly TDMA based schemes. The most prominent feature of the CDMA scheme is the improved spectral efficiency due to the use of Quadrature Phase Shift Keying (QPSK) as a modulation scheme. Theoretically, it increases the spectral efficiency three to four times higher than that of the GSM system. In addition, CDMA allows to use the same channel frequency in adjacent cells and an improved handover reliability by supporting a so called “soft handover” mechanism. Furthermore, the use of spread spectrum and multiple spreading codes for CDMA makes the transmission resistant to signal jamming, significantly reduces the chances of eavesdropping and allows flexible allocation of resources.

In contrast to the GSM system, UMTS networks were not designed just for voice but for a flexible delivery of any type of service where each new service does not require particular network optimization. With such provisions UMTS networks were capable of providing high data rates up to 384kbps in Rel-3 and beyond 2Mbps in Rel-5. The packet round trip time was reduced below 200ms, a seamless mobility was managed for data applications, quality of service support was improved, simultaneous transmission of voice and data was made possible, and interworking with existing GSM/GPRS networks was made feasible. Person-to-person services of UMTS include voice telephony with wideband codec to improve speech fidelity, video telephony using new multimedia architecture, an enhanced SMS service termed MMS (Multimedia Messaging Service) which is capable of delivering messages with embedded multimedia contents, Push-to-talk over Cellular (PoC) service which is similar in nature to walkie-talkie, and Voice over IP (VoIP) support which can also be complemented with streaming video, images, content sharing, gaming etc. Content-to-person services of UMTS are web browsing, content download (e.g., ringing tone, video clips, MP3 music etc.), Multimedia Broadcast Multicast Service (MBMS) as well as other multimedia streaming services like web broadcasting, video streaming on demand etc.[HT04]

2.3.1 System Architecture

In order to meet the design targets, the UMTS network architecture was required to provide significantly higher performance than that of the original GSM network. However, owing to the fact that many networks had migrated through the use of GPRS and EDGE, they already had the ability to carry data. Therefore, many of the elements required for the UMTS network architecture were seen as a migration. This substantially reduced the cost of implementing the UMTS network as many elements were already in place or needed upgrading.

As depicted in Figure 2.2, the UMTS network comprises three interacting domains: Core Network (CN), UMTS Terrestrial Radio Access Network (UTRAN) and User Equipment (UE).

a) Core Network (CN)

The main function of Core Network (CN) is to provide switching and routing for the user traffic as well as some network management functions. The basic CN architecture for UMTS is based on the GSM network with GPRS and can be further divided in circuit switched and packet switched domains. MSC, VLR,

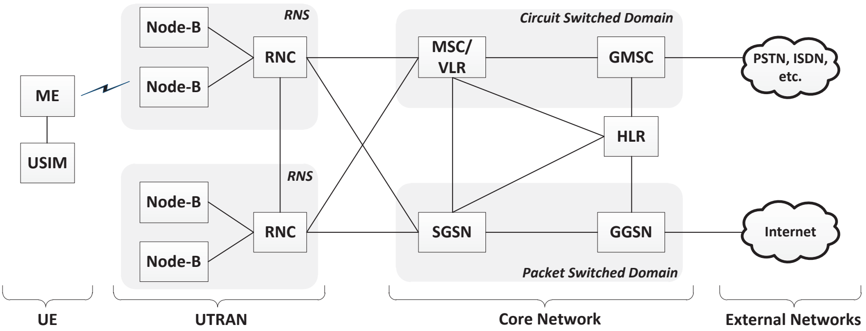


Figure 2.2: UMTS system architecture [HT04].

and gateway MSC are among elements of the circuit switched domain. The packet switched domain includes elements like Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). Moreover, some network elements, like EIR, HLR, and AUC are shared by both domains. SGSN functionality is similar to that of MSC but is typically used for Packet Switched (PS) services. In the same way, GGSN functionality is close to that of the gateway MSC but is related to PS services.

b) UMTS Terrestrial Radio Access Network (UTRAN)

UTRAN consists of one or more Radio Network Subsystems (RNS). An RNS is a sub-network within UTRAN comprising one Radio Network Controller (RNC) and one or more base stations referred as Node-Bs. The Node-B contains the transmitter and receiver to communicate with the UEs within the cell. The RNC controlling one Node-B is responsible for the load and congestion control of its own cells, and also executes the admission control and code allocation for new radio links to be established in those cells.

c) User Equipment (UE)

The UE works as an air interface counter part of the Node-B and has two components: Mobile Equipment (ME) and UMTS Subscriber Identity Module (USIM). The ME is the radio terminal used for radio communication over the air interface. The USIM is a smartcard that holds the subscriber identity, stores authentication and encryption keys, and performs the user authentication.

2.3.2 Data Service Enhancements

Similar to GSM, system enhancements also followed UMTS to achieve higher data rates and improved system capacity. This includes HSDPA and HSUPA standards from 3GPP specifications of Rel-5 and Rel-6, respectively. The use of both of these enhancements is often referred to as HSPA. HSPA increases the spectral efficiency by using a higher order modulation scheme (16QAM) to achieve up to 14Mbps data rate in downlink and 5.8Mbps in uplink. Using a shorter Transmission Time Interval (TTI) (i.e., 2ms instead of 10ms in Rel-3), HSPA reduces the packet round trip time and improves link adaptation to fast channel variations. In addition, moving the packet scheduling function from RNC to Node-B along with the adaptive coding and modulation enables the system to quickly respond to the varying radio channel and interference conditions. Furthermore, Node-B based Hybrid ARQ (HARQ) provides reduced retransmission round trip time and adds robustness to the system by allowing soft combining of retransmissions.

An enhanced version of HSPA termed HSPA+ or Evolved HSPA was defined in Rel-7 and Rel-8 of the 3GPP standards. Using HSPA+ the data transfer rates were increased further to provide download speeds comparable with fixed broadband lines. Some of the major HSPA+ features include up to 42Mbps data rate, Multiple Input Multiple Output (MIMO) transmission, higher order modulation scheme (64QAM), enhancements to layer 2 protocols, and faster call set-up time etc.

2.4 LTE

For many years, voice calls dominated the traffic in mobile communication systems. Though the growth of mobile data was initially slow, its use has been increasing dramatically for the last few years. This is mainly due to widespread use of smartphones which are more attractive and user friendly than their predecessors and facilitate the creation of applications by third party developers. The result was an explosion in number and use of mobile applications accompanied with flat rate charging schemes that led to a situation where neither developers nor users were motivated to limit their data consumption. Owing to their limited capacity, 2G and 3G networks soon started to become congested. This made network operators and developers realize the demand of increase in system capacity.

Earlier generations of mobile communication systems were built only for circuit switched services. The first data services over GSM were provided by packet based GPRS in a later addition. The demand of data services also influenced the development of 3G which was based on circuit switched data with packet switched services as an add-on. Provided that voice calls can be transported over packet switched networks (i.e., through VoIP), operators can move everything to the packet switched domain and hence reduce their capital and operational expenditure. This encouraged the concept for an all IP network architecture for mobile communication systems.

In 3G networks, the packet delays between network elements and across the air interface are of the order of 100ms. This a is big hurdle in providing good quality of experience to the users of VoIP and other real time interactive services. Thus another driver was to reduce the latency in the network.

Maintaining the backward compatibility and incorporating the above described features in the existing overly complex specifications for UMTS would have been a cumbersome task. Therefore, a fresh start was required by the designers in order to improve the system performance without the need to support legacy devices.

2.4.1 LTE Standardization

The 3G evolution continued in 2004, when 3GPP organized a workshop to initiate work on the Long-Term Evolution (LTE) radio interface. The aim was to make LTE competitive over timescales of at least 10 years. Afterwards this task was handled as a study item in a technical specification group of 3GPP for almost six months. The result was a technical report approved in June 2005 which defined the requirements or design targets for LTE. The main requirements included higher data rate, enhanced cell edge coverage, lower latencies, improved system capacity, and spectrum flexibility. An extensive study of different physical layer technologies by a 3GPP working group suggested OFDM as the LTE radio access technology. In December 2007, 3GPP approved first LTE specifications in its Rel-8 standards. Work has since then continued on LTE with new features added in each release. Figure 2.3 shows the way in which the new architecture has been developed from that of UMTS.

In the new architecture, the Evolved Packet Core (EPC) replaces the packet switched domain of UMTS while there is no equivalent to the circuit switched domain. This is because voice calls are supposed to be transported over EPC using Voice over IP (VoIP). UTRAN in the UMTS network has been replaced by the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) which handles the EPC's radio communications with the user equipment. Actually, the new architecture was developed as part of two 3GPP work items, (i) System Architecture Evolution (SAE), which covers the core network, and (ii) Long-Term Evolution (LTE) which covers the radio access network, air interface, and user equipment. Officially, the whole system is termed Evolved Packet System (EPS). However,

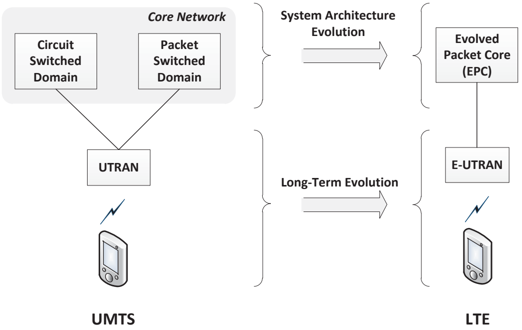


Figure 2.3: Evolution of system architecture from UMTS to LTE [Cox12].

LTE has become the colloquial name for the whole system and is also being regularly used in this way by 3GPP.

2.4.2 LTE Key Features

In the following a number of the LTE key features are discussed.

2.4.2.1 Enhanced Air Interface

LTE is built on an all-new radio access network based on OFDM (Orthogonal Frequency-Division Multiplexing) technology with higher order modulation schemes such as 64QAM. It also allows the use of MIMO and Beam Forming supporting up to four antennas per station as the complementary radio techniques. In addition, LTE exploits highly sophisticated Forward Error Correction (FEC) schemes like tail biting, convolution coding, and turbo coding etc. With all these enhancements, LTE manages to provide up to five times higher throughput than that offered by HSPA networks. This accounts for downlink and uplink peak data rates of 100Mbps and 50Mbps, respectively, when operating in 20MHz spectrum allocation. Moreover, LTE can support at least 200 mobile terminals in the active state when operating in 5MHz spectrum allocation.

2.4.2.2 Spectral Efficiency

LTE substantially improves the spectral efficiency and cell edge coverage while maintaining the same site locations. For example, spectral efficiency in the downlink is targeted at 5 bps/Hz/cell and 2.5 bps/Hz/cell in the uplink. This implies a three to four times improvement over the HSPA technology. A better spectral efficiency allows network operators to support more customers with the reduced cost of delivery per bit.

2.4.2.3 Latency

LTE significantly reduces the transition times from idle and dormant states to the active state. This means a transition time of less than 100ms from a camped state to the active state and less than 50ms from dormant to the active state. Radio access network latency is reduced to below 5ms under the unloaded condition for small IP packets. This is four to five times less than the delays experienced in HSPA networks. These enhancements help LTE deliver a more responsive user experience for interactive, real time service such as high quality audio/video telephony and multi-player gaming etc.

2.4.2.4 Mobility

The E-UTRAN of LTE networks provides optimum performance for mobile speed 0–15km/h, whereas a slight degradation is permitted for higher speeds. For a speed between 15 to 120km/h, LTE provides high performance and for speeds above 120km/h, the system is capable of maintaining the connection across the cellular network. The maximum speed supported by LTE is 350km/h.

2.4.2.5 An All-IP Environment

LTE supports a ‘flat’ all-IP based core network with much simplified architecture and open interfaces. This enables an improved interworking with other fixed and non-3GPP wireless communication networks. A complete packet oriented network also enables more flexible service provisioning.

2.4.2.6 Flexible Radio Planning

LTE can deliver optimum performance in a cell size of up to 5km radius. It is still capable of delivering effective performance for cells of size up to 30km radius. However, a limited performance should be expected for a cell with radius up to 100km. LTE can be deployed with scalable spectrum allocations, e.g., 1.25, 1.6, 2.5, 5, 10, 15, and 20MHz. It can operate in all 3GPP specified frequency bands in paired and unpaired spectrum allocations. In this way, when deployed at higher frequencies, LTE is attractive for strategies focused on network capacity. On the other hand, when operating at lower frequencies LTE provides ubiquitous cost-effective coverage.

2.4.3 LTE Radio Access

The most important technologies used by LTE radio access include transmission schemes, scheduling, and multi-antenna support as discussed in the following.

2.4.3.1 Transmission Schemes

For the LTE downlink, the Orthogonal Frequency Division Multiplexing (OFDM) transmission scheme is used, while the uplink employs a single-carrier transmission based on DFT-Spread OFDM (DFTS-OFDM). The selection of the OFDM scheme for LTE downlink transmission is due to its inherent high degree of robustness to frequency selective fading especially when used in conjunction with spatial multiplexing. OFDM enables LTE to perform channel aware scheduling with an additional degree of freedom by providing access to the frequency domain. In addition, OFDM makes flexible resource allocation a possibility by varying the number of OFDM sub-carriers used for transmission. Moreover, inherent properties of OFDM makes brodcast/multicast transmissions a simple task. Similarly, the choice of DFTS-OFDMA as the LTE uplink transmission scheme is mainly because of its lower power requirements for transmission and straightforward channel equalization.

OFDM Transmission Scheme

OFDM is a broadband multi-carrier modulation method where the total bandwidth is split into a large number of smaller and narrower bandwidth units termed sub-carriers. OFDM offers superior performance and benefits over traditional single-carrier modulation methods. This is because its sub-channels are of narrow bandwidths and therefore not vulnerable to frequency selective fading. This property helps simplify equalization techniques.

The term Orthogonal Frequency Division Multiplexing is because of the fact that two modulated OFDM sub-carriers x_{k_1} and x_{k_2} are mutually *orthogonal* over the time interval $mT \leq t < (m+1)T$, i.e.,

$$\int_{mT}^{(m+1)T} x_{k_1}(t) x_{k_2}^{*}(t) \mathrm{d}t = 0 \quad \text{for } k_1 \neq k_2. \tag{2.1}$$

where $(\cdot)^*$ denotes the complex conjugate operator. The T is the per sub-carrier modulation-symbol time and m is an OFDM symbol number [DPS11]. In this way, basic OFDM transmission can be seen as the modulation of a set of orthogonal functions. Due to this property guard intervals between the sub-carriers are not required which help increase the spectral efficiency of the system.

In case of OFDM transmissions, the 'physical resource' can be illustrated as a time-frequency grid where each column corresponds to one OFDM symbol and each row corresponds to one OFDM sub-carrier. This time-frequency grid is shown in Figure 2.4.

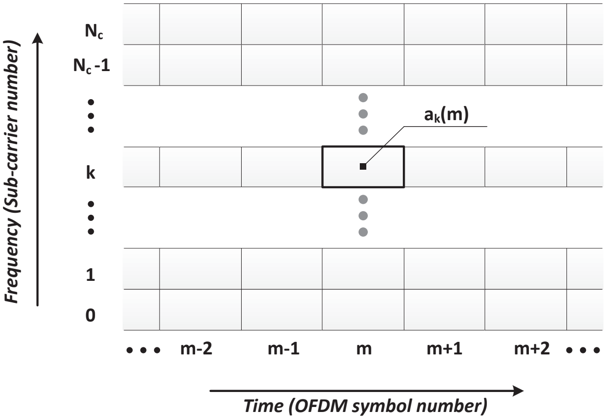

Frequency (Sub-carrier number)
N_c
N_c-1
k
1
0
a_k(m)
• • • m-2 m-1 m m+1 m+2 • • •
Time (OFDM symbol number)

Figure 2.4: OFDM time-frequency grid where N_c represents the number of sub-carriers, $a_k(m)$ is an OFDM symbol, and m is the symbol index.[DPS11].

OFDM can also be used as a multiple-access scheme, allowing simultaneous frequency-separated transmissions to/from multiple mobile terminals. This implies that in each OFDM symbol interval, different subsets of available sub-carriers are used for transmission to/from different mobile terminals. This scheme is often referred to as Orthogonal Frequency Division Multiple Access (OFDMA).

Single Carrier OFDMA Transmission Scheme

LTE uplink transmission is based on Single-Carrier FDMA (SC-OFDMA) which is a modified form of OFDMA. It inherits all benefits of OFDMA with the additional advantage of low peak-to-average power ratio which makes it suitable for uplink transmission by mobile terminals. This is because low peak-to-average power ratio is a property desired to employ efficient power amplifiers in order to save battery power of the mobile terminal.

SC-OFDMA is often viewed as a DFT-coded OFDM where time domain data symbols are transformed to frequency domain by a Discrete Fourier Transform

(DFT) before feeding them to the standard OFDM modulator. In a standard scheme of OFDMA, each data symbol is carried on a separate sub-carrier. However, in SC-OFDMA, multiple sub-carriers carry each data symbol due to mapping of the symbol's frequency domain samples to sub-carriers. Owing to the fact that each data symbol is spread over multiple sub-carriers, SC-OFDMA offers frequency diversity gain or spreading gain in a frequency selective channel. That is why SC-OFDMA is also called frequency spread OFDM or DFT-spread OFDM.

2.4.3.2 Channel Aware Scheduling

The LTE transmission scheme dynamically shares the overall time-frequency resources among the users. It is often termed shared-channel transmission where the scheduler controls, for each time slot, to which users the different parts of the shared resource should be allocated. The scheduler also performs the rate adaptation to determine the data rate to be used in each transmission. Thus, in determining the overall system performance the scheduler plays a key role. In order to improve system capacity, the scheduler may also consider the channel conditions in the scheduling decisions which is called channel aware scheduling. For example, due to the use of OFDM, the scheduler has access to both the time and frequency domains and therefore for each time instant and frequency region, it can select the user with the best channel conditions, as shown in Figure 2.5. In principle, a scheduled user can be allocated with an arbitrary combination of resource blocks in 1ms scheduling intervals.

2.4.3.3 Fast Hybrid ARQ With Soft Combining

LTE uses fast hybrid ARQ with soft combining to allow mobile terminals request retransmission of erroneously received data and to provide a way to control rate adaptation implicity. Retransmissions are requested rapidly for each erroneously received packet in order to minimize its impact on end user application performance. Moreover, incremental redundancy is used as the soft combining strategy where incorrectly received data blocks are buffered at the receiver instead of being discarded, and when the retransmitted block is received, the two blocks are combined. In practice, multiple sets of coded bits are generated for the same set of information bits. Each retransmission uses a different set of coded bits with different redundancy versions generated by puncturing the decoder output. In this way, at every retransmission the receiver gains extra information to perform the decoding correctly.

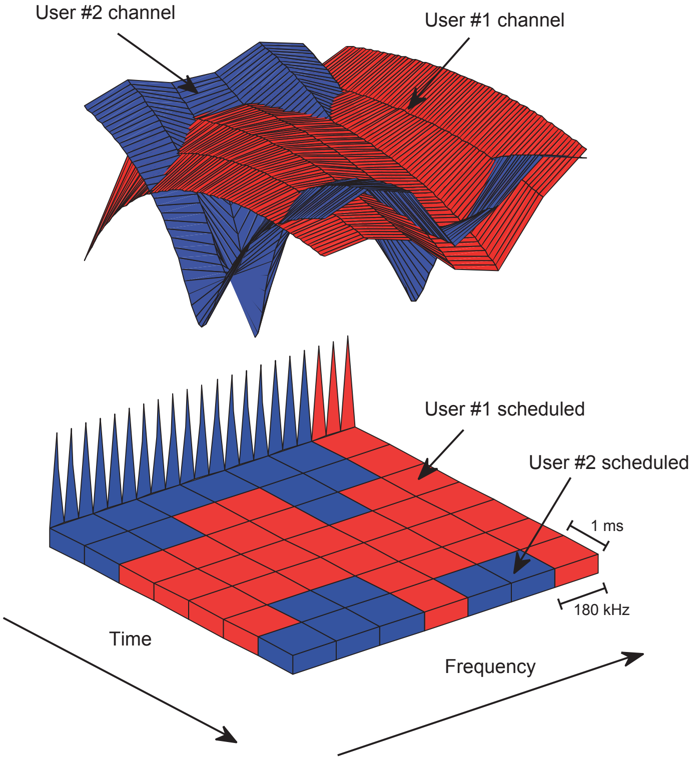


Figure 2.5: Downlink channel aware scheduling in the time and frequency domains [DPS11]. The upper part of figure represents the user channel conditions in terms of SINR measurements. The lower part shows the radio spectrum allocations along the time.

2.4.3.4 Multi-Antenna Transmission

The use of multi-antenna transmission techniques is the key feature of LTE in order to achieve aggressive performance targets. LTE supports multiple antennas both for uplink and downlink transmissions. Multiple transmit antennas at the base station are employed for receive diversity and beam-forming to improve the received SINR. Similarly, multiple receive antennas can be used to attain additional gains in interference-limited situations if the antennas are used not only for the diversity but also to suppress interference. In addition, multiple antennas at the transmitter and receiver are used for ‘spatial multiplexing’ which helps create multiple parallel channels in order to substantially improve data rates. Spatial multiplexing is also termed multi-user MIMO.

2.4.4 Overall System Architecture

As mentioned earlier, in the evolution from third generation the overall system architecture of both Core Network (CN) and Radio Access Network (RAN) was revised, including a split of functionality between the two network parts. This functional split allows different radio-access technologies to be served by the same core network. The RAN is responsible for radio-related network functionalities including scheduling, coding, radio transmission, and radio resource handling. The core network or Evolved Packet Core (EPC) takes care of setting up end-to-end connections, authentication, mobility management, billing and also other complementary functions to provide a complete mobile broadband network. Figure 2.6 depicts some of different node types from the overall system architecture.

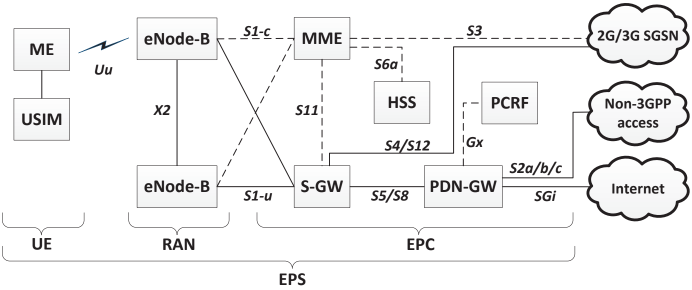


Figure 2.6: LTE overall system architecture.

2.4.4.1 Core Network

The main logical nodes of the core network (or EPC) are listed and described below:

- **PCRF**: The Policy Control and Charging Rules Function is responsible for detecting service flows and enforcing charging policy. The PCRF also provides the QoS authorization that specifies how a certain data flow will be handled in accordance with the user's subscription.

- **MME**: The Mobility Management Entity is a control-plane node of EPC which is responsible for establishment, maintenance, and release of bearers as well as handling of security keys.
- **HSS**: The Home Subscriber Service keeps a database of subscriber information, e.g., QoS profile, any access restrictions for roaming, identity of the MME to which the user is attached or registered etc. The HSS may also integrate the authentication center (AUC).
- **PDN-GW**: The Packet Data Network Gateway allocates the IP address to the UEs, as well as, performs QoS enforcement and per flow-based charging in accordance to rules from PCRF. It also acts as mobility anchor for interworking with non-3GPP access technologies like WiMAX, WLAN etc.
- **S-GW**: The Serving Gateway is a user-plane node which connects the EPC to the RAN. The S-GW serves as a local mobility anchor for data bearers when the UE moves between eNode-Bs as well as a mobility anchor for other 3GPP access technologies such as GSM/GPRS, HSPA etc.

2.4.4.2 Radio Access Network

The access network of LTE called E-UTRAN, is a simple network of eNode-Bs. Owing to the fact that there is no centralized controller in E-UTRAN, this architecture is said to be flat. An eNode-B is connected to the EPC by means of a S1 interface, more specifically to the MME by means of the S1 control-plane part (S1-c), and to the S-GW by means of the S1 user-plane part (S1-u). It is allowed for one eNode-B to be connected to multiple MMEs/S-GWs for the purpose of load sharing and redundancy. Furthermore, the eNode-Bs are normally interconnected with each other by means of the X2 interface which is mainly used to support intra-LTE handovers.

In contrast to second and third generation mobile communication systems, the radio controller function is integrated into eNode-B itself. This accounts for the tighter integration between different protocol layers of RAN resulting in reduced latency and improved efficiency. The advantage of having distributed control is that the need for a highly reliable, processing intensive central unit is eliminated which in turn avoids ‘single point of failure’. However, the disadvantage lies in the fact that during the UE handover, all UE information together with any buffered data must be transported using the X2 interface between the involved eNode-Bs.

Considering the fact that the eNode-B is a logical node, a typical implementation of the eNode-B is a three sector site, where one base station handles transmission in three cells. Another common implementation involves one baseband processing unit to which a number of remote radio heads are connected.

2.4.5 Protocol Architecture

Figure 2.7 shows the user-plane protocol architecture of the E-UTRAN together with two nodes from core network. In downlink, the user data flow in the form of IP packets has to pass through a number of protocol layers as described below:

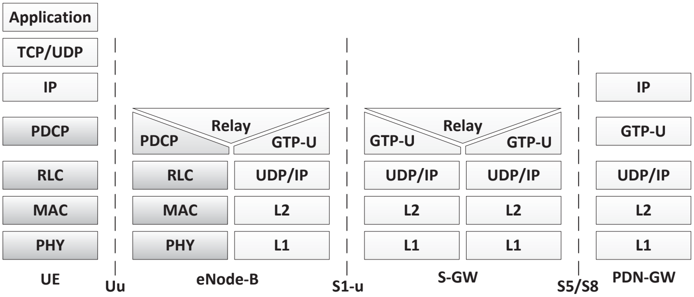


Figure 2.7: LTE user-plane protocol stack.

- **PDCP**: The Packet Data Convergence Protocol mainly performs the IP header compression using the ROCH (Robust Header Compression) standard in order to reduce the overhead of protocol header bits. For each radio bearer there must be a PDCP entity which is also responsible for other functions such as compression / decompression and ciphering / deciphering of the data flow [3GP11c].

- **RLC**: The Radio Link Control provides services to the PDCP in the form of radio bearers. Similar to PDCP there exists one RLC entity per radio bearer which takes care of in-sequence packet deliveries to the upper layer, necessary retransmissions as well as segmentation/reassembly [3GP10].

- **MAC**: The Medium Access Control layer offers its services to the RLC in the form of logical channels. The MAC is primarily responsible for scheduling of radio resources in uplink and downlink. In addition, it also handles fast Automatic Repeat Request (i.e., HARQ) retransmissions. For each cell the eNode-B maintains one MAC entity [3GP11b].

- **PHY**: The Physical layer provides its services to the MAC in the form of transport channels. It handles several typical physical layer functions which include coding/decoding, modulation/demodulation etc.

A summary of functions can be seen in Figure 2.8 for radio interface protocols of LTE in downlink communication.

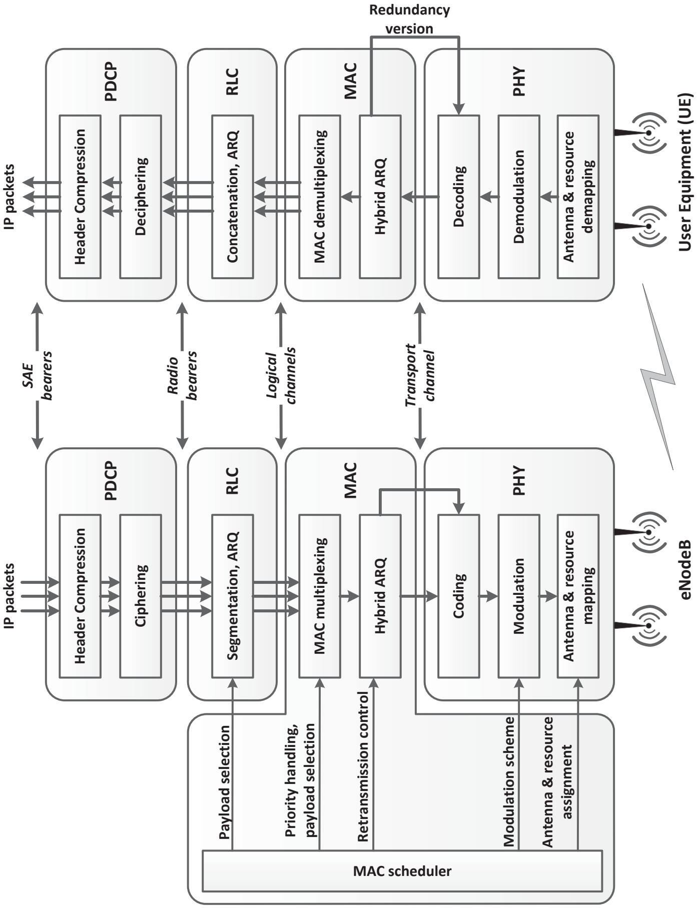


Figure 2.8: Detailed LTE downlink protocol architecture [DPS11].

2.4.5.1 Medium Access Control (MAC)

The MAC layer should be seen as the lowest sub-layer in the Layer 2 of the E-UTRAN protocol architecture. It accesses the services of physical layer through transport channels and connects to RLC layer above through logical channels. In this way, the MAC has to perform multiplexing/demultiplexing between logical and transport channels. However, the most important functionality of the MAC layer is scheduling of air interface resources in both uplink and downlink. A detailed discussion about the MAC scheduling procedures will be carried out later in this section. In addition, the MAC is also responsible for the transmit and receive HARQ operations, QoS based prioritization of logical channels, medium access control as well as various other control functions.

Logical Channels and Transport Channels

A logical channel is defined by the type of information it carries and is classified as Control Logical Channel, which is used to transmit control data, or as Traffic Logical Channel which carries user-plane data. The data from logical channels are multiplexed into transport channels depending on how it should be transmitted over the air. In other words, a transport channel is defined by how and with what characteristics the information should be transmitted over the radio interface. The data on a transport channel is organized into Transport Blocks of dynamic size. In each Transmission Time Interval (TTI) up to two transport blocks are transmitted over the radio interface if spatial multiplexing is employed or at most one transport block in the absence of MIMO. A list of logical and transport channels is presented in Table 2.1, however a detailed description can be found in [DPS11].

Table 2.1: A list of logical channels and transport channels [3GP11b].

Logical channels	**Transport channels**
Broadcast Control Channel (BCCH)	Broadcast Channel (BCH)
Paging Control Channel (PCCH)	Paging Channel (PCH)
Dedicated Control Channel (DCCH)	Downlink Shared Channel (DL-SCH)
Multicast Control Channel (MCCH)	Multicast Channel (MCH)
Common Control Channel (CCCH)	Uplink Shared Channel (UL-SCH)
Multicast Traffic Channel (MTCH)	Random Access Channel (RACH)
Dedicated Traffic Channel (DTCH)	

Hybrid ARQ

As introduced earlier, LTE employs hybrid ARQ with soft combining to achieve robustness against the transmission errors. Hybrid ARQ is part of the MAC layer, while the soft combining is performed at the physical layer. Hybrid ARQ is not applicable for broadcast transmissions and therefore only supported for uplink and downlink shared channels, i.e., DL-SCH, UL-SCH.

The LTE hybrid ARQ protocol is based on multiple ‘stop and wait’ processes. In order to minimize the overhead, a single bit is used to report ACK/NAK. Therefore, timing of the ACK/NAK bit is used to determine the associated hybrid ARQ process at the transmitter and receiver. The use of multiple parallel hybrid ARQ processes may give rise to out-of-sequence data delivery as shown in Figure 2.9. For example, transport block 5 was successfully decoded before transport block 3, which required retransmission. Hence, the MAC layer must take care of proper reordering of data using the sequence numbers before performing de-multiplexing into the logical channels.

The hybrid ARQ mechanism may occasionally fail to deliver error free data blocks to the RLC due to erroneous feedback signalling, for example, a NAK is incorrectly interpreted as an ACK. Though the probability of having such incidents are of the order of 1% [Dah07], it is too high for TCP based services which virtually require error-free delivery of TCP packets. In order to avoid retransmissions at higher layers which cause excessive delays and performance degradation, another data integrity check is performed at the RLC layer. The necessary retransmissions may also be carried out after this integrity check when operating in ‘RLC Acknowledged Mode (RLC-AM)’.

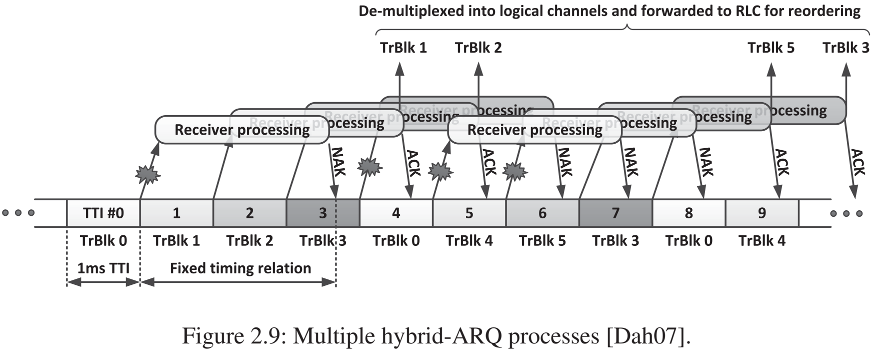


Figure 2.9: Multiple hybrid-ARQ processes [Dah07].

Scheduling

The scheduling function of the MAC layer controls the assignment of uplink and downlink time-frequency resources dynamically. The basic time-frequency resource unit in the scheduler is called ‘resource block’ which spans 180kHz in the frequency domain and 1ms in the time domain. In each, 1ms scheduling interval, the scheduler allocates resource blocks to one or more terminals together with specifying the transport block size, the modulation and coding scheme as well as the antenna mapping for MIMO transmissions. In LTE, uplink and downlink scheduling decisions can be taken independent of each other.

Owing to the fact that 3GPP has not specified any scheduling strategy, numerous scheduler designs have been proposed by the research community. The goals of most schedulers is to improve the spectral efficiency by exploiting the channel variations between the mobile terminals and schedule the transmissions accordingly. The use of OFDM allows LTE downlink schedulers to take advantage of channel variations in both time and frequency domain (see Figure 2.5). The possibility of exploiting frequency domain channel variations, in addition to time domain variations, bears a significant importance in order to support the larger bandwidths of LTE where frequency selective fading turns out to be a major problem. The downlink scheduler relies on channel-quality reports from mobile terminals to incorporate channel conditions in the scheduling decisions. The channel-quality report, termed Channel Quality Indicator (CQI), has its basis in the measurement on the downlink reference signals and conveys not only the instantaneous channel quality in frequency domain but also the information regarding spatial multiplexing mechanism. Moreover, as LTE supports QoS aware scheduling, it implies that a high performance scheduler should also consider the buffer status and priorities of individual bearers in the scheduling decisions.

In contrast to the LTE downlink, where the scheduling decisions are taken per radio bearer basis, the uplink scheduling decision is taken per mobile terminal basis. For every TTI, the uplink scheduler at eNode-B assigns the time-frequency resources to the mobile terminal and also determines the transport format (e.g., transport block size, modulation and coding scheme etc.) which shall be used by the mobile terminal. As the scheduler already has the knowledge about the transport format of mobile terminal’s transmission, it eliminates the need for outband control signalling from the mobile terminal to the eNode-B. This is beneficial from the coverage viewpoint as the transmission of outband control information with higher reliability requires significantly more resources compared to the transmission of user-plane data.

In principle, the uplink scheduler can also take the channel conditions of mobile terminals into account. However, estimating the uplink channel quality is not as simple as for the downlink. This is because, in downlink all mobile terminals share the same reference signal transmitted by the eNode-B for channel quality estimation purpose. In uplink, this reference signal must be transmitted by all mobile terminals to allow the eNode-B, to estimate the channel quality. Though the transmission of such a reference signal is supported by LTE, it comes at the cost of overhead. An overview of downlink and uplink scheduling mechanisms can be seen in Figure 2.10

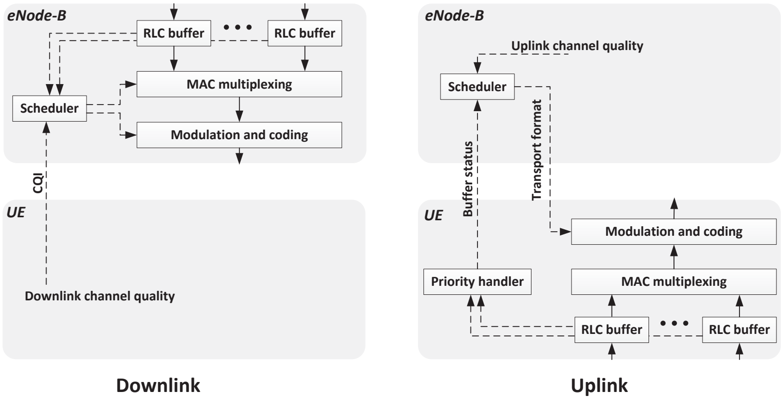


Figure 2.10: An overview of the downlink and uplink scheduling [Dah07].

2.4.6 LTE Mobility

LTE supports mobility not only within LTE but also to other networks of 3GPP and non-3GPP technologies. As far as intra-LTE mobility is concerned, there are two types of handover procedures for UEs in active mode, i.e., handover using the X2 link and the handovers using the S1 link. Typically, the X2-based handover procedure is preferred for inter-eNode-B handover, however, if there is no X2 link between the two eNode-Bs then an S1-based handover is triggered.

2.4.6.1 X2-based Handovers

The X2-based handover procedure has been shown in Figure 2.11. In this case, based on measurement reports from the UE, the source eNode-B determines the target eNode-B and also queries the target eNode-B if it has sufficient resources to accommodate the UE. After successful completion of this phase, the target eNode-B sets aside the radio resources before the UE is commanded to start the actual handover procedure. In addition to this negotiation, the two eNode-Bs have to make necessary arrangements to avoid data loss during the handover. This is because, in LTE, data buffering in downlink occurs at the PDCP and RLC layers of the eNode-B's E-UTRAN protocol stack. Therefore, once the handover decision is taken, the source eNode-B must forward the buffered data to the target eNode-B through a mechanism called 'Buffer Forwarding'. It is up to the source eNode-B to decide which data of which traffic type would be forwarded, e.g., it may forward the data belonging to non-realtime traffic (lossless handover) and no forwarding for realtime traffic data (seamless handover).

If the source eNode-B selects the seamless handover mode for a bearer, it requests the target eNode-B to establish a GTP tunnel over the X2 interface in order to perform downlink data forwarding. On receiving the acknowledgement message, it starts forwarding the freshly arriving data from the S-GW toward the target eNode-B in parallel to sending the handover trigger to the UE over the radio interface. The forwarded data is then delivered to the UE by the target eNode-B as soon as the radio bearer is established between the UE and the target eNode-B. In order to support in-sequence delivery of packets to the UE, the target eNode-B first delivers the packets received over the X2 interface and then delivers the packets received over its S1 interface. The end of the forwarding of data over the X2 interface is signaled to the target eNode-B using special packets termed as 'End Markers'.

In case, the source eNode-B selects the lossless handover mode for a bearer, it has to additionally forward the buffered data over the X2 interface before forwarding the freshly arriving data from the S-GW. The buffered data includes the PDCP packets that are buffered locally because they have not yet been delivered to the UE. These packets are forwarded along with their sequence number assigned by the PDCP layer. In this way, PDCP sequence numbers are continued at the target eNode-B, which helps the UE to reorder packets to ensure in-sequence delivery of packets to the higher layers. The rest of the procedure is the same as described for the seamless handover mode.

2.4.6.2 S1-based Handovers

In some situations the X2-based handover is not possible, e.g., because there is no X2 connectivity to the target eNode-B or an error is indicated by the target eNode-B after an unsuccessful attempt of X2-based handover, or it is dynamically learned by the source eNode-B using the 'status transfer' procedure. In these situations, the source eNode-B initiates the handover process by sending control messages over the S1-MME reference point. The handover procedure in this case is very similar to that of X2-based handover, except the involvement of the MME in relaying the handover signaling between the source eNode-B and target eNode-B. Furthermore, in S1-based handovers, the target eNode-B needs not to inform the MME to switch

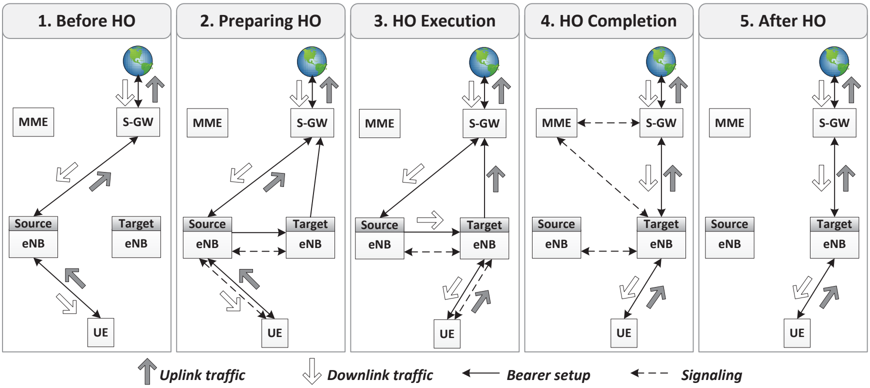


Figure 2.11: Steps of the X2-based handover procedure in LTE [NCG13].

the user traffic path at S-GW from source eNode-B to the target eNode-B, as MME is already aware of the handover. In addition, data forwarding has to be performed between the source eNode-B and target eNode-B via the S-GW because of the unavailability of the direct forwarding path.

2.4.7 LTE Quality of Service

Generally speaking, Quality of Service (QoS) involves the data delivery between two nodes with certain constraints on latency, error rate, jitter, and bit rate etc. In LTE, an end-to-end class based QoS architecture has been defined based on data flows and bearers as shown in Figure 2.12. Data flows are mapped to bearers so that an end-to-end QoS in the LTE network is provided via an EPS bearer which itself relies on the services of its constituent bearers i.e., Radio, S1, and S5/S8 bearers. Table 2.2 shows the QoS information for EPS bearers which must be supported by the network nodes. In order to achieve complete end-to-end QoS, the support of external bearers is also required which is not within the scope of LTE standards.

Owing to the fact that network services are usually classified into realtime services and non-realtime services, the bearers are also classified into two categories based on their offered QoS. These two bearer types are ‘GBR’ (Guaranteed Bit Rate) and ‘non-GBR’ bearers. As the name implies, a GBR bearer guarantees to offer a minimum bit rate for which the dedicated transmission resources are permanently allocated during the bearer establishment or modification. GBR bearers

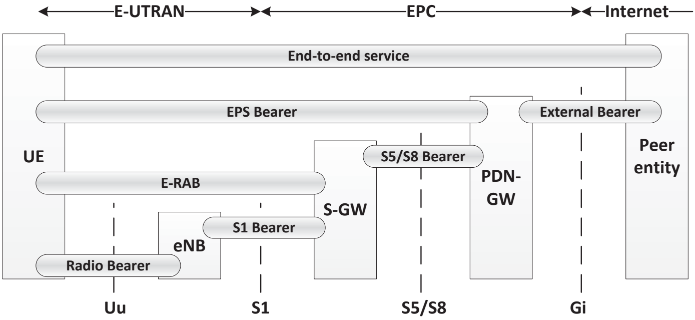


Figure 2.12: LTE bearer architecture [36.11].

are suitable to provide services like voice and video telephony. In contrast to this, non-GBR bearers do not guarantee a minimum bit rate and therefore are used to support applications such as web browsing, FTP file transfer, email etc.

In order to meet intended QoS requirements at the radio interface, each bearer is assigned a class identifier (i.e., QCI) and an 'Allocation and Retention Priority (ARP)' at the eNode-B. Each QCI is characterized by priority, packet delay budget, and acceptable packet loss rate. 3GPP has standardized a number of QCIs which ensures the uniform traffic handling behavior throughout the network irrespective of the equipment manufacturers. The set of standardized QCIs and their QoS requirements are provided in Table 2.2.

The ARP of the bearer is used in relative prioritization and preemption decisions such as in call admission control and new bearer establishment requests. However, an established bearer's ARP has no influence on bearer-level packet forwarding treatment (e.g., scheduling policy, queue management policy, rate control policy etc.). Instead, such packet forwarding treatments must be determined by the other bearer-level QoS parameters such as QCI, GBR etc.

2.5 Beyond LTE

The evolution of mobile communication systems did not stop after the development of the LTE network and the eyes turned towards the next development, that is, the true 4G technology named LTE-Advanced. The proposal of LTE-Advanced was submitted as a candidate 4G system to ITU-T in 2009 which was approved and later on its standardization was finalized by 3GPP in April 2011. The key requirements of LTE-Advanced are listed below:

Table 2.2: LTE standardized QCIs and their parameters [3GP12].

Bearer type	QCI	Priority	Packet delay budget (ms)	Packet error loss rate	Example services
GBR	1	2	100	10^{-2}	Conversational voice
	2	4	150	10^{-3}	Conversational video (live streaming)
	3	3	50	10^{-3}	Real time gaming
	4	5	300	10^{-6}	Non-conversational video (buffered streaming)
non-GBR	5	1	100	10^{-6}	IMS signaling
	6	6	300	10^{-6}	Video (buffered streaming)
	7	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
	8	8	300	10^{-6}	TCP based (e.g., www, e-mail, chat, FTP, p2p)
	9	9	300	10^{-6}	

- Peak data rate of 1Gbps for downlink and 500Mbps for uplink.
- 30bps/Hz and 15bps/Hz as the peak spectral efficiency for downlink and uplink, respectively. This is 3 times greater than that of LTE.
- Less than 50ms transition time from Idle to Connected state and less than 5ms radio network delay for individual packet transmissions.
- Scalable bandwidth and spectrum aggregation with transmission bandwidths up to 100MHz in downlink and uplink. The spectrum aggregation allows non-contiguous spectrum to be used.
- Backward compatibility with the existing LTE standards. This implies that LTE user terminals should be supported in an LTE-Advanced networks.
- Enhanced cell edge coverage which provides two times higher user throughput than that of LTE.
- The mobility and coverage requirements are the same as mentioned for LTE in 3GPP Rel-8 standard.

In order to fulfill the above mentioned requirements, LTE-Advanced makes use of several recently developed cutting-edge technologies. Over the next subsections an overview of a few of the LTE-Advanced key technologies is provided.

2.5.1 Carrier Aggregation

LTE-Advanced targets 1Gbps as the downlink data throughput which cannot be achieved with 20MHz bandwidth despite significant improvements in the spectral efficiency. The only way to achieve the higher data rates is to increase the overall spectrum bandwidth available to the system. Though LTE-Advanced allows to use up to 100MHz bandwidth, it is difficult to find a contiguous frequency band of this size. In many areas only small bands are available which are of smaller size such as 10MHz. As a result LTE-Advanced has to rely on carrier aggregation, a technique of ‘bonding’ together separate frequency bands. To an LTE terminal, each component frequency band appears as an individual LTE carrier, while an LTE-Advanced terminal, using the carrier aggregation, can exploit the total aggregated bandwidth. See Figure 2.13 for a graphical representation of three types of carrier aggregation.

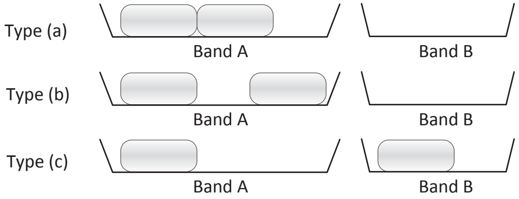


Figure 2.13: Three types of carrier aggregation. Type (a) and (b) represent intra-band carrier aggregation with contiguous and non-contiguous components, respectively. Type (c) represents inter-band carrier aggregation.

2.5.2 Enhanced Uplink Multiple Access

LTE is based on SC-FDMA which requires carrier allocation across a contiguous block of spectrum and hence prevents the scheduling flexibilities inherent in pure OFDM. LTE-Advanced adopts clustered SC-FDMA which is similar to SC-FDMA but allows non-contiguous groups of sub-carriers to be allocated for transmission by a single UE. This enables uplink frequency-selective scheduling and consequently improves the uplink spectral efficiency while maintaining the backward compatibility with LTE.

2.5.3 Enhanced Multiple Antenna Transmission

According to 3GPP Rel-8 standard, LTE supports a maximum of four spatial layers of transmission in downlink (4x4 MIMO) and a maximum of one spatial layer per UE (1x2 MIMO) in uplink. LTE-Advanced supports downlink transmission using up to eight spatial layers and the UE supports up to four transmitters allowing the possibility of up to 4x4 MIMO transmission in uplink. This significantly improves the single user peak data rate and helps achieve the target spectral efficiency.

2.5.4 Coordinated Multipoint

Coordinated multipoint (CoMP) is an advanced variant of MIMO which promises improved data rates, cell-edge throughput, and system performance in both high load and low load scenarios. CoMP is essentially a range of different techniques that enable the dynamic coordination of transmission and reception over a variety of different base stations. In CoMP, a number of geographically separated eNode-Bs dynamically coordinate to achieve joint scheduling and transmission as well as joint processing of the received signals. In this way, a UE at the cell-edge can be served by two or more eNode-Bs turning the inter-cell interference into useful signals to enhance the coverage at the cell-edge.

2.5.5 Relaying

Relaying is mainly used to improve urban or indoor throughput, to add dead zone coverage, and to extend coverage in rural areas. Relaying in LTE-Advanced is different from conventional repeaters which just re-broadcast the signal. A relay delivers much more by receiving the actual transmission, demodulating and decoding the data, applying error correction, etc. and then retransmitting a new signal. As a result, the signal quality is significantly enhanced by the use of an LTE relay. Typically, the UEs communicate with the relay node which in turn wirelessly communicates with a donor eNode-B (see Figure 2.14).

2.5.6 Self Organizing Network

LTE self-organizing and self-optimizing network (SON) enhancements substantially simplify and automate many tasks related to radio planning and operation & maintenance (O&M). The main aspects of SON are as follows [Agi11]:

- Self configuration: With SON, a range of specific events (e.g., introducing a new femto-cell) can be automated using the O&M interface and the network management module.

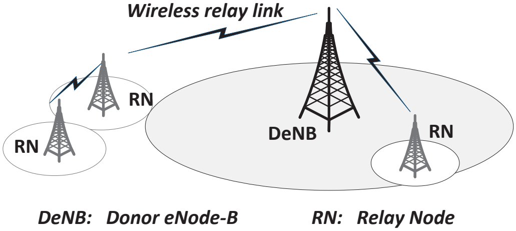


Figure 2.14: Relaying in LTE-advanced.

- Self optimization: SON recursively strives for an optimized network setting with the help of continuous analysis of environmental data such as UE and eNode-B measurements.

- Self healing: SON is capable of recovering from exceptional events triggered by unusual circumstances, e.g., dramatic changes in interference conditions, or a ping pong situation in which a UE continuously switches between macro and femto cells.

2.6 IEEE 802.11 Networks

802.11 belongs to the IEEE 802 family of standards for Local Area Network (LAN) technologies. The LAN technologies mainly encompass the lowest two layers of the OSI (Open Systems Interconnection) reference model, i.e., data link layer (MAC) and physical (PHY) layer. Figure 2.15 shows various components of the 802 family and their relationship with the OSI reference model. It is obvious from the figure that the members of the 802 family are assigned an identification number which is appended to the family number (i.e., ‘802’) and separated by a dot. For example, 802.1 specifies the management features of the network, 802.2 specifies the common link layer, termed as Logical Link Control (LLC) layer which can be employed by lower layer LAN technology, 802.3 is a standard related to Ethernet LAN technology, 802.5 is the specification for the Token Ring etc. In the same way, 802.11 is just another link layer technology that provides its services to the 802.2 LLC layer to allow mobile network access using radio waves.

The primary goal of the 802.11 standard was the design of a wireless LAN technology which is simple and robust. The MAC layer should be able to support various physical layers, each of which exhibits different medium sense and radio transmission characteristics. Initially, Infra-red and spread spectrum radio trans-

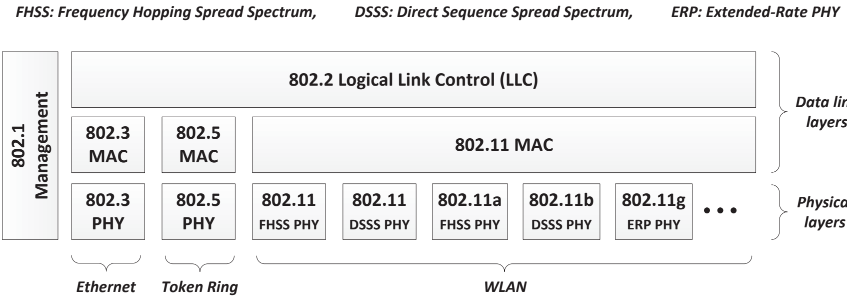


Figure 2.15: IEEE 802 family components and their place in the OSI reference model [Gas05].

mission techniques were chosen as the physical layer candidates. In addition, the aim was that WLAN should provide built-in power management functions to save battery power and it should be able to operate worldwide. This is the reason why the ISM (Industrial, Scientific and Medical) band, which is reserved internationally as license free spectrum, was chosen for the radio transmissions. The standard targeted 1Mbps data rate as mandatory and 2Mbps as optional.

2.6.1 System Architecture

In the 802.11 standard, the term station (STA) is used for a node which is equipped with a wireless LAN interface or adapter. The stations are arranged in logical groups called Service Sets. A Basic Service Set (BSS) is a group of stations which communicate with one another via a specialized station known as an Access Point (AP). In this system architecture, which is called infrastructure-based architecture, the stations in a BSS do not communicate directly with one another. Instead, they rely on the AP to forward their communication to the destination stations (See Figure 2.16). The AP connects the stations insides its coverage to the Distribution System (DS) which is a backbone network providing access to external networks such as the Internet. In common practice, an AP connects to the DS using a wired link, however, the 802.11 specification leaves the potential for this link to be wireless.

In infrastructure mode, multiple BSSs can be connected to the same DS to form an Extended Service Set (ESS). In an ESS a station can move between two APs without interrupting an ongoing connection using the roaming feature of 802.11 (See Figure 2.17).

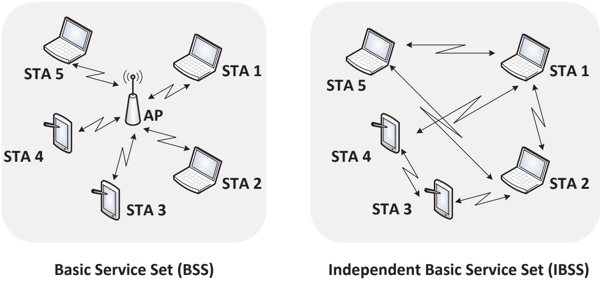


Figure 2.16: Infrastructure based (BSS) and ad-hoc based (IBSS) system architectures.

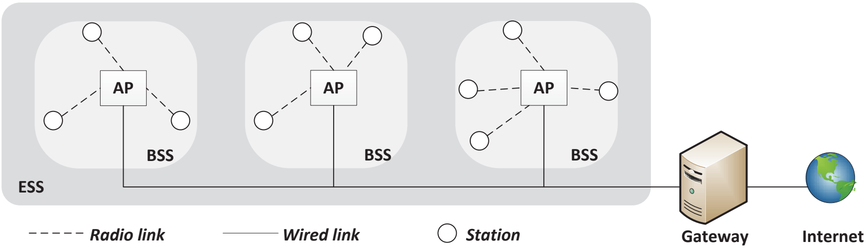


Figure 2.17: Connection of BSSs to form an Extended Service Set (ESS).

The 802.11 standard also supports another system architecture referred to as ad-hoc based architecture where the stations communicate directly without requiring the presence of any AP. A group of stations operating in ad-hoc mode form an Independent Basic Service Set (IBSS) as shown in Figure 2.16. The ad-hoc mode communication is useful when no WLAN infrastructure is available, e.g., in outdoor meetings or in natural disaster scenarios.

2.6.2 Protocol Architecture

The protocol architecture specified by the 802.11 standard has been shown in Figure 2.18. The standard covers the MAC and PHY layers. The MAC layer is responsible for medium access, fragmentation and reassembly of user data frames, as well as, for the data encryption. The physical layer is further divided into two sublayers which are not specified in the OSI reference model: the Physical Layer Convergence Protocol (PLCP) and Physical Medium Dependent sublayer (PMD). The PLCP sublayer provides a carrier sense signal and a transmission technology independent common service access point (SAP) to the PHY layer. PMD sublayer takes care of encoding/decoding of signals.

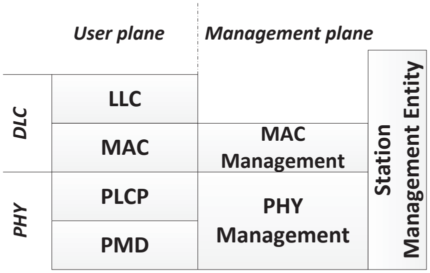


Figure 2.18: IEEE 802.11 protocol architecture and management.

The 802.11 standard also specifies management layers and the station management entity in the protocol architecture. The MAC management handles the roaming of a station between the access points, facilitates the association of a station to an access point, performs power management to save battery power, controls the authentication mechanism, encryption, and synchronization of a station with respect to an access point. The PHY management mainly supports the channel tuning and maintains the Management Information Base (MIB). The station management entity interacts with both aforementioned management layers to control frequency channel selection and to adjust the transmission power as well as to carry out additional higher layer functions, e.g., control of bridging and interaction with the distribution system in the case of an access point.

2.6.3 Physical Layer

The initial revision of 802.11 was released in 1997 which standardized three physical layer technologies: Frequency Hopping Spread Spectrum (FHSS), Direct Sequence Spread Spectrum (DSSS) and Infrared light (IR). The standard specified the 2.4GHz ISM band and a radio spectrum of 22MHz for the transmissions when using FHSS and DSSS schemes. The transmission power was limited to 100mW in order to reduce interference to other communications taking place in the same ISM band. The standard also defined two data rates at the physical layer, i.e., 1 and 2Mbps.

A few months later, an extension of 802.11 standard, termed 802.11b, was published which offered enhanced data rates of 5.5 and 11Mbps using the DSSS scheme. In the earlier standard, DSSS used BPSK (Binary Phase Shift Keying) and QPSK (Quadrature Phase Shift Keying) modulation schemes for 1 and 2Mbps data rates, respectively. The modulation scheme proposed for 802.11b was also QPSK but the higher data rates were achieved by reducing the spreading factor in DSSS.

The 2.4GHz ISM bands are heavily in use by non-802.11 traffic. In 1999, the 802.11 working group released another extension, 802.11a, which operates in the 5GHz ISM band. This frequency band has far fewer non-802.11 devices which reduces the interference and helps achieve higher data rates. 802.11a is based on the Orthogonal Frequency Division Multiplexing (OFDM) transmission scheme which has been explained in Section 2.4.3.1. Using the same radio spectrum size as in 802.11 and 802.11b, but using higher order modulation schemes data rates of up to 54Mbps are supported in 802.11a. The new modulation schemes include 16QAM (Quadrature Amplitude Modulation) and 64QAM.

In 2001, 802.11a products were commercially available but the users still had a desire to obtain higher data rates while retaining the backward compatibility with the installed 802.11b infrastructure. This resulted in the 802.11g standard which uses OFDM in the 2.4 GHz ISM band and offers a bit rate comparable to 802.11a. Due to lower operating frequency, 802.11g networks have better coverage compared to that of 802.11a networks. Most of the physical layer specifications of 802.11g are built on existing work of 802.11b with slight modifications in order to provide backwards compatibility.

In response to growing market demand for higher performance WLAN, the task group IEEE 802.11n was formed in 2004. The new standard, released in 2009, improves upon the previous 802.11 standards by introducing Multi-Input Multi-Output (MIMO) technology. The OFDM technology used for 802.11a/g proved to be resilient against the multi-path nature of the channel and, therefore, also adopted for 802.11n. The final system is categorized as a MIMO-OFDM system. 802.11n mandates the interoperability with the legacy 802.11a/g systems and can operate both in 20 and 40MHz bandwidth. With four transmit antennas and 40MHz bandwidth a maximum data rate of 500Mbps can be achieved at the physical layer. Another flexibility feature of 802.11n is its capability to operate in both 2.4GHz and the less crowded 5GHz ISM bands.

Other mentionable 802.11 variants include 802.11ac and 802.11ad which are currently under development. 802.11ac will use wider bandwidths up to 160MHz, up to eight MIMO antennas, and higher order modulation schemes like 256QAM. 802.11ad is a tri-band WLAN which will operate in 2.4, 5, and 60GHz ISM bands and provide a maximum throughput of 7Gbps.

2.6.3.1 802.11a PHY

The 802.11a uses an OFDM transmission scheme with 52 sub-carriers (48 data and 4 pilot) that can be modulated using BPSK, QPSK, 16QAM, or 64QAM. All sub-carriers are modulated with the same modulation scheme and the duration of each modulation symbol is independent of the selected modulation scheme. In this way, within one OFDM symbol duration, $b \cdot C$ bits can be transmitted when C sub-carriers are used and each sub-carrier transports one modulation symbol carrying b bits.

In order to minimize the interference several operating channels have been standardized in 802.11a. These channels start from 5GHz and have 20MHz spacing. In reality the occupied bandwidth of a channel is only 16.6MHz instead of 20MHz. This is because 802.11a defined 64 sub-carriers each of size 20MHz/64 = 312.5kHz. However, only 52 sub-carriers are used in practice and the rest 12 sub-carriers are reserved for other purposes. In 5GHz ISM bands, there are of the 12 channels available which can support interference-free operation of overlapping 802.11a cells.

For the transmission of a PHY PDU, the first step is to perform data scrambling which evens the distribution of 0 and 1 bits and after which a Forward Error Correction (FEC) based on convolution coding is employed. The use of FEC introduces redundancy in the transmission to make it resilient against bit errors. The amount of added redundancy can be quantified with the help of a 'coding rate' which is a ratio of number of data bits to the number of code bits. 802.11a defines three coding rates which are as follows: $\frac{1}{2}, \frac{2}{3}$, and $\frac{3}{4}$. Combinations of modulation schemes and coding rates produce different data rates at the physical layer as listed in Table 2.3

Table 2.3: Physical layer data rates and dependent parameters for 802.11a.

Data Rate (Mbps)	Modulation	Coding rate	Coded bits per sub-carrier	Coded bits per OFDM symbol	Data bits per OFDM symbol
6	BPSK	1/2	1	48	24
9	BPSK	3/4	1	48	36
12	QPSK	1/2	2	96	48
18	QPSK	3/4	3	96	72
24	16-QAM	1/2	4	192	96
36	16-QAM	3/4	4	192	144
48	64-QAM	2/3	6	288	192
54	64-QAM	3/4	6	288	216

The data delivered from the MAC layer (MAC PDU) is considered as a payload of the PHY PDU (PPDU) which has to go through the scrambler and convolution encoder before the transmission as described earlier. The basic structure of the PPDU for 802.11a is shown in Figure 2.19. The 'PLCP preamble' consists of 12 symbols and is used for frequency acquisition, synchronization and channel estimation. The 'signal' contains the control information required by the physical layer, e.g., the 4bit 'rate' field determines the modulation scheme, the 'length' field indicates the payload size in bytes, the 'parity' bit represents the even parity of the first 16bits, and the 6 'tail' bits are always set to zero. The 'data' field contains user data and a 'service' subfield which helps synchronize the descrambler of the receiver. In addition, the 'tail' bits are used to reset the encoder and the 'pad' bits may be used so that the PPDU can be mapped to an integer number of OFDM symbols.

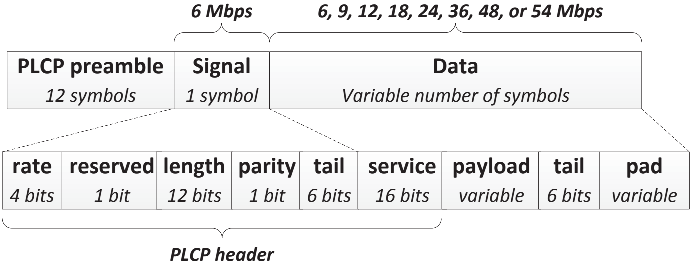


Figure 2.19: IEEE 802.11a physical layer PDU format.

2.6.4 Medium Access Control Layer

The MAC layer has to support several tasks including roaming, authentication, power control etc. However, the most important task is the control of medium access so that a station transmits only if the radio channel is free. IEEE 802.11 standard specifies three MAC access mode. The mandatory basic access based on a version of CSMA/CA is provided by the Distributed Coordination Function (DCF) and contention-free service is provided by the Point Coordination Function (PCF). Between the free-for-all of the DFC and the precision of the PCF there is a third mode termed as Hybrid Coordination Function (HCF). The contention-free services of PCF are supported only in infrastructure networks, however, DCF and HCF may be employed in any network. As illustrated in Figure 2.20 both PCF and HCF are built on top of DCF.

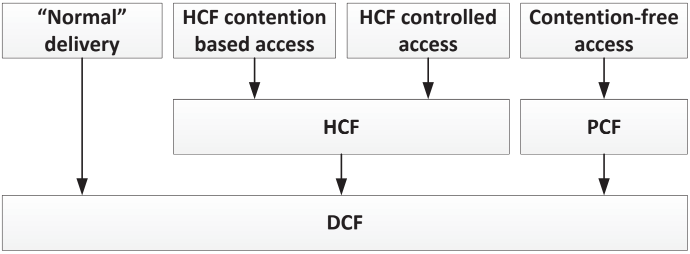


Figure 2.20: IEEE 802.11 MAC coordination functions [Gas05].

For all access modes, several parameters for controlling the waiting time and priorities of medium access have been defined. Inter-frame spacing is one of these parameters which plays a vital role in coordinating access to the transmission medium. 802.11 defines four different types of inter-frame spaces, three of which are used to determine medium access. As a part of collision avoidance mechanism of the 802.11 MAC, the stations delay their transmission until the medium is sensed idle. Once the medium is free, the stations have to wait a certain period of time before they can take hold of the channel. This time period is determined by the inter-frame spaces. Selecting an inter-frame space of a shorter time period, the high priority traffic takes hold of the channel before low priority traffic has a chance to try. An inter-frame space represents a fixed amount of time independent of the transmission speed.

Figure 2.21 shows the relationship between different types of inter-frame spaces. Short Inter-Frame Space (SIFS) is used for the highest priority control traffic, such as acknowledgements of data packets or polling responses. PCF Inter-Frame Space (PIFS) is used by the PCF during contention-free operation. For example, an access point polling other nodes has to wait PIFS for medium access. DCF Inter-Frame Space (DIFS) represents the longest time period which is used by contention-based services to access the medium. Stations may have immediate access to the medium if it has been free for a time period longer than the DIFS. Finally, Extended Inter-Frame Space (EIFS) is used only when an error is encountered in frame transmission. The values of inter-frame spaces is defined in relation to 'slot time'. Slot time is derived from the PHY dependent parameters such as medium propagation delay, transmitter delay etc.

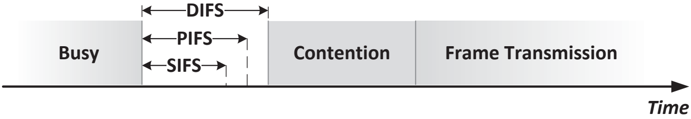


Figure 2.21: Medium access and inter-frame spaces.

2.6.4.1 Basic DCF with CSMA/CA

The mandatory basic access mechanism of 802.11 is based on a random access scheme with carrier sense and collision avoidance through random backoff. This scheme is termed Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). In this scheme, if the medium remains idle for at least the duration of DIFS, a node can access the medium immediately. However, if the medium is sensed as busy after DIFS duration, the node enters a contention phase. In this phase the node chooses a ‘random backoff time’ within the ‘contention window’ and defers the medium access for this random amount of time. After this period of time, if the medium is still busy then the node has lost this cycle and has to wait for the next chance, i.e., until the medium is idle again for at least DIFS duration. However, if the medium is sensed idle after the random time period is elapsed, the node can access the medium immediately. The purpose of randomized waiting time before accessing the medium is to avoid that all stations access the medium at the same time resulting in a collision situation.

Such a basic CSMA/CA scheme is considered as unfair because of the following reasons. Each node gets the equal probability of accessing the medium irrespective of the overall time this node has already waited for the transmission. To help this situation, 802.11 introduced a ‘backoff timer’. With this modification, each node chooses a random waiting time and continuously senses the medium. If a certain node senses the medium as busy during its random waiting time, it stops its backoff timer, waits for the medium be free again for DIFS and resumes its backoff timer. In this way the deferred stations do not select a randomized backoff time again, but continue to count down. The stations that have already waited for a long time get advantage over the stations that have just entered the contention phase.

The randomized backoff time is measured in terms of the earlier mentioned slot times. The contention window starts with a size of 7 slots and on each collision its value doubles up to a maximum of 255 slots. This algorithm is called ‘exponential backoff’ and has been used in IEEE 802.3 standard for Ethernet.

Figure 2.22 shows a sender accessing the medium and sending unicast data in the basic DCF access mechanism. On successfully receiving the data, the receiver answers directly with an acknowledgement (ACK) after SIFS time duration. The reception of the ACK at the receiver ensures the correct reception of a frame on the MAC layer, which is particularly important in error-prone wireless connections. In case no ACK is received, the sender has to retransmit the frame. But now the sender has to compete for the medium access as described earlier. After a limited number of retransmissions, the MAC layer aborts the operation and reports the failure to the higher layers.

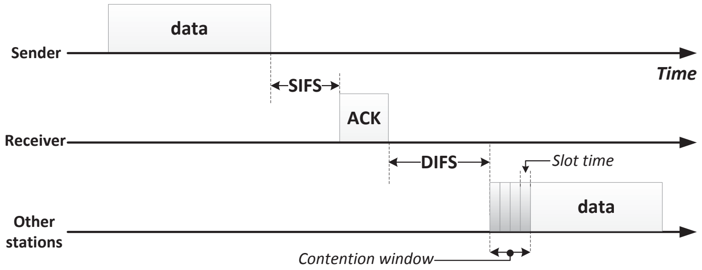


Figure 2.22: Unicast data transmission in IEEE 802.11 using basic DCF access mechanism with CSMA/CA.

2.6.4.2 DCF with RTS/CTS Extension

The basic DCF access mechanism suffers from the 'hidden terminal problem'. This problem occurs if one station can receive two others, but those stations cannot receive each other. The two stations may sense the medium idle, send a frame, and cause a collision at the receiver in the middle. To overcome this situation, 802.11 extends the basic DCF access mechanism by introducing two additional control packets: Request To Send (RTS) and Clear To Send (CTS). The use of RTS/CTS is illustrated in Figure 2.23. After receiving the medium access, the sender issues RTS control packet which includes information about the intended receiver of the data transmission and its duration. Every node receiving this RTS sets its Network Allocation Vector (NAV) based on the information contained in the RTS. The NAV determines the earliest time at which the stations can try to access the medium again. The intended receiver of the data transmission answers with a CTS control message which also contains similar information about upcoming data transmissions and helps stations adjust their NAV. Now all stations within receiving distance around the sender and receiver are aware of the imminent data transmission and hence the medium is reserved exclusively for that data transmission. Finally, the sender sends the data and receives the ACK if the transfer was correct. This concludes the transmission and stations can start competing for the medium access again.

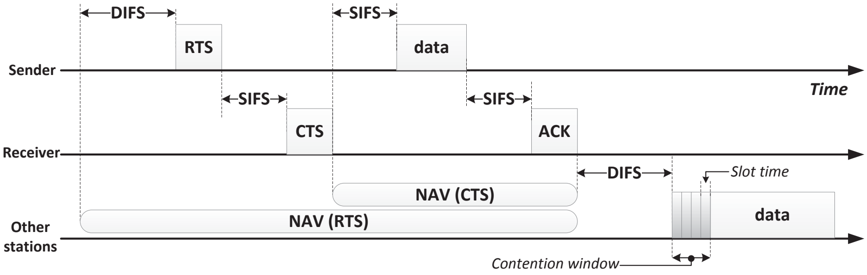


Figure 2.23: Unicast data transmission in IEEE 802.11 using DCF access mechanism with RTS/CTS extension.

2.6.4.3 Point Coordination Function (PCF)

The PCF access mechanism can be used to offer time-bounded services where a guarantee of maximum access delay is required. The aforementioned two access mechanisms can offer only 'best effort' service due to the involved contention phase and random backoff timer. Using PCF, an access point controls the medium access of its associated stations through polling. The access point splits the access time into 'super frame' periods which comprise a contention-free period and a contention period. The contention period is used for the two access mechanisms based on DCF as presented above. During the contention-free period the access point sends downstream data to the first station. This station has to reply immediately after SIFS, otherwise, the access point can poll the next station after waiting for PIFS. This cycle continues until all stations are polled. Finally, the access point issues an end-marker to indicate the end of the contention-free period. Afterwards, the contention period may start. Alternating periods of contention-free service and contention-based service repeat at regular intervals, which are called the contention-free repetition interval.

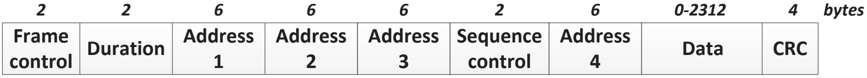
2.6.4.4 Hybrid Coordination Function (HCF)

Within HCF there are two channel access mechanisms: Enhanced Distributed Channel Access (EDCA) and HCF Controlled Channel Access (HCCA). With

EDCA, a station with higher priority traffic waits a little less before transmitting as compared to a station with low priority traffic. To achieve this 802.11 has introduced a shorter Arbitrary Inter-Frame Space (AIFS) for high priority traffic. The HCCA works in a similar as the PCF with the difference that here contention-free period can be initiated at almost anytime during a contention period. This kind of contention-free period is termed Controlled Access Phase (CAP). Moreover, HCF also defines the Traffic Class (TC) and the Traffic Streams (TS) to provide respective Quality of Service (QoS). The access point with HCF support, maintains a summary of the queue lengths of each TC of each station. This information helps the access point determine which stations will be allocated transmission opportunities during the contention-free period. Finally, the access point polls the stations according to their traffic class priority in the same way as described above for the PCF access mechanism.

2.6.4.5 MAC Frame Format

Figure 2.24 shows the basic structure of the IEEE 802.11 MAC frame format. The two byte long ‘frame control’ field contains several sub-fields used for control, power management, and security mechanisms. The ‘duration’ field indicates the channel busy time used to adjust the NAV value of stations in the access mechanism with RTS/CTS. There are four address fields each of which can hold a MAC address of 48 bit size. The significance of these address fields is determined using the information in the ‘Frame control’ field. The ‘sequence control’ field holds the frame sequence number in order to avoid duplicate frames. A 32bit checksum is included in ‘CRC’ field and the user data of an arbitrary length is inserted in the ‘data’ field of the frame.



2	*2*	*6*	*6*	*6*	*2*	*6*	*0-2312*	*4*	*bytes*
Frame control	Duration	Address 1	Address 2	Address 3	Sequence control	Address 4	Data	CRC	

Figure 2.24: IEEE 802.11 MAC frame structure.

2.7 3GPP Networks–WLAN Interworking

WLAN and 3GPP wireless access technologies (i.e., WWANs) may be seen to compete but in reality they complement each other. For example, WLANs are suitable for providing hotspot coverage where very high data rate wireless services are required in a small area with limited mobility. Such deployments of WLAN to offer hotspot coverage have emerged in dense populated areas including restaurants, hotels, convention centers, railway stations, airports, etc. On the other hand, 3GPP access technologies are designed to support wide coverage and high mobility and therefore well suited to the areas with low-density of demands for wireless services requiring high mobility. Intuitively, an integration of these two types of access technologies can bring significant advantages in providing wireless multimedia and other high data rate services to large populations. In a simple use case, a user terminal can take advantage of these integrated heterogeneous networks by accessing high bandwidth data services where WLAN coverage is offered, while accessing 3GPP based wide area networks at the other places. However, in order to offer an effective heterogeneous network access, the proposed solution must provide seamless mobility between the access technologies allowing the continuity of ongoing sessions.

In general there are three approaches to integrate WLAN with 3GPP access networks. Considering 3G networks as an example, their interworking with WLAN is illustrated in Figure 2.25 and described as follows.

- **Mobile IP approach**: This approach which is also called loose coupling, introduces ‘Mobile IP’ to manage the mobility of a user terminal between two network types. In Mobile IP, a user terminal connects to a visited network and establishes a connection to the home network using IP-over-IP tunneling. To all corresponding hosts, this user terminal appears to be in the home network even when it does handover from one visited network to another. A detailed overview of Mobile IP can be found in [C. 96] and [PJA04]. This approach requires the Mobile IP mechanism implemented in the user terminal as well as in some of the network entities. The benefit of this approach lies in its support for all IP network types where one network can evolve without interfering the integration architecture. The drawback is the delay involved in Mobile IP control signaling during the handover.

- **Gateway approach**: In this approach, a logical node termed Gateway, connects two wireless access networks. The information between two networks is always exchanged through the Gateway. The Gateway is responsible for inter-conversion of signalling and assists in handover procedures by forwarding the packets of roaming users. This approach allows independent operation of two networks and a seamless inter-system roaming without excessive handover delays introduced by Mobile IP.

- **Emulator approach**: This is also called tight coupling approach. This approach views WLAN as an access stratum in 3G networks just like another Serving GPRS Support Node (SGSN). All packets routing and forwarding as well as the handovers are carried out by a 3GPP core network. Though this approach offers reduced packet loss and delay during the handover, it lacks the flexibility due to tight coupling and requires both networks to be owned by the same operator. Another drawback is the potential bottleneck at the Gateway GPRS Support Node (GGSN) through which all traffic to the Internet must be routed.

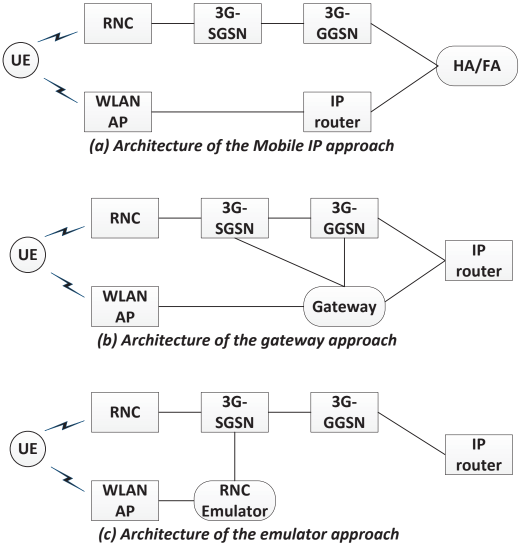


Figure 2.25: Integration of WLAN with 3G networks [Gar07].

2.7.1 Integration in SAE

As discussed earlier in Section 2.4, 3GPP introduced an evolved core network architecture termed System Architecture Evolution (SAE) in Rel-8. An important feature of SAE is the standardization of system architecture to integrate non-3GPP access technologies, e.g., WLAN, WiMAX, etc. in 3GPP networks. In fact, the architecture design described in [3GP11a], allows the interconnection with just about any access technology whether it is wireless or fixed. This has been achieved by making the access to the PDN gateway generic so that a terminal's association to the network, access to general IP services as well as other network features like user subscription management, billing, encryption, policy control, and VPN connections can be made independent of the access technology.

SAE also takes care of seamless user mobility during the handovers between different access technology types. Two possible mobility management options are using either 'host-based' or 'network-based' Mobile IP. Host-based mobility means that the user terminal implements Mobile IP functionality and IP tunnels are established between the user terminal and the PDN gateway across the access network. Network-based mobility means that there are Mobile IP aware entities in the access network which assist the mobility by acting on behalf of the user terminal. The benefit of the host-based approach is that it can work with any access network as long as the user terminal supports the Mobile IP functionality. On the other hand, the network-based approach simplifies the user terminal implementation but requires the support of Mobile IP functions in the network itself.

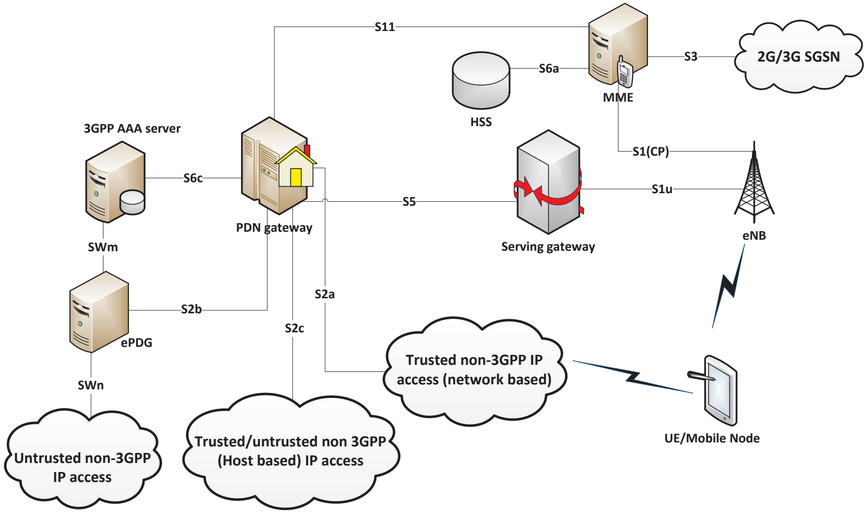

S11
S6a
S3
2G/3G SGSN
MME
HSS
3GPP AAA server
S6c
S1(CP)
S1u
S5
PDN gateway
Serving gateway
eNB
SWm
S2a
S2b
S2c
ePDG
Trusted non-3GPP IP access (network based)
SWn
UE/Mobile Node
Untrusted non-3GPP IP access
Trusted/untrusted non 3GPP (Host based) IP access

Figure 2.26: Integration of 3GPP and non-3GPP technologies in the SAE architecture [3GP11a].

At the time of integration, the network operator has to decide whether the access technology to be integrated is 'trusted' or 'un-trusted' in terms of network security. Generally speaking, an un-trusted access includes any network type that is not under direct control of the operator (e.g., public hotspot, office WLAN etc.) or a network which does not provide sufficient security like authentication, encryption, etc. Trusted access networks are usually operator owned WLAN or WiMAX networks which support over-the-air encryption and authentication methods.

Figure 2.26 represents the integration architecture proposed by 3GPP [3GP11a]. It can be seen that non-3GPP technologies are integrated with 3GPP technologies through one of the three interfaces (S2a, S2b, S2c) provided by SAE. The description of the each interface is as follows:

- **S2a** - provides the integration path between the trusted non-3GPP IP networks and 3GPP networks. In this case the mobility is handled by the network-based mobility solution, i.e., Proxy MIPv6 [GLD$^+$08].
- **S2b** - provides the integration path between the un-trusted non-3GPP IP networks and 3GPP networks. In this case also the mobility is handled by the network-based mobility solution. The S2b interface is connected to the un-trusted access network via a new logical node called evolved Packet Data Gateway (ePDG). The user terminals exchange the traffic with the ePDG in a secure way over the encrypted tunnels. This creates a logical association between the ePDG and the user terminals termed SWu interface. The ePDG then connects to the PDN gateway using the S2b interface. Another interface of the ePDG, called SWm, connects it to the 3GPP AAA server. It is used to fetch authentication, authorization, and accounting related parameters from the AAA server in order to support IPsec [KS05] tunnel setup between the ePDG and the user terminals.
- **S2c** - provides the integration path between both trusted and un-trusted non-3GPP IP networks and 3GPP networks. In this case the mobility is handled by the host based mobility solution, i.e., Dual Stack MIPv6 [H. 09]. This implies an overlay solution which does not require any specific support from the underlying access network.

This work follows 3GPP's proposal for integration of non-3GPP access technologies into the SAE architecture. For the purpose of the simulation based study of such 3GPP compliant heterogeneous networks, the next chapter discusses the development of a simulation environment, where LTE and 802.11a networks are integrated.

3 Simulator for Heterogeneous Network Access Technologies

3.1 Introduction

Network designers and developers are constantly being challenged to satisfy the user demands of high performance networking capabilities. In order to keep up with the demands, service providers are expanding their networks, network researchers are developing revolutionary communication technologies, and equipment manufacturers are rapidly improving the performance of their devices. This fast evolution has brought the whole communication network industry to a complex landscape of ever growing complexity of communications protocols and their large scale deployments.

There are several available tools and feasible methods which a network engineer can use to perform research and development in the area of communication networks. They include mathematical analysis, prototype implementation in a test-bed, modeling and simulation, as well as, other hybrid simulation approaches. Mathematical analysis based modeling has the advantage of usually solving the problems faster than the simulations. However, mathematical analysis based modeling cannot be used to evaluate end-to-end communication network path unless a decomposition is carried out via Kleinrock's assumption of independence or the problem is solved using hop-by-hop single system analysis. Both of these approaches demand a significant amount of development time to achieve an exact modeling of the system. Even so, a slight complexity in network protocol operation can bring large scale modeling difficulties and loss of accuracy. Therefore, researchers have to rely on approximate models obtained through reducing the original generic model to a typical and representative analytical path [Kle75].

Prototype implementation in a test-bed provides insight on feasibility of a proposed protocol or algorithm to an actual situation by considering the aspects of both real world and protocols. However, there are several disadvantages to this approach, like, the ability to emulate only small scale scenarios, significantly high cost of hardware as well as other real world difficulties and hardware engineering problems.

Network simulation is a technique where network behavior is modeled either by computing the interaction between network entities with the help of mathematical relations or playing back the captured data composed of observations from a real network. The network model based on discrete event simulations can be used to study large scale networks in a realistic and extremely accurate way. The broad applicability of discrete event simulations to real world problems and its ability to evaluate network behavior to a desired level of details makes it a favorite tool for network researchers and engineers. The main disadvantage involved in the study of network modeling through simulations is the requirement of computing power which, in many cases, increases rapidly with the scale of the investigated networks and the required level of modeling details. In some cases mathematical analysis can be used along with the simulations for faster speed without losing much accuracy. Such techniques are typically called hybrid simulations [LY12].

Owing to the above mentioned facts, network modeling through simulations turns out to be the most appropriate method for investigations of this work. Although there are many popular discrete event network simulators like NS [NS-13], OMNET++ [Env13], NetSim [Net13] etc. but a distinct place belongs to OPNET [OPN13]. The reason is OPNET's well documented and rich library of already developed and tested simulation models of common network entities and protocols. This helps considerably shorten the development cycle of custom simulation environment. Though OPNET is a commercial tool but an academic license can be obtained for free. The modeling concepts discussed in this chapter consider OPNET Modeler as a network simulation tool; however, the provided guideline is general enough to be used in conjunction with any other discrete event simulator.

The rest of this chapter has been organized as follows: Section 3.2 provides the general introduction to the OPNET modeler which also serves as a prerequisite towards the better understanding of the tool. A reference heterogeneous network architecture is discussed in Section 3.3 and its implementation details are described in Section 3.4 . An overview of user traffic models is given in Section 3.5 and finally, Section 3.6 details the common techniques used to perform statistical analysis of the simulation output data.

3.2 OPNET Modeler

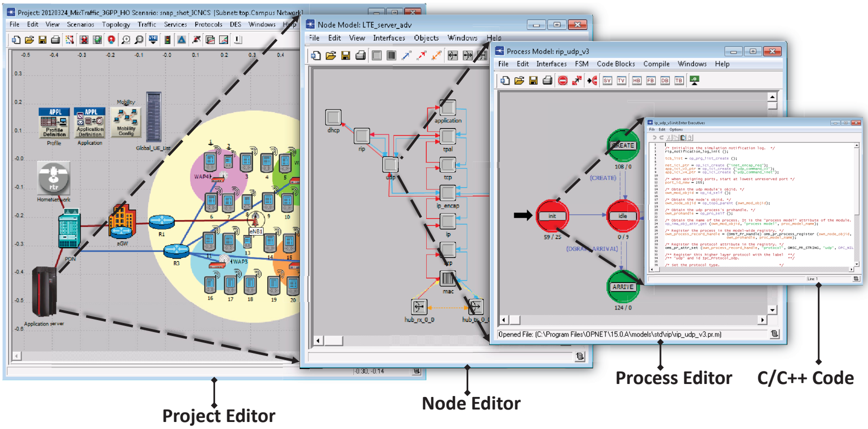
OPNET Modeler is a commercial product that provides network modeling and simulation tools. An advanced user friendly graphical interface, objected oriented and hierarchical modeling approach, and comprehensive tools for analysis and debugging are among the salient features of the OPNET Modeler. It also provides a programming interface to integrate external object files, 3rd party libraries, and other simulation tools like MATLAB etc. With the passage of time, the OPNET Modeler has evolved to incorporate more and more features in order to support new protocols, devices, and applications. Its model library consists of hundreds of protocols and vendor device models which can work together following the plug-and-play principle.

The user interface of OPNET Modeler consists of several types of editors. For example, the Project Editor is used to construct the topology of a communication network model, configure network entities, run the simulations, and view the results. The Node Editor creates the internal structure of the model of network entities (e.g., computers, routers, base stations etc.), specifies the configuration attributes as well as available statistics. The device models, which are called node models in OPNET terminology, can be instantiated as node objects in a network domain. A node model may consist of several types of modules. Each module represents some particular functions of the node's operation and they can be active concurrently. The modules within a node model may communicate with each other with the help of several types of connections, e.g., Packet Streams, Statistic Wire, or Logical Association.

At node level the internal description or functionality of a module is not visible. The behavior of these modules can be specified using the Process Editor by going one step down in the hierarchy. The modeling of a module inside a node is performed using process models which are instantiated as independent processes that execute both general communications and data processing functions. They can represent functionality that would be implemented both in hardware and in software. For example, modeling of a protocol stack inside a node can be carried out with the help of several process models each representing the functionality of one protocol layer.

A process model inside the Process Editor is seen as a finite state machine using state transition diagrams to model a certain behavior. The states of the process and the transitions between them are depicted as graphical objects. Each finite state contains C language code to express processing tasks which are performed immediately after entering this state, or immediately before leaving the state. Figure 3.1 provides a graphical overview of the aforementioned editors in the OPNET simulator.

In addition, there are also a number of other editors in the OPNET Modeler to create communication links, packet formats, interface control information etc. Further information about the OPNET Modeler can be obtained using product documentation which comes along with the software installation package.


Project: 20120324_MixTraffic_3GPP_HO Scenario: snap_shot_ICNCS [Subnet: top.Campus Network]
File Edit View Scenarios Topology Traffic Services Protocols DES Windows Help
Profile Definition
Application Definition
Mobility Config
Global_UE_List
HomeNetwork
PDN
aGW
R1
R3
eNB1
WAP4
WAP3
Application server
Node Model: LTE_server_adv
File Edit View Interfaces Objects Windows Help
dhcp
rip
udp
application
tpal
tcp
ip_encap
ip
mac
hub_rx_0_0
hub_tx_0_0
Process Model: rip_udp_v3
File Edit Interfaces FSM Code Blocks Compile Windows Help
CREATE
init
idle
ARRIVE
(CREATE)
(ARRIVAL)
Opened File: (C:\Program Files\OPNET\15.0.A\models\std\rip\rip_udp_v3.pr.m)
ip_udp_v3.init.Enter Executives
File Edit Options
Line: 1
Project Editor
Node Editor
Process Editor
C/C++ Code

```
/* Initialize the simulation notification log.  */
rip_notification_log_init ();

tcb_list = op_prg_list_create ();

net_ici_ptr = op_ici_create ("inet_encap_req");
app_ici_v3_ptr = op_ici_create ("udp_command_v3");
app_ici_v4_ptr = op_ici_create ("udp_command_inet");

/* when assigning ports, start at lowest unreserved port */
port_id_new = 255;

/* Obtain the udp module's objid. */
own_mod_objid = op_id_self ();

/* Obtain the node's objid. */
own_node_objid = op_topo_parent (own_mod_objid);

/* Obtain the udp process's prohandle. */
own_prohandle = op_pro_self ();

/* Obtain the name of the process. It is the "process model" attribute of the module.
op_ima_obj_attr_get (own_mod_objid, "process model", proc_model_name);

/* Register the process in the model-wide registry. */
own_process_record_handle = (OmsT_Pr_Handle) oms_pr_process_register (own_node_objid,
                            own_prohandle, proc_model_name);

/* Register the protocol attribute in the registry. */
oms_pr_attr_set (own_process_record_handle, "protocol", OMSC_PR_STRING, "udp", OPC_NIL

/** Register this higher layer protocol with the label   **/
/** "udp" and id IpC_Protocol_Udp.                        **/

/* Set the protocol type.                          */
```

Figure 3.1: OPNET Modeler editors.

3.3 Simulation Framework

The main objective behind the development of this simulation model is to analyze, evaluate and study several aspects of the heterogeneous wireless networks which include end user application performance, air interface performance, and the transport network performance. To be able to achieve the above targets, two types of network access technologies are selected; LTE from 3GPP networks and IEEE 802.11a from non-3GPP networks. Selection of IEEE 802.11a standard is just for the sake of an example and the discussion holds for other WLAN standards too, like, IEEE 802.11b/g/n etc. Integration of these two access technologies is performed following 3GPP proposal [3GP11a] of integrating non-3GPP access technologies to existing 3GPP networks. This proposal has been discussed in Section 2.7.1 and shown in Figure 2.26.

The following section explains the steps of developing a heterogeneous network simulator according to 3GPP specifications, where two types of access technologies coexist. The integration of the two network types is performed using the S2c interface of the Packet Data Network Gateway (PDN-GW) as shown in Figure 2.26. Figure 3.2 shows the reference architecture for the target simulation environment where two access networks namely LTE and WLAN are connected to a common core network of the operator.

Though the 3GPP SAE architecture enables mobile users to roam seamlessly

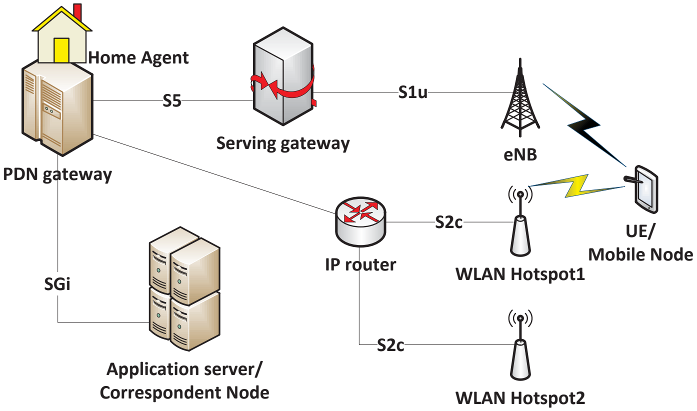


Figure 3.2: Reference architecture for simulation model.

between 3GPP and non-3GPP access technologies, it does not support the user multi-homing, i.e., simultaneous user connection to more than one access network. In order to investigate the achievable advantages through the support of user multi-homing, the existing SAE architecture is extended by adopting some of the IETF (Internet Engineering Task Force) standards. A brief overview of these extensions and the process to incorporate them in the SAE architecture is also outlined in the following section.

3.4 Simulation Model

The simulation model of the LTE access technology used in this study has been developed in a collaboration with other colleagues [Wee11] [Zak12]. It includes the detailed modeling of E-UTRAN and Evolved Packet Core (EPC) with particular focus on the important features and functionalities of the nodes and protocols. Figure 2.7 shows the LTE user-plane protocol structure which has been followed to develop the LTE simulator. The protocols are categorized into three groups: radio (Uu), transport, and user terminal protocols. The radio (Uu) protocols include peer to peer protocols such as PDCP, RLC, MAC and PHY between the UE entity and the eNode-B entity. The PDCP, RLC and MAC (including the air interface scheduler) layers are modeled in detail according to the 3GPP specifications. However, as the PHY (physical) layer is not the focus of this study, therefore, it is modeled in full detail. Nevertheless, the effect of the radio channels and PHY characteristics are modeled at the MAC layer in terms of user data rate performance.

The SAE (or LTE transport) network is based on all-IP technology, i.e., no ATM links. The user-plane transport protocols are implemented for both the S1 interface (i.e., the interface between the eNode-B and the S-GW) and the X2 interface (i.e., the interface between the eNode-Bs) according to 3GPP specification. The S1/X2 interface mainly includes the user-plane protocols such as GTP, UDP, IP and layer 2 (L2) protocols. Gigabit Ethernet is used as the layer 2 protocol in this simulation model. Moreover, UE mobility has been modeled using simple mobility models such as the random directional and random way point models.

The simulation model for IEEE 802.11 access technologies is readily available from the OPNET Standard Model Library. The only shortcoming of this model is the inability to adapt the PHY data rate according to user channel conditions. This issue will be addressed further later in this section. Some of the device models required to realize the reference architecture are also imported from the OPNET Standard Model Library without any modification, e.g., WLAN access point, application server, Ethernet links, IP Routers, and the Home Agent (HA). However, large scale modifications were needed in the models of the following network entities in order to carry out the integration of the two access technologies and realize multi-homing support for users:

- Mobile node or User Equipment (UE),
- e-NodeB (eNB),
- Serving Gateway (S-GW) and
- Packet Data Network Gateway (PDN-GW).

The next subsections describe and discuss the modeling details of the above mentioned network entities.

3.4.1 UE Node Model

This simulation environment provisions the UE as a multi-interface terminal, which enables users to be associated with both WLAN and LTE simultaneously. Figure 3.3 shows the protocol stack for developing such a UE node, it further indicates the respective protocols of each interface, e.g., PDCP, RLC, MAC, and PHY layers for the LTE interface, and ARP, MAC, and PHY for the WLAN interface. The OPNET UE node model shown in Figure 3.3 is actually a modified node model of the Mobile IPv6 capable mobile device from the OPNET model library. The original node model can perform network access only through WLAN. The LTE protocol stack has been added by the author to enable simultaneous access to WLAN and LTE networks. The implementation details about the LTE protocol layers are available in [Wee11] [Zak12], whereas the protocol layer entities for the WLAN access technology are used from the standard OPNET model library.

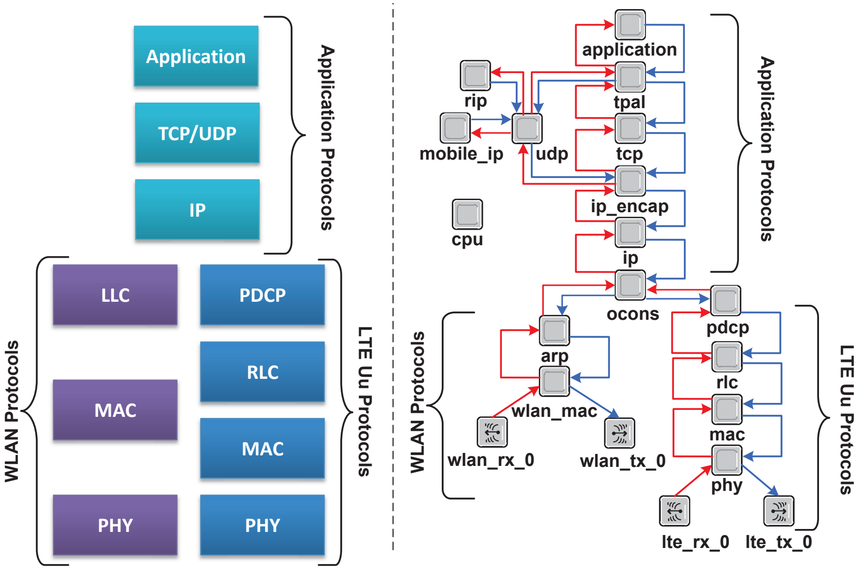


Figure 3.3: UE protocol stack (left) and OPNET UE node model (right).

The standard OPNET modeling of the WLAN MAC protocol does not consider the effect of the received signal strength on the transmission data rate of physical layer (PHY). This implies that the PHY data rate of a UE is configured as a static attribute which does not change dynamically based on a user's received signal strength. This does not correctly reflect the behavior of a user device in the real world. In order to eliminate this shortcoming, the user received signal strength is computed based on its distance to the access point and the transmission power configuration on the associated access point. This information is used in conjunction with the 'free space path loss formula' to compute the signal to interference (SIR) value at the UE. The SIR value is then mapped to a corresponding PHY data rate according to Table 3.4.1. This PHY data rate value is used to determine the transmission speed between that user and its associated access point. The PHY data rate value for each user is computed periodically after the reevaluation of received

SIR value. This allows adapting the PHY data rate of a mobile UE during the simulation time. Using this modification access point coverage is seen available within an area of approximately 100m radius.

Another deficiency encountered in the standard OPNET model library was its lack of support for Mobile IPv6 (MIPv6) capable multihomed device. In order to overcome this shortcoming, a sub-layer “OconS” is introduced between the ARP (Address Resolution Protocol) and the IP layer. One of its functionalities is to make the IP layer believe that in addition to WLAN there exists another network interface (i.e., LTE). In this way, the IP layer assigns an IPv6 address to each active network interface of the UE. The “OconS” entity is also responsible for receiving IP packets from two network interfaces and forwarding them to the IP layer during the downlink communication. While in uplink communication this entity can decide through which network interface a packet will be transmitted based on the source IP address. These decisions are part of the flow management functionality which are discussed in greater detail in Chapter 5.

Table 3.1: Minimum receiver sensitivity requirements to achieve a certain PHY data rate for 802.11a[Gas05].

PHY data rate (Mbps)	6	9	12	18	24	36	48	54
Minimum receiver sensitivity (dBm)	-82	-81	-79	-77	-74	-70	-66	-65

As stated earlier, the OPNET model library implements only the basic MIPv6 functionality. This implies that a mobile node may have several care-of addresses but only one, called the primary care-of address, can be registered with its home agent and the correspondent nodes. In order to achieve multihoming, this basic support has been extended according to the IETF RFC for multiple care-of address (MCoA) registration[WDT$^+$09]. This enables the user to register the care-of addresses from all of its active network interfaces with its home agent. As a part of the heterogeneous network architecture, it has been assumed that the user never attaches to its home network, and both LTE and WLAN networks are seen as foreign networks by the user. Therefore, a user configures one IPv6 care-of address when it is in the coverage of LTE and still another care-of address is obtained when WLAN access is available. The care-of address configuration at the UE is facilitated by enabling the eNode-B and the WLAN access point to respond to the UE’s router solicitation message with an appropriate router advertisement reply. This helps the UE to perform stateless auto-configuration of an IPv6 address on the respective network interface. More information about stateless auto-configuration of an IPv6 address can be accessed from [TNJ07].

Though above mentioned MCoA extension enables a user to register multiple care-of addresses with its home agent, the user cannot communicate over the two network interfaces simultaneously. This is because MCoA recommends using only that single care-of address which has been most recently registered/refreshed. This calls for the need of another MIPv6 extension namely Flow Binding Support [G. 10] that permits the UE to bind one or more traffic flows to a care-of address. A traffic flow, in this extension, is defined as a set of IP packets matching a traffic selector [TGSM11]. Traffic selectors help identify the flow to which a particular packet belongs through the matching of the source and destination IP addresses, transport protocol number, the source and destination port numbers and other fields in IP and higher-layer headers. Traffic selector information is carried as a sub-option inside the new mobility option “Flow Identification Mobility Option” introduced by the flow binding support extension.

The implementation of the aforementioned MIPv6 extensions has been incorporated through a patch to the “mobile ip” protocol entity in the UE and home agent OPNET node models. In this way, a user is enabled not only to use all of its active network interfaces simultaneously but also to assign a certain traffic flow to the desired network interface. Depending upon the scenario configuration user traffic flows can either be switched from one network interface to another or one traffic flow can be split into two sub-flows carried to the user over the two network interfaces. This will be discussed in greater detail in section 5.2.

3.4.2 eNode-B Node Model

Figure 3.4 shows the eNode-B architecture as implemented in the OPNET simulator. The eNode-B has two interfaces, (i) the LTE Uu interface which provides wireless network access to the users and (ii) an Ethernet based network interface towards the serving gateway which resides in the core network of the operator. Two static IPv6 addresses are configured on these two network interfaces of the eNode-B. As a part of the realization of the multihoming functionality, the LTE Uu network interface not only responds to user router solicitation messages but also transmits the router advertisement messages periodically.

The eNode-B operates two protocol stacks, one for each of its network interfaces. The LTE protocol stack implements PDCP, RLC, MAC and PHY layers according to the 3GPP specification. In the transport protocol stack, three protocol layers, i.e., UDP, IP, and Ethernet are taken from the OPNET standard model library. However, the GTP-U entity has been implemented by the author to es-

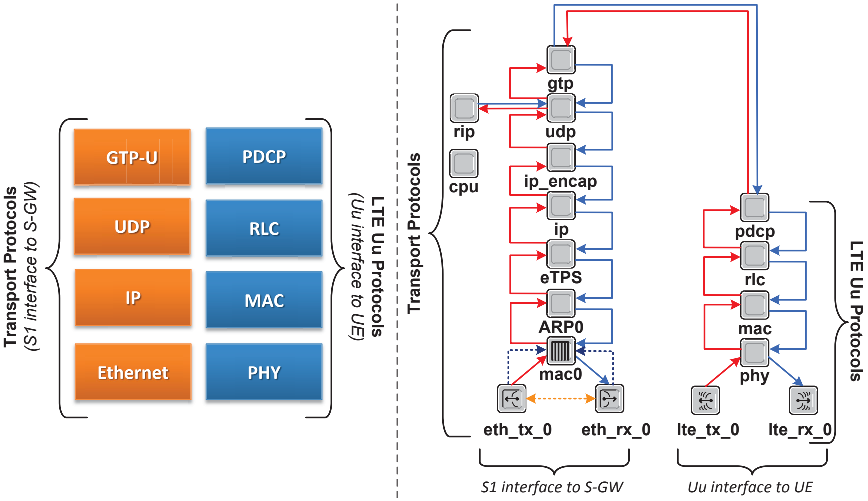


Figure 3.4: eNB protocol stack (left) and OPNET eNB node model (right).

tablish the GTP tunnel over UDP between the eNode-B and the serving gateway following 3GPP recommendations [3GP08].

In uplink communication, an IP packet from the PDCP protocol entity is encapsulated in GTP and is sent to the UDP layer. At the UDP layer, the GTP packet acts as the payload of a UDP datagram which is forwarded to the IP layer. The IP packets carrying these UDP datagrams are then routed in the transport network up to the serving gateway. In downlink communication, the GTP-U receives the GTP packets from the UDP layer. After processing the GTP headers, the de-capsulated IP packets are forwarded to the PDCP layer.

3.4.3 Serving Gateway (S-GW) Node Model

According to the 3GPP specifications, the serving gateway is mainly responsible for creation, deletion, and modification of bearers for individual users connected to the Evolved Packet System (EPS). These functions are performed based on per PDN connections for each UE. In this way, the S-GW provides the local anchor functionality for a single terminal for all of its bearers and manages them towards the PDN gateway.

The S-GW has two network interfaces, (i) S1 interface towards the eNode-B and, (ii) S5 interface towards the PDN gateway. From a traffic routing viewpoint,

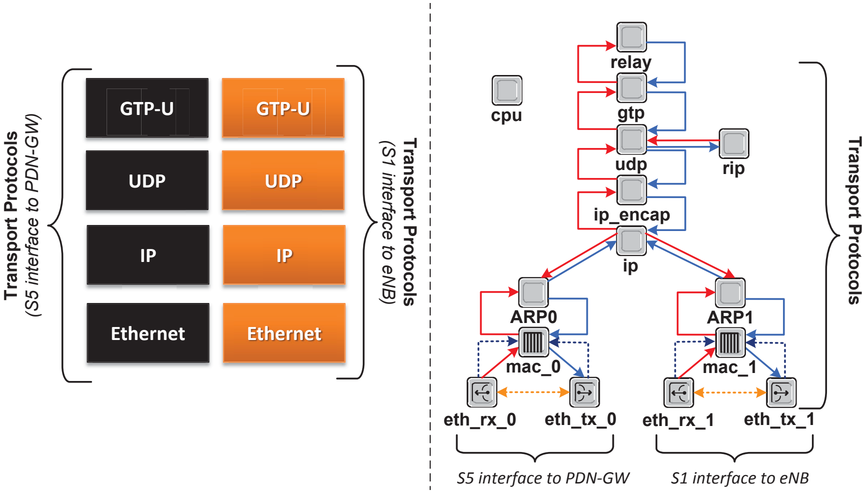


Figure 3.5: Serving gateway protocol stack (left) and the OPNET node model (right).

the S-GW serves as an end-point for two GTP tunnels; one from the eNode-B and the other from the PDN gateway. The "relay" entity in the OPNET node model (see Figure 3.5) of the S-GW receives IP packets from the PDN gateway (downlink traffic) and eNode-B (uplink traffic). The received packets are then forwarded to the respective GTP tunnel based on the destination address of the packet (which is either the UE care-of address or the application server address).

3.4.4 PDN Gateway (PDN-GW) Node Model

Figure 3.6 depicts the user plane transport protocols of the PDN gateway (PDN-GW) and its implementation as a node model in OPNET. The PDN-GW is a central entity for connecting the external IP networks through the SGi interface, 3GPP networks through S5, and non-3GPP networks through interface S2a/b/c. However, within the scope of this implementation, only the trusted non-3GPP networks with host-based mobility are considered. As evident from Figure 3.6 (right side), other than the aforementioned three interfaces, there exists still another interface which is linked with the home network of the UE. As the UE is always associated with LTE/WLAN networks and never attaches with its home network, it implies that the UE is always in a foreign network.

The "OconS" entity in the OPNET node model of the PDN-GW plays an impor-

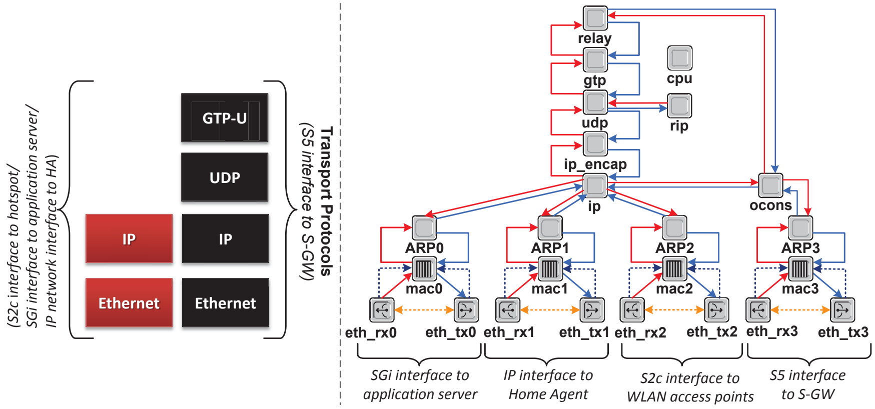


Figure 3.6: PDN gateway protocol stack (left) and OPNET PDN gateway node model (right).

tant role in proper routing of the traffic according to protocol stack requirements of each network interface. In downlink direction, IP packets coming from the home agent are destined to one of UE's two care-of addresses. This destination address determines whether this packet will be delivered to the user over the LTE or WLAN access network. If the destination address belongs to the LTE network, the IP protocol entity forwards it to the "OconS" entity. This packet is then forwarded to the "relay" entity which encapsulates it in the GTP tunnel with destination end-point at S-GW. These encapsulated packets are then sent by UDP to the IP protocol entity. The IP protocol entity sees these encapsulated packets destined to the S-GW and forwards them to the "OconS" entity which lets them leave the node through the S5 interface towards S-GW.

Similarly in uplink direction, the user traffic coming from S-GW reach the IP protocol entity via the "OconS" entity. Owing to the fact that the PDN-GW is the end-point of the GTP tunnel, these packets are forwarded to the upper layers after the packet header removal process. Eventually, these packets reach the "relay" entity in the form of original IP packets as sent by the UE. These packets have the home agent as their destination address because of being encapsulated in the IP-over-IP tunnel as a part of MIPv6 functionality. The "relay" entity sends them to the "OconS" entity so that they can be forwarded to IP protocol entity for further routing to the home agent.

There is no GTP tunnel established on the S2c interface between the PDN-GW and the WLAN access points. Therefore, IPv6 packets destined to the care-of address of the UE are simply routed to the access point using the Ethernet link. Similarly, uplink traffic coming from access points is received by the IP protocol entity and routed to the home agent without any further handling.

The home agent is connected to the PDN-GW through the IP link and is the anchor point for all MIPv6 communication. As the “route optimization” option of MIPv6 is not used here, all uplink and downlink traffic routes through the home agent on its way to the destination. This is required to perform flow management as explained in Section 5.2. In downlink direction, the traffic from the application server or correspondent node (CN) is received by PDN-GW at the SGi interface. This traffic is destined to the home addresses of the respective UEs and therefore routed to home agent. The home agent performs IPv6 encapsulation of these packets in order to tunnel them to a care-of address of the UE. The encapsulated packets are then sent to the PDN-GW from where they are routed to the UE through either LTE or WLAN as explained above. In uplink direction, UE sent IPv6 encapsulated traffic reaches the home agent via the PDN-GW. These packets are de-capsulated by the home agent and then forwarded to the PDN from where they are routed to the application server.

3.4.5 Sub-flow Aggregation

In multi-path communication packets may arrive out of order at the destination [EM02]. Real time applications usually deploy a play-out (or de-jitter) buffer which is mainly intended to eliminate the jitter associated with packet delays. However, it also performs packet reordering if the packets arrive within the time window of play-out buffer length. In this way, real time applications face no problem when receiving out of order packets in multi-path communication. This holds as long as delay of each involved network paths is less than the play-out buffer length $\gamma_{\text{de-jitter buffer}}$.

On the other hand, TCP based applications are very sensitive to packet reordering. This is because an out-of-sequence packet can lead TCP to overestimate the congestion of the network. This, in turn, causes a substantial degradation in application throughput and network performance [LG02]. A literature survey shows that there are several proposals to make TCP robust against packet re-ordering, e.g., [FMMP00], [LG02], [S. 03], [FY02], [M. 03], [LK00]. However, the analysis and implementation of such schemes are currently not within the focus of this research work. Instead a simple TCP re-ordering buffer is implemented at the receiver which is very similar in functionality to a play-out buffer. Simulation analysis shows that the re-ordering buffer length of such a play-out buffer must be less than the TCP protocol timeout value. In this work, a feasible value of reordering buffer length $\gamma_{\text{tcp reorder buffer}}$ found to fall in a range from 50 to 300ms. Further details about TCP re-ordering buffers will be provided in Chapter 5.

3.5 User Traffic Models

The tasks of designing a communication network or performance optimization of an existing network involves the efforts to maximize capacity, minimize latency, and offer high reliability using limited bandwidth resources. The main factors that affect the performance of any communication network are packet end-to-end delay, packet loss, and throughput. In order to design high performance networks and guarantee user application performance the detailed analysis of the above factors is a crucial step. Often the foremost step in such an analysis is the study of the traffic demands on the network. As a consequence, the types of traffic models used to understand the flow of user traffic in the network and their ability to depict the realistic characteristics of the network bear fundamental importance. Traffic models not only enable a network designer to make assumptions about the networks being designed based on the past experience but also enable prediction of performance for future requirements. They are of paramount importance in network architecture comparisons, network resource allocations, and the performance evaluation of protocols.

This section describes the traffic models used to carry out simulation studies in this work. This includes the traffic models for Voice over IP (VoIP), File Transfer Protocol (FTP), Hypertext Transfer Protocol (HTTP), and Video Streaming applications.

3.5.1 FTP Model

The users in the OPNET simulator can use FTP applications in order to transfer a file. Two basic commands for file transfer are modeled, i.e., GET and PUT. The GET command triggers a file download from the application server to a user. The PUT command initiates file upload from a user to the application server. The model does not implement separate control and data channels. For each file transfer a new TCP connection is opened which is used to transfer the data as well as the control commands. The two main parameters of an FTP session are:

- **File size**: Size of a file being transferred in bytes.
- **Inter-request time**: Time between subsequent file requests in seconds.

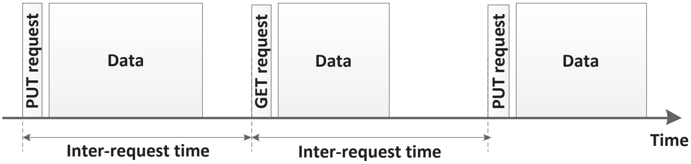


Figure 3.7: FTP application modeling.

Owing to the fact that the FTP protocol uses TCP as an underlying transport protocol, it is important to list the main modeling parameters of the TCP model.

- **TCP flavor**: New Reno with fast recovery and fast retransmit enabled.
- **Maximum segment size (MMS)**: 1300Byte. This is to avoid packet segmentation at the IP layer.
- **TCP receive window size**: 1MByte. The window scaling option is enabled.
- **Duplicate ACK threshold**: Number of consecutive duplicate ACKs after which Fast Retransmit will trigger a retransmission. The default value of 3 is used for this purpose.
- **Maximum ACK delay**: Maximum time in seconds that TCP waits after receiving a segment before sending an ACK. The default value is 200 ms.

3.5.2 HTTP Model

The HTTP application models web browsing uses TCP as a default transport protocol. A web page contains text in HTTP format as well as embedded image/video objects. When a user downloads a web page from an application server, it results in opening multiple TCP connections for downloading the inline objects embedded in the page. Once the whole page is downloaded, the user needs time to read the page, before the next request is made. This time is called user reading time. The model can follow HTTP version 1.0 & 1.1 specifications with the important parameters listed below:

- **Main object size**: Size of web page in bytes excluding inline objects.
- **Embedded objects size**: Size of the object in bytes embedded in a page.
- **Number of embedded objects**: Number of inline objects contained in a page.

- **Reading time**: Time between end of a page download and start of the next page download.

The default values of the TCP model parameters described above under FTP model also hold for HTTP model.

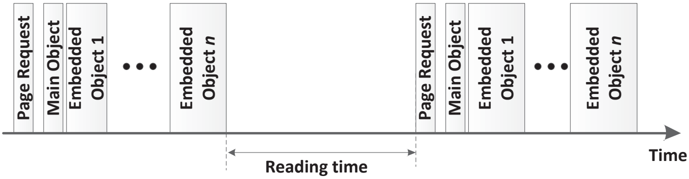


Figure 3.8: HTTP application modeling.

3.5.3 VoIP Model

The VoIP application models a virtual channel between a user and the application server over which digitally encoded voice signals are transported. The voice data can be encoded using one of several supported encoding schemes. VoIP packets are transported over the UDP protocol using the Real-Time Protocol (RTP). The voice communication is modeled using ON/OFF periods. An ON period models a voice spurt followed by an OFF period representing the silence period.

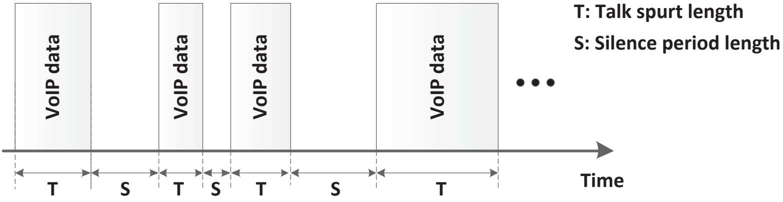


Figure 3.9: VoIP application modeling.

The VoIP model supports narrowband as well as wideband codecs. The main difference between the two codec types is the sampling rate of the voice signal. In narrowband encoding the voice signal is sampled at 8kHz, resulting in an effective voice pass-band of about 200 to 3300Hz. Wideband voice codecs offer double the sample rate, providing an effective pass-band of 50 to 7000Hz. As a result, wideband codec can achieve higher voice call fidelity at the expense of computational processing power.

The important parameters of the VoIP model are as follows:

- **Silence and talk spurt lengths**
- **Encoder scheme**: The supported codecs include G.711 [IT88], G.729a [IT12], GSM EFR [72699] and G.722.2 [IT03b]
- **Voice frames per packet**: Number of encoded voice frames packed in a single RTP/UDP packet.
- **Compression delay and decompression delay**: Specifies the delays in compressing a voice packet at the source and decompressing the voice packet at the destination.
- **De-jitter buffer size**: It is also referred to as play-out delay. This is the amount of time a packet could be delayed at the destination in order to minimize packet delay variations. The default value of the static de-jitter buffer size is taken as 50ms.

3.5.4 Video Model

The popular video codecs rely on complex algorithms in order to provide better visual quality while keeping the demands of hardware processing power and bandwidth resources to a minimum level. This makes it extremely difficult, if not impossible, to model video streaming traffic using stochastic processes. That's why in this work, a trace from video communication over a real IP network is used to generate video traffic in simulations. For this purpose, a reference video clip encoded with the MPEG-4 codec with a certain bit rate and resolution is selected. The video clip is then transmitted over the IP network using a video streaming software (like VLC media player [Pla13]) and the IP packets are captured using a packet sniffing tool like Tcpdump [Ana13]. This helps generate a trace file containing the size of all transmitted packets and their inter-arrival time. Figure 3.10 shows a graphical representation of such a trace file.

In the OPNET simulation environment, this trace file can be used to generate video traffic with the same characteristics as the reference video streaming traffic in the real IP network. These video packets are transported to the receiver over the UDP protocol. Table 3.2 lists two video applications used in this work.

3.6 Statistical Analysis of Simulation Output Data

In order to study the behavior of a simulated system properly and draw valid conclusions about its performance an appropriate analysis of simulation output data

Application name	Codec	Resolution	Mean bit rate	Frame rate
Skype video	MPEG-4	640x480	512kbps	30 fps
Live News video	MPEG-4	720x480	1Mbps	30 fps

Table 3.2: Video application models.

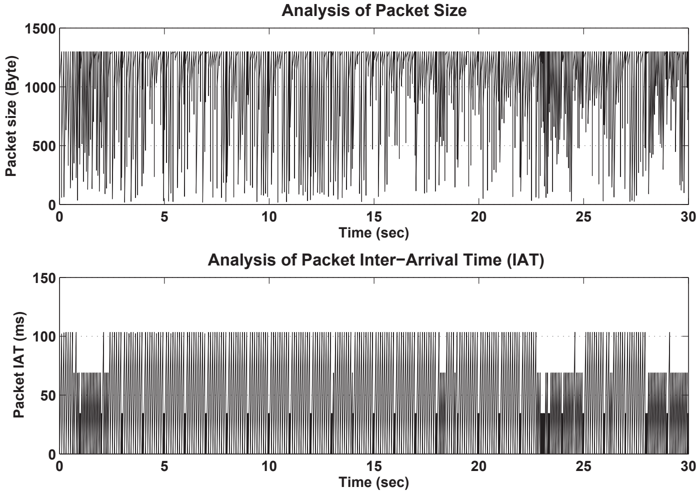


Figure 3.10: Packet size and packet inter-arrival time analysis of 30 second long 'Skype video' clip.

is of utmost importance. This is because typically a simulation model is driven through the time using random samples from probability distributions. Therefore, runs of the simulation yield the estimates of system performance which are themselves random variables and, hence, subject to sampling error. This leads to a situation where a significant probability exists that erroneous inferences may be made about the system under investigation. Such a situation can be avoided by using state-of-the-art techniques to statistically analyze simulation output data as described in this section.

3.6.1 Confidence Interval

Let's assume $X_1, X_2, X_3, ..., X_n$ are the independent identically distributed random variables with finite population mean μ and finite population variance σ^2. The goal here is to estimate μ and σ^2. An estimator $\hat{\theta}$ is unbiased for the parameter θ if $E(\hat{\theta}) = \theta$. In our case sample mean $\overline{X}(n)$ of the random variables is an unbiased point estimator for μ so that

$$\overline{X}(n) = \frac{1}{n}\sum_{j=1}^{n} X_j$$

$$s^2(n) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}(n))^2$$

where s^2 is sample variance and an unbiased estimator for σ^2 [AM07]. Now the problem is to assess how close the estimator $\overline{X}(n)$ is to μ. This is because $\overline{X}(n)$ has a certain value of variance Var[$\overline{X}(n)$] which could lead to a situation that on one simulation run $\overline{X}(n)$ may be close to μ while on another it may have large difference from μ. This problem can be resolved if a confidence interval is computed for μ which helps assess the accuracy of $\overline{X}(n)$ as an estimator of μ. In other words, confidence interval estimation quantifies the confidence (probability) that μ falls within an interval whose boundaries are calculated using appropriate point estimates.

The first step in computing a confidence interval is to estimate Var[$\overline{X}(n)$]. Considering the fact that X_j's are independent and identically distributed Var[$\overline{X}(n)$]=σ^2/n which leads to an unbiased estimator of Var[$\overline{X}(n)$] as

$$\widehat{Var}[\overline{X}(n)] = \frac{s^2(n)}{n}$$

Now if the X_j's are normally distributed random variables then a $100(1-\alpha)\%$ confidence interval for μ is as follows

$$\overline{X}(n) \pm t_{n-1,1-\alpha/2}\sqrt{\frac{s^2(n)}{n}}$$

where $t_{n-1,1-\alpha/2}$ is the upper $1-\alpha/2$ critical value for a t-distribution [Sta13] with $n-1$ degrees of freedom . Hence it can be claimed that the above computed confidence interval contains μ with the probability $(1-\alpha)$.

If the X_j's are not normally distributed, the actual coverage of a confidence interval can be less than the desired coverage $1-\alpha$. However, the central limit theorem guarantees that if the sample size n is sufficiently large then the actual coverage will be close to $1-\alpha$ [Law83].

3.6.2 Types of Simulations

There are two types of simulations with respect to output analysis:

1. **Terminating simulations**: A terminating simulation model starts in a specific state and is run until a termination point T_E when a predefined event E occurs. T_E may be a random variable itself. Such models are used to study short time system dynamics within a system's natural time horizon. The output process is not expected to achieve a steady state behavior. Therefore, any measure of performance estimated from output data explicitly depends on the initial system state. In general, the initial conditions in these simulations are set to represent the initial conditions for the corresponding system. An example is the simulation of a computer network starting from idle state until n jobs are served.

2. **Steady-state simulations**: A steady-state simulation model has no natural event E for termination and could be potentially run forever. Such models are used to study the long-term behavior of the system. A performance measure of a system is called a steady-state parameter if it is a characteristic of the equilibrium distribution of an output stochastic process [LK91]. The value of a steady-state parameter does not depend upon the initial conditions. An example is the simulation of a continuously operating IP network router where the objective is the estimation of mean queue length or mean queuing delay experienced by the packets. The simulation studies presented in this work follow steady-state simulation model.

Though in steady-state simulations long-run system behavior is of particular interest, initial system conditions may exert bias on long-term system statistics. Two most common ways to eliminate biasing effect are as follow

(a) To start statistics collection after an initial period of system warm-up, namely, after the biasing effect of the initial conditions decays substantially. The problem in this case is to determine an accurate system warm-up time until when simulation output data should be truncated. If the output data is truncated too early, a significant bias could still exist in the remaining data. On the other hand, truncating it too late could lead to the waste of valid observations. In literature several procedures have been proposed to determine an adequate length of system warm-up time [LK91] [CS92] [Fis72] [GAM78] [Wel81].

(b) To run a simulation for a very long time so that the biasing effect becomes imperceptible. This is a rather simple way to control biasing effect which may yield point estimators with lower mean squared error compared to the estimators obtained through above described method [Fis72]. The only problem with this approach is that an excessive simulation run might be required before which biasing effects are rendered negligible.

3.6.3 Steady-State Analysis

A number of methods have been developed to estimate steady-state system parameters which include Batch Means, Independent Replications, Standardized Time Series, Spectral Analysis, and Regeneration Cycles. Keeping the primary focus on the estimation of steady-state mean value of a discrete-time simulation output process, the most frequently used methodologies are Batch Means and Independent Replications as discussed in the following [Ban98].

3.6.3.1 Batch Means

The simplicity of computation and effectiveness of the batch means method makes it a popular way to estimate the steady-state mean and variance. The method is based on the idea of splitting one long simulation run into a number of contiguous non-overlapping batches and using sample means of these batches to produce point or interval estimators. As an example, consider a vector of observations $(X_1, X_2, X_3, \dots .X_n)$ of a system parameter obtained from an enormously long simulation run. Now splitting this vector into non-overlapping contiguous batches each of length m would generate the following b batches so that $n = b \cdot m$

Batch 1: $X_1, X_2, X_3, \dots X_m$
Batch 2: $X_{m+1}, X_{m+2}, X_{m+3}, \dots X_{2m}$
Batch 3: $X_{2m+1}, X_{2m+2}, X_{2m+3}, \dots X_{3m}$
$\vdots$
Batch b: $X_{(b-1)m+1}, X_{(b-1)m+2}, X_{(b-1)m+3}, \dots X_{bm}$

for $i = 1, 2, 3, \dots b$, the i^{th} batch mean is given by

$$Y_i = \frac{1}{m} \sum_{j=1}^{b} X_{(i-1)m+j}$$

Similarly the batch means estimator for variance is calculated as follows

$$\text{Var}[Y_i] = \frac{1}{b-1} \sum_{i=1}^{b} (Y_i - \bar{Y}_b)^2$$

where $\bar{Y}_b$ is the grand sample mean, i.e.,

$$\bar{Y}_b = \frac{1}{b}\sum_{i=1}^{b} Y_i$$

The only problem in applying the batch means method in practice is the right choice of batch size m. If m is small, the batch means Y_i can have high correlation. Alternatively making m of very large results in only a few batches b and potential problems with the application of central limit theorem to compute the confidence interval. A number of proposals have been made to compute the smallest batch size which can pass the test of statistical independence, e.g., see [Fis72] [SA97].

3.6.3.2 Independent Replications

The problems of possible correlation among the batch means can be avoided by the method of independent replications. In this method, M independent replications of the system simulation are run. Each replication starts in the same state and uses a portion of random number stream that is different from the portions used to run the other replications. Assuming $X_{i1}, X_{i2}, X_{i3}, ..., X_{iN}$, as the output data obtained from replication i, the sample means are given by

$$Y_i = \frac{1}{N}\sum_{j=1}^{N} X_{ij} \qquad i = 1, 2, 3, ...M$$

In this way M approximately independent sample means can be obtained whose estimator for variance is calculated as shown below

$$\mathrm{Var}[Y_i] = \frac{1}{M-1}\sum_{i=1}^{M} (Y_i - \bar{Y}_M)^2$$

where the grand sample mean $\bar{Y}_M$ is given by

$$\bar{Y}_M = \frac{1}{M}\sum_{i=1}^{M} Y_i$$

Owing to the fact that each replication should be started properly, the biasing effects must be handled carefully as explained earlier in this section.

The estimators for the variance and sample means obtained from batch means or the independent replication method can be used to compute the confidence interval as discussed in the beginning of this section.

4 LTE Access Interface Enhancements

The OFDMA based air interface of LTE delivers the flexibility and increased spectral efficiency required by the next generation of high-speed, all-IP mobile networks. The air interface scheduler is, therefore, a key component of the LTE access technology which intelligently schedules the radio resources to deliver required QoS to the active radio bearers. The algorithms employed by the air interface scheduler for this purpose have a significant impact on the performance of the individual base station and overall LTE radio access network. These algorithms have not been standardized by 3GPP, leaving an opportunity for vendors to craft them according to the requirements of specific deployment and usage scenarios. Though numerous MAC scheduling schemes have already been proposed in scientific literature, this work highlights two important performance optimization aspects of MAC scheduling which have not been well addressed by the research community. Section 4.1 and 4.2 of this chapter are dedicated to these enhancements of the MAC scheduler.

A detailed discussion about the LTE protocol stack has been made in Chapter 2. It has already been explained that the PDCP, RLC, and MAC layers together constitute layer-2 of the LTE air interface. The RLC entity resides between the PDCP layer and the MAC sub-layer in the LTE protocol stack. It reformats PDCP PDUs, referred to as segmentation and concatenation process, to fit the size required by the MAC layer transport block. The size of transport block is determined by the MAC scheduler considering the bandwidth requirements, distance, power requirements, modulation scheme, and type of application. Owing to the fact that RLC buffers are of limited capacity, the user data is mainly held in PDCP buffers and delivered to RLC on demand. In this way, packet queues at PDCP layer act as the main buffer for LTE air interface. An analysis of MAC scheduler operation indicates that a large performance gain may be realized by proper management of buffers at the PDCP layer. In section 4.3 of this chapter some simple buffer management techniques are discussed as candidates for deployment at the PDCP layer.

4.1 Coordinated Uplink Radio Interface Scheduling

In LTE the bandwidth demands of time varying user traffic cannot always be fulfilled by the transport networks either due to their limited capacity or because of the inaccuracies involved in the bandwidth dimensioning process. This often creates a congestion situation in the transport network which could substantially degrade the system performance. In this work, a novel congestion control scheme is introduced for the LTE uplink which functions based on the coordination between transport network interface and LTE radio interface capacities. The proposed mechanism preferably operates at the eNode-B where two network interfaces are monitored in order to efficiently minimize the congestion situations.

The overall performance of a wireless access network is found to be particularly sensitive to the uplink congestion in the transport network. This is because a congested transport uplink leads to excessive packet delays and packet losses which have to be recovered by the higher layer protocols (e.g., TCP) through retransmissions. If the congestion is not properly controlled, it results not only in poor user QoE but also in the wastage of radio resources and the UE's limited battery power consumed in transmissions over the air interface. To mitigate the adverse effects of transport congestion, this work introduces a novel mechanism, termed back-pressure coordination. The back-pressure manager operates between the radio and transport schedulers and manages the resources of the both networks. In order to avoid the congestion, the back-pressure manager sanctions only a sufficient amount of traffic from the radio interface which can be carried over the bandwidth limited transport links. This strategy brings gain in twofold manner: first, it circumvents the congestion at transport uplink and helps achieve the end-to-end target QoS. Second, it saves the UE power and radio resources which otherwise would be wasted in retransmitting packets dropped due to congestion in the transport network. In order to avoid transport network congestion, the back-pressure manager manages the resources of traffic only from non-GBR bearers. Whereas, GBR bearer traffic is always provided with the required resources so that its strict QoS demands can be fulfilled.

An overview of the uplink back-pressure coordination scheme is given in Figure 4.1. The figure shows three main components: a set of uplink MAC schedulers for three cells, an uplink transport scheduler with DiffServ capabilities, and a back-pressure manager unit which coordinates between the two schedulers. The back-pressure manager processes input signals from both schedulers and adaptively estimates the maximum allowed data rate per cell over the radio interface. In order to achieve the objective of simulating the most realistic behavior of networks, the processing/transmission delays of different entities are taken into account. The

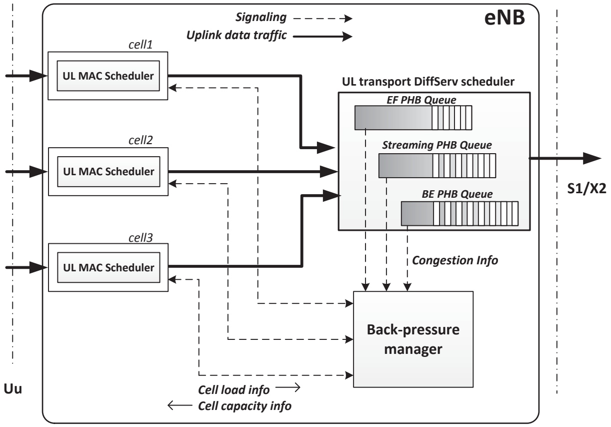


Figure 4.1: Overview of coordinated uplink radio scheduling.

individual delay components considered for this study are listed below and also shown in Figure 4.2.

- **d1**: Signaling & processing delay of conveying the congestion indication / release event information from the transport scheduler to the back-pressure manager. In addition to these triggers, the transport scheduler also sends the other information, e.g., the transport link capacity available for non-GBR traffic. The mean value of this delay component is assumed to be 5ms based on the observations in real world networks.
- **d2:** Signaling & processing delay of requesting "TD Priority Term (described later)" and "individual cell throughput" from MAC scheduler by the back-pressure manger. The mean value of d2 set as 8ms considering the measurements in real world networks.
- **d3:** Signaling & processing delay of sending TD priority term and cell throughput info to back-pressure manager by the MAC scheduler which has a mean value of 10ms as experienced in real world networks.
- **d4:** Signaling & processing delay of sending allocated cell throughput information to the MAC scheduler by the back-pressure manager. It also includes delay for sending updated MAC grants to the UEs by the MAC scheduler, as well as, the UE processing delays. The mean value of this delay element is considered as 12ms based on real world network experiments.

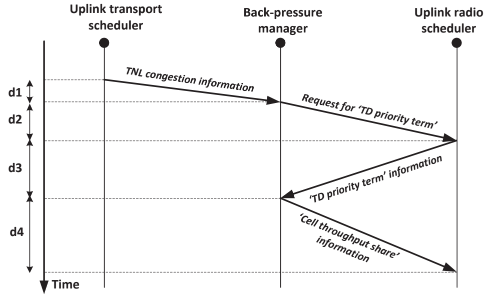

Uplink transport scheduler
Back-pressure manager
Uplink radio scheduler
TNL congestion information
Request for 'TD priority term'
'TD priority term' information
'Cell throughput share' information
d1
d2
d3
d4
Time

Figure 4.2: Exchange of control signaling by the back-pressure manager.

4.1.1 DiffServ Scheduler Implementation

The structure of the DiffServ transport scheduler is shown in Figure 4.3. The traffic data coming from the radio interface for different services is identified and classified into the following PHB queues; (i) conversational voice traffic is directed to 'Voice (GBR)' PHB, (ii) real time streaming video traffic is handed over to 'Streaming (GBR)' PHB queue, and (iii) rest of the traffic, e.g., FTP, HTTP etc. is received by 'non-GBR' or 'Best Effort (BE)' PHB queue. Voice and streaming video traffic belongs to the GBR traffic class and, therefore, has a higher priority over the non-GBR traffic. Following the the common practice of traffic prioritization, the conversational voice traffic is attributed with a strict priority and assigned to the EF PHB queue. Moreover, the streaming user traffic is assigned to any PHB queue with higher priority than that of the non-GBR user traffic. However, any other traffic prioritization scheme can also be configured depending upon the specific service requirements.

Due to service prioritization, mainly the lowest priority non-GBR or best effort (BE) traffic suffers from the bandwidth limitation imposed by the transport network. Often, when the transport uplink gets congested, the buffer occupancy of BE PHB queue grows rapidly leading to packet drops. Therefore, the transport congestion in uplink direction can be easily identified by monitoring the buffer occupancy of BE PHB queue. For this purpose, two suitable threshold values of buffer occupancy are configured which help identify congestion indication and congestion release events. For example, when the buffer filling level surpasses the upper threshold, it implies a congestion situation and when it goes below the lower threshold, it gives an indication of congestion release. Although it is logical for the back-pressure manager to act only when the congestion is detected, but it can

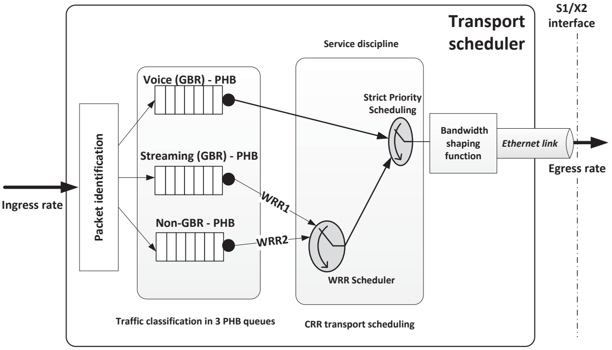


Figure 4.3: Structure of the DiffServ transport scheduler for uplink.

be also be kept operational during other times. This is because, it will not restrict the radio interface capacity until there exists an actual congestion in the transport network.

As service discipline, a Weighted Round Robin (WRR) scheduler has been used along with a strict priority scheduler for voice PHB. The Egress rate of the transport uplink scheduler can be controlled with the help of a bandwidth shaping function. However, in this study the traffic shaping rate is fixed to a pre-configured value during the whole simulation time.

4.1.2 Congestion Control Algorithm

The congestion control algorithm used by the back-pressure manager is shown in Table 4.1 in the form of pseudocode. In order to operate according to this algorithm, the back-pressure manager gathers the following pieces of information from the transport scheduler and the MAC scheduler.

a) Available capacity for non-GBR traffic

The available capacity $C^{\text{n-GBR}}$ for non-GBR bearer traffic at the transport scheduler is estimated from the measured egress rate of voice PHB E_r^{voice} and video PHB E_r^{video} as follow:

$$C^{\text{n-GBR}} = C_{S1_{UL}} - E_r^{\text{voice}} - E_r^{\text{video}} \tag{4.1}$$

where $C_{S1_{UL}}$ is uplink bandwidth of last-mile S1 link.

Depending on the offered traffic load, the S1 uplink capacity is shared between GBR and non-GBR bearers according to their assigned priority. The GBR traffic transmission often has on-off nature and uses the uplink capacity only when data is available at the PHB queues. During the other times, this capacity is used by the traffic of non-GBR bearers. Therefore, in the transport congestion situation an accurate estimation of available S1 uplink capacity for traffic of non-GBR bearers is important. To achieve this, the mean egress rate of GBR traffic is measured regularly at the transport scheduler. Using this information the capacity available to non-GBR traffic $C^{\text{n-GBR}}$ is then computed with the help of equation 4.1. The value of $C^{\text{n-GBR}}$ is sent to the back-pressure manager periodically (e.g., every 10ms) by the transport scheduler.

b) TD Priority Term

In order to calculate the throughput allocation for each cell, the back-pressure manager requires a TD priority term from the MAC scheduler. This term represents the QoS share information of the active bearers over the radio interface. This term is computed based on the type of the time domain scheduler deployed in the radio interface scheduling. The TD priority term κ_c is calculated per cell basis for non-GBR bearers and is given as follows:

For ‘Proportional Fair’ (PF) time domain scheduler

$$\kappa_c^{PF} = \sum_i \widehat{R}_i \cdot Q_{k,i} \tag{4.2}$$

And for ‘Blind Equal Throughput’ (BET) time domain scheduler

$$\kappa_c^{BET} = \sum_i Q_{k,i} \tag{4.3}$$

here i is the index of non-GBR bearers of QoS class k in a cell c. $\widehat{R}_i$ is the instantaneous achievable data rate based on the channel quality or CQI value of the corresponding UE being served with bearer i. $Q_{k,i}$ is the QoS weight which represents the relative priority of the respective bearer i of class k in the MAC scheduling function. It can be noted that κ_c^{PF} includes both the channel quality and QoS weight of the traffic class whereas the κ_c^{BET} relies only on the QoS weight.

c) Average radio cell throughput

An important piece of information required by the back-pressure manager from the radio interface side is the mean cell throughput. This throughput value is measured at the RLC layer to exclude any HARQ retransmissions of the MAC layer and, therefore, it is also called effective cell throughput. A bearer’s throughput measured at the RLC layer differs from its measurement at the transport network link (TNL). This is due to different protocol overheads at the radio and transport network interfaces. For the use of the back-pressure manager, the effective cell throughput at the radio interface must be determined from the transport point of view by considering all relevant transport protocol headers.

Owing to the fact that all radio bearers terminate at the PDCP layer and all pertaining cell information of the traffic is lost at this point, a direct measurement of the cell throughput is unfeasible at TNL. Therefore, the throughput change ρ^{OH} due to TNL protocol overhead bits has to be estimated mathematically at the RLC layer for each cell. The estimated value of ρ^{OH} is then added to the measured cell throughput value ρ at the RLC layer to get an estimation of the effective cell throughput $\widetilde{\rho}$ at TNL, i.e.,

$$\widetilde{\rho} = \rho + \rho^{\mathrm{OH}} \tag{4.4}$$

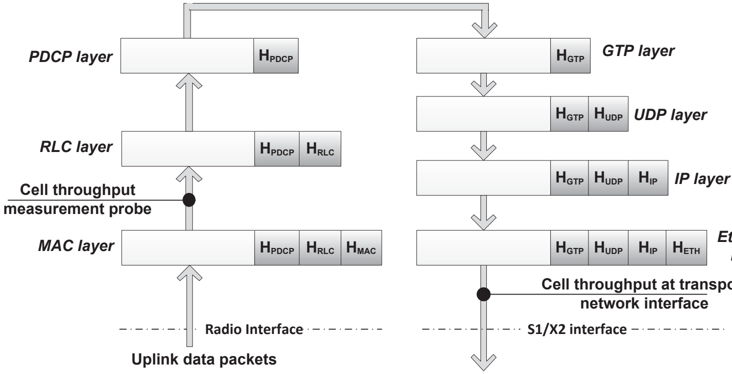


Figure 4.4: An overview of cell throughput estimation process. The protocol headers of a packet are shown for different layers of protocol stacks at the radio and transport network interfaces.

The estimation of ρ^{OH} involves the implementation of two counters for non-GBR traffic at the RLC layer. Within a certain time period T, one of the counters is used to count the number of RLC PDUs N_{RLC} received from the MAC layer and the other counter monitors the number of PDCP PDUs N_{PDCP} sent to the PDCP layer after the reassembly process at the RLC layer. Considering the size of the PDCP PDUs, N_{PDCP} consists of two components: i) N_{PDCP_L}, number of packets with a large size which cause IP fragmentation at the IP layer, and ii) N_{PDCP_S}, number of packet with small size and need not to be fragmented at the IP layer. The protocol overhead OH at the transport network interface is computed for each packet based on its size as given below:

$$OH_S = H_{GTP} + H_{UDP} + H_{IP} + H_{ETH}, \quad \text{and} \tag{4.5}$$

$$OH_L = H_{GTP} + H_{UDP} + 2 \cdot H_{IP} + 2 \cdot H_{ETH}; \tag{4.6}$$

where H with a protocol name as its subscript represents the number bits of that protocol header. Finally, the value of ρ^{OH} is computed as follows

$$\rho^{\mathrm{OH}} = \frac{N_{PDCP_L} \cdot OH_L + N_{PDCP_S} \cdot OH_S - N_{RLC} \cdot H_{RLC} - N_{PDCP} \cdot H_{PDCP}}{T} \tag{4.7}$$

Figure 4.4 elaborates the process of ρ^{OH} estimation. It shows the point where the cell throughput ρ is measured at the radio interface. It also explains how ρ differs at the transport network interface due to the protocol overheads.

Operation of the algorithm

On receiving the required information from the radio interface scheduler and transport scheduler, the back-pressure manager computes the throughput allocation for non-GBR traffic of each cell by executing the instructions shown in Table 4.1. The throughput allocation value $\overset{\smile}{\zeta}$ for a cell determines the bandwidth share of its non-GBR traffic from the limited transport link capacity. Using this information the MAC scheduler restricts the uplink radio capacity for non-GBR traffic accordingly in order to circumvent the transport congestion situations.

The algorithm listed in the execution section of Table 4.1 is rather simple. In the first line, it is specified that the basic throughput share of a cell from the available uplink capacity $C^{\text{n-GBR}}$ is proportionate to its traffic load which is indicated by its TD priority term. The rest of the instructions outline the procedure of redistributing the surplus throughput fairly. Surplus throughput ν is in-excess throughput from the basic allocated share which will be left unused by the cell. A positive value of ν implies that the measured throughput amount less than the allocated basic throughput share of the cell. Such a situation may occur if the cells of an eNB are unevenly loaded with the user traffic.

At line 3, the surplus throughput ν is computed for each cell. Positive values of ν are summed at line 4 to get the total surplus throughput of the system. Heaviside function $H(\cdot)$ is used here to ensure that the negative values are disregarded in the summation process. At line 5, TD priority terms of the needy cells are added up for later use at line 10. The FOR loop, at line 6, redistributes the overall surplus throughput $\widehat{\nu}$ among the needy cells in proportion to their traffic load or TD priority term. It may happen that the redistribution again results in some surplus throughput, therefore, the process at lines 3–12 is repeated until all surplus

Table 4.1: Pseudocode for algorithm implemented at back-pressure manager

Given	
S	A set of all cells in the eNB
$C^{\text{n-GBR}}$	Uplink capacity available to non-GBR traffic at eNB transport scheduler
κ_c	TD priority term of a cell c; $\forall c \in S$
ρ_c	Mean throughput of a cell c as seen at Uu interface; $\forall c \in S$
$\widetilde{\rho_c}$	Mean throughput of a cell c as seen at TNL ; $\forall c \in S$
ρ_c^{OH}	Difference of ρ_c and $\widetilde{\rho_c}$ due to protocol overheads for a cell c; $\forall c \in S$
Δ_{SM}	Safety margin value
Define variables	
ζ_c	Throughput to be allocated to a cell c as seen at Uu interface; $\forall c \in S$
$\widetilde{\zeta_c}$	Throughput to be allocated to a cell c as seen at TNL; $\forall c \in S$
ν_c	Surplus from the allocated throughput to a cell c;$\forall c \in S$
$\widehat{\nu}$	Overall surplus throughput of all cells in eNB;
$\widehat{\kappa}$	Sum of κ_c of cells having surplus throughput i.e., $\nu_c > 0$;$\forall c \in S$

Execute

```
SET $\widetilde{\zeta_c}$ to $\left(C^{\text{n-GBR}} \cdot (\kappa_c / \sum_{c \in S} \kappa_c)\right); \quad \forall c \in S$
REPEAT
    SET $\nu_c$ to $(\widetilde{\zeta_c} - \widetilde{\rho_c}); \quad \forall c \in S$
    SET $\widehat{\nu}$ to $(\sum_{c \in S} H(\nu_c) \cdot \nu_c)$
    SET $\widehat{\kappa}$ to $(\sum_{c \in S}(1 - H(\nu_c)) \cdot \kappa_c)$
    FOR $\forall i \in S$
        IF $\widetilde{\zeta_i} \geq \widetilde{\rho_i}$ THEN
            SET $\widetilde{\zeta_i}$ to $\widetilde{\rho_i}$
        ELSE
            ADD $(\widehat{\nu} \cdot (\kappa_i / \widehat{\kappa}))$ to $\widetilde{\zeta_i}$
        ENDIF
    ENDFOR
UNTIL $\widehat{\nu} > 0$ OR $\widetilde{\zeta_c} = \widetilde{\rho_c};\forall c \in S$
SET $\zeta_c$ to $\left(\widetilde{\zeta_c} - (\rho_c^{\text{OH}} / \widetilde{\rho_c}) \cdot \widetilde{\zeta_c}\right); \quad \forall c \in S$
SET $\zeta_c$ to $(\zeta_c \cdot \Delta_{SM}); \quad \forall c \in S$
```

throughput is consumed or no needy cell is found. At this point, the throughput share of each cell $\widetilde{\zeta}$ at TNL becomes available. However, the MAC scheduler still needs to know how much this throughput will amount at the radio interface. At line 14, an estimation of TNL protocol overhead is subtracted from $\widetilde{\zeta}$ to obtain ζ, the throughput share of a cell at the radio interface. Finally, the effect of a predefined safety margin value is added to ζ. As mentioned earlier, ζ represents the cell capacity only for the non-GBR traffic. The GBR traffic is not affected by the back-pressure manager.

The safety margin is a unit less multiplicative factor. A value of safety margin less than 1.0 reduces the allowed cell throughput from the allocated throughput share. Similarly a value of the safety margin greater than 1.0 implies that more data traffic is being accepted from the radio interface than the calculated share. In this way, the safety margin value provides an additional control to fine tune the radio interface capacity which in turn helps regulate the buffer occupancy of the BE PHB queue in the transport scheduler.

The back-pressure manager periodically sends the computed cell throughput shares to the radio scheduler in order to limit the non-GBR traffic over the radio interface. The radio scheduler imposes the cell throughput share value as the maximum cell throughput for non-GBR traffic and allocates the MAC grants to the UEs accordingly. Limiting the radio interface capacity not only avoid the transport network congestion but also helps users with bad channel conditions to transmit with lower modulation and coding schemes (MCS) which are more robust against the transmission errors.

4.1.3 Simulation Scenarios and Results

The simulation analysis presented in this work can be divided into three stages. First of all, the implementation of the proposed algorithm is verified with simple configurations and the configuration values to obtain an optimum performance are discovered. In the second analysis, a modification in the back-pressure algorithm is suggested which makes it usable in various system load conditions. Finally, the modified algorithm is tested in an environment where the eNode-B cells are unevenly loaded and it is shown that scarce S1 uplink bandwidth resources are distributed fairly among the cells in proportion to their offered traffic load.

Table 4.2 lists the LTE simulator configuration parameters used to get the results presented in this work. Distribution of the users with respect to their running applications is shown in Table 4.3. Each user accesses one application at a time. The VoIP users generate traffic flows both in uplink and downlink to mimic full duplex voice conversation.

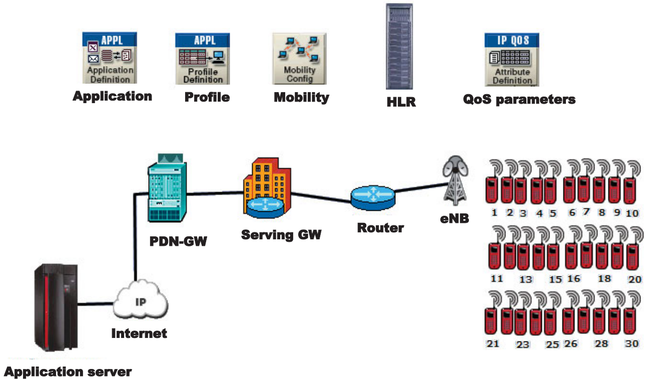


Figure 4.5: An overview of the considered simulation scenario in the OPNET simulator.

Table 4.2: Simulation configurations for coordinated uplink radio interface scheduling

Parameter	Configurations
User Profile Definition	
Number of active users	30 per cell (3 cells per eNode-B)
FTP traffic model	File size: constant 5MBytes , Inter-request time: exp(45) sec
VoIP traffic model	GSM EFR codec (12.2kbps) [72699], Call length: 90sec Inter-arrival time: exp(30) sec
HTTP traffic model	Number of pages per session: 5, Inter-arrival time: exp(12) sec Average page size: constant 100KB
User Mobility model	Random direction (50km/h)
TCP configuration	new Reno, receive window size 64KB, MSS: 1460Byte
Network Configuration	
Total Number of PRBs	50 PRBs (10MHz spectrum)
MAC scheduler	Time domain: Proportional Fair Frequency domain: Round Robin
Maximum scheduled user	6 users per TTI
S1 link type	Ethernet 100BaseX (100Mbps)
RED parameters for DiffServ uplink transport scheduler	PHB queue size: 512KB; Queue filling thresholds–min:33%, max:100%; maximum discard probability: 20%
Congestion detection threshold parameters	Congestion indication at 80% of PHB queue filling Congestion release at 20% of PHB queue filling
Simulation run time	2000 seconds

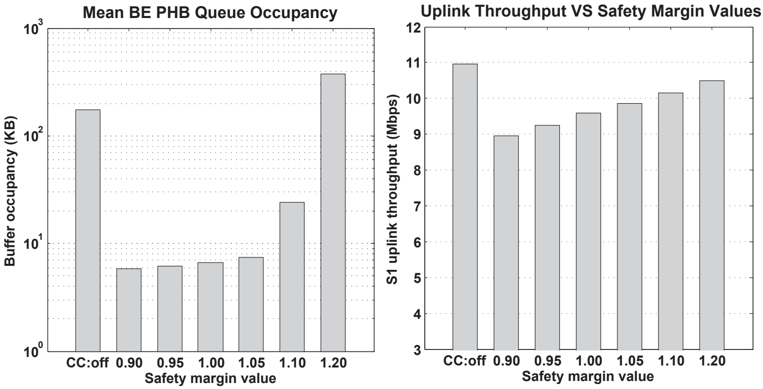
Table 4.3: User distribution for simulation analysis 1 and 2.

Number of UEs/cell – downlink			**Number of UEs/cell – uplink**		
QCI 1	*QCI 8*	*QCI 9*	*QCI 1*	*QCI 8*	*QCI 9*
10 VoIP	10 HTTP	2 FTP	10 VoIP	none	8 FTP

4.1.3.1 Simulation Analysis 1

In this simulation analysis the performance of the back-pressure algorithm is studied using different values of the safety margin. Due to the fact that there exists a certain signaling/processing delay in the information collection process of the back-pressure manager, there could be some time lag in the computation of the S1 uplink capacity available to the non-GBR traffic of the cells. Therefore, in this analysis the offered load to the S1 link of limited capacity is tuned using a range of safety margin values from 0.95 to 1.20. Moreover, the congestion control mechanism of the back-pressure manager is kept activated irrespective of the congestion detection/release triggers received from the transport scheduler. This is because a simulation study of the system shows that frequent congestion detection/release triggers could lead to unstable system states. For example, when the system is in the congestion state the PHB queue buffer occupancy increases and activates the congestion detection trigger. In response to this, the back-pressure manager controls the congestion and brings the buffer occupancy down to the lower threshold value. This generates the congestion release trigger and makes the back-pressure manager stop all traffic regulation. If the offered traffic load remains unchanged then after a short period of time the buffer occupancy again surpasses the upper threshold leading to a new congestion detection trigger. This cyclic behavior continues and, as a result, the buffer occupancy fluctuates between the two threshold values. Further study of this behavior reveals that such oscillations can be rectified if the congestion control mechanism of the back-pressure manager is always kept functional regardless of the congestion detection/release triggers. This strategy not only stabilizes the system but also help regulate the uplink traffic preemptively before the transport network fully enters in congestion state.

Figure 4.6(a) shows how the buffer occupancy of the BE PHB queue varies against different values of safety margin and also when the proposed algorithm is not used, i.e., CC:off. In the CC:off case, the Random Early Detection (RED) [FJ93] scheme performs the buffer management at the DiffServ scheduler. The details of the RED scheme will be described in Section 4.3.2. During the congestion situations the RED scheme discards a large number of packets transmitted by the UEs in uplink. For example, with the current configuration of RED pa-



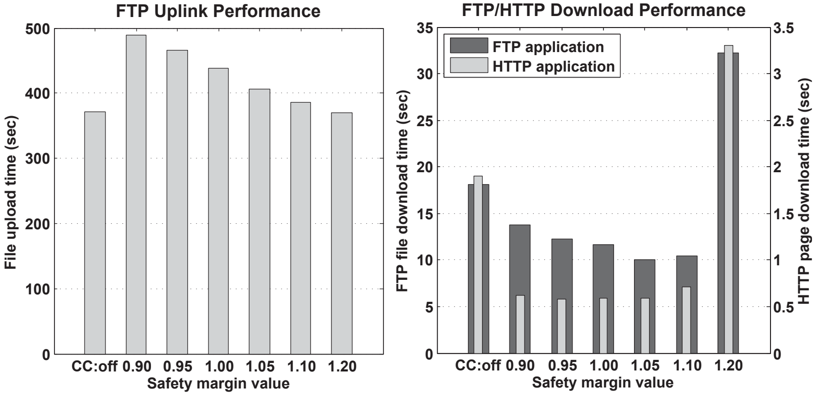
(a) (b)

Figure 4.6: Mean buffer occupancy of BE PHB queue and uplink throughput for different safety margin values.

rameters and offered load approximately 12 IP packets are discarded per second. These losses have to be recovered by the upper layer protocols (e.g., TCP) through the retransmissions. As a consequence, the battery power of the UE as well as the radio interface capacity is wasted. The proposed mechanism avoids the packet discards during the congestion and keeps the BE PHB buffer occupancy at the lowest possible level to enhance the system performance.

The role of the safety margin factor can also be studied both in Figure 4.6(a) and Figure 4.6(b). The safety margin provides a fine control on the BE PHB buffer occupancy level. This control is important because of the fact that too low buffer occupancy could cause buffer under-runs while too high buffer occupancy could lead to packet drops. Therefore, an appropriate value of the BE PHB queue buffer occupancy must be achieved by tuning safety margin value. Figure 4.6(a) indicates that using the safety margin value of 1.20 (which allows 20% more traffic from radio interface than the available capacity at transport network) could lead to a full buffer situation. On the other hand, an interesting observation can be made from Figure 4.6(b) that with the lower values of safety margin the available uplink transport capacity cannot be utilized as much as in the CC:off case. For example, a safety margin value of 0.90 makes uplink transport carry 10% less traffic compared to the CC:off case just because of buffer under-runs. Therefore, it is expected that lower values of safety margin can degrade uplink application performance. This effect can be observed in Figure 4.7.

Figure 4.7 shows the performance of uplink and downlink applications. One can observe the application performance for different values of safety margin and compare them against the case when the congestion control mechanism is not used, i.e., CC:off. As far as the FTP uplink performance is concerned, the CC:off case provides the lowest mean file upload time by fully utilizing the available uplink transport network capacity. The safety margin value of 1.20 delivers the FTP uplink performance close to the best case while all other investigated values of safety margin degrade user experience of file upload. On the other hand, the performance of downlink applications, i.e., HTTP/FTP is superior for nearly all safety margin values compared to CC:off case. The safety margin value of 1.20 represents the only scenario where downlink application performance is worse than the CC:off case. This is because all safety margin values except 1.20 create lower buffer occupancy as compared to CC:off case. Relatively high safety margin values make the BE PHB queue buffer occupancy increase and hence TCP acknowledgement packets have to wait longer in the queue leading to additional delays in file download. Studying the presented results, the safety margin value of 1.10 can be considered as the best choice which provides a considerable gain in downlink application performance at the cost of negligible degradation in uplink application performance.



(a)

(b)

Figure 4.7: FTP and HTTP application performance for different values of safety margin.

4.1.3.2 Simulation Analysis 2

It can be inferred from previous simulation analysis that the buffer under-runs of the BE PHB queue at the uplink transport scheduler could be avoided by carefully regulating the safety margin value. However, the problem is to determine an appropriate value of the safety margin for each scenario. Owing to the fact that an optimum value of safety margin depends on the current carried traffic load, the feasible safety margin value should be selected automatically in response to the offered traffic load level. This can be achieved by defining a range of safety margin values against a range of buffer occupancy threshold levels. An operative value of the safety margin is then determined by linear interpolation between the two boundary values in response to the buffer filling level. For example, when there is low buffer occupancy, a large safety margin value should be used to allow more traffic from the radio interface and vice versa.

In the above proposed modification of algorithm, buffer occupancy serves an indication of congestion level in the network. However, a literature survey reveals that buffer occupancy alone cannot always be used as a measure of the congestion level at a network interface [J. 05]. Therefore, it is not recommended to regulate the safety margin using bare buffer occupancy values. This issue is resolved by devising a sophisticated approach that considers the egress rate of the transport scheduler for the BE PHB queue along with its buffer occupancy. In other words, allow high buffer occupancy when egress rate is high so that buffer under-runs can be avoided and keep buffer occupancy low when egress rate is small to circumvent the buffer overflows and the excessive queuing delays. A ratio of buffer occupancy level and egress rate, termed as effective buffer filling level, can be defined as below

$$\text{Effective buffer filling level} = \frac{\text{Buffer occupancy (bits)}}{\text{Egress rate (bps)}} \tag{4.8}$$

Effective buffer filling level or EBFL is measured in units of second and it hints at the maximum queuing delay experienced by a packet in the PHB queue. The safety margin value varies linearly with the effective buffer filling level as indicated below:

$$\text{Safety margin} \propto f(\text{effective buffer filling level}) \tag{4.9}$$

In this simulation analysis different value ranges for EBFL are studied while keeping the range of the safety margin value fixed, i.e., 0.9–1.5. Moreover, the uplink bandwidth shaping rate is reduced from 11Mbps to 10Mbps in order to evaluate the system performance at an even higher level of transport network congestion.

Figure 4.8 shows the mean buffer occupancy and the uplink throughput for three different value ranges of the EBFL. It is noticed that the EBFL values can be used to achieve the desired level of PHB queue buffer occupancy. Moreover, with the help of this new approach buffer under-runs are eliminated and full capacity of uplink transport network is utilized for all three investigated EBFL value ranges. As a result, it can be expected that the FTP uplink performance will no longer be compromised due to buffer under-runs.

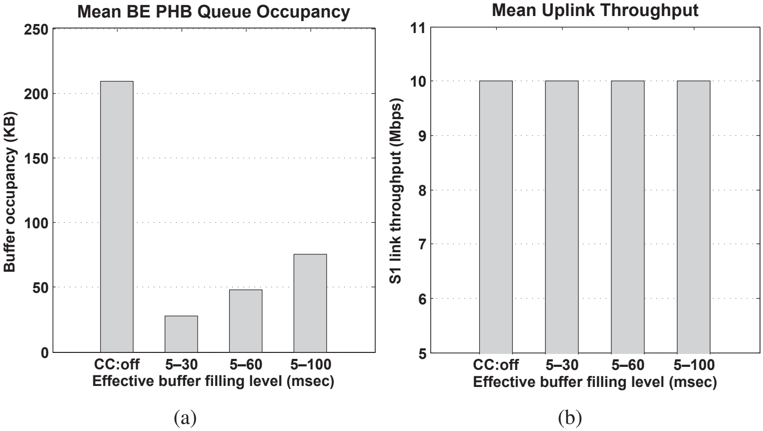
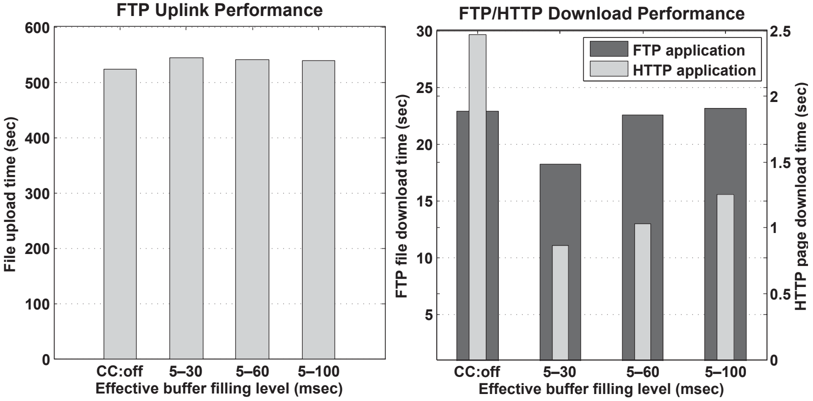


Figure 4.8: Mean buffer occupancy of BE PHB queue and uplink throughput for several EBFL values.

Figure 4.9 shows the performance of uplink and downlink applications in the system. As anticipated the FTP uplink performance is almost identical for all cases. In addition, the back-pressure manager avoids the packet losses which, otherwise, might be encountered due to high buffer occupancy and hence saves UE battery power. Moreover, lower PHB queue buffer occupancy by the congestion control mechanism also helps improve the downlink application performance as shown in Figure 4.9(b). It can be noticed that the HTTP application benefits more from low buffer occupancy and shorter round trip time compared to the FTP downlink application. This is because of small object sizes (100KB) which download completely during the TCP slow start phase. In slow start phase, shorter round trip time improves user throughput significantly higher compared to the congestion avoidance phase. Moreover, it can be seen that FTP uplink performance is not affected by the shorter TCP round trip time. This is due to the reason that uplink Uu interface capacity has been limited by MAC scheduler to avoid transport network congestions. The best choice of the EBFL value range can be accredited as 5–30msec. In principle this choice is independent of the offered traffic load as shown in the next subsection.



(a)

(b)

Figure 4.9: FTP and HTTP application performance for different values of effective buffer filling level.

4.1.3.3 Simulation Analysis 3

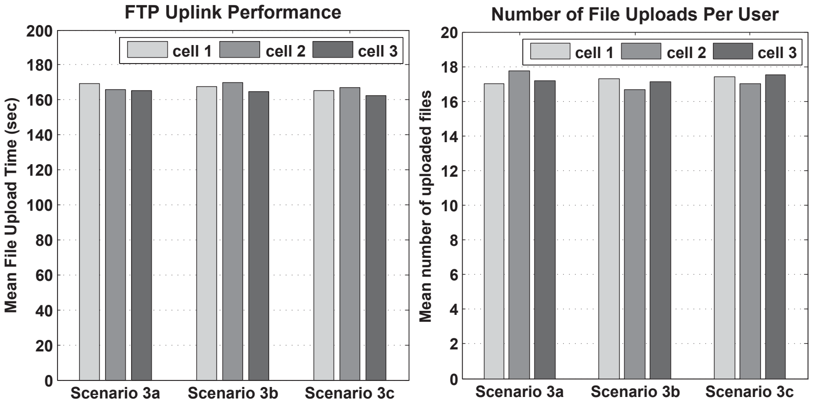
In this subsection the performance of the back-pressure manager will be evaluated when the three cells of the eNode-B are unevenly loaded with uplink user traffic. With the help of this analysis it will be shown that the proposed congestion control algorithm distributes transport network bandwidth resources among all the cells in proportion to their carried traffic load. Table 4.4 shows the distribution of FTP uplink users in three cells of eNode-B for the investigated scenarios. Apart from the FTP uplink users, each cell has 10 VoIP users who are generating both uplink and downlink traffic. UEs with HTTP and FTP downlink application are not included in this investigation. The S1 uplink capacity is still limited to 10Mbps. Moreover, based on the findings of previous simulation analyses, the following parameter settings are used for the back-pressure manager: Effective buffer filling thresholds of (5–30msec) with safety margin value thresholds (0.9–1.5).

Figure 4.10 shows the FTP uplink performance as experienced by the users in the three cells of the eNode-B. It can be seen that the FTP file upload time is inde-

Table 4.4: Number of FTP uplink users in the three cells of the eNode-B.

	Cell 1	Cell 2	Cell 3
Scenario 3a	10	10	10
Scenario 3b	8	14	8
Scenario 3c	6	14	10

pendent of the user traffic load in the cells. A similar statement can also be made for the number of completed file uploads per user. Hence it proves that the proposed algorithm fairly distributes the limited transport bandwidth resources among the cells of the eNode-B in accordance to their traffic demands. For example, when cells are equally loaded they receive the equal share of the transport bandwidth resources and when the cells are offered with unequal traffic load they are provided with the proportional share of the S1 link capacity.



(a) (b)

Figure 4.10: FTP uplink application performance when the eNode-B cells are unevenly loaded with user traffic.

It can be concluded from the study of simulation results that the proposed mechanism performs an effective congestion control and enhances the system performance in several ways. For example, it saves battery power of UEs by avoiding the packet drops of uplink traffic at the congested transport network, it brings performance gain to user applications, and it also improves the system stability in congestion situations. In addition it was also revealed that the uplink congestion at the last mile can adversely affect the downlink application performance by delaying the higher layer acknowledgements in the uplink direction. The proposed algorithm also completely mitigates such adverse effects and helps achieve the sustainable system performance under the variable traffic load conditions.

The back-pressure mechanism is effective only for the uplink communication. The next section discusses an adaptive radio MAC scheduling algorithm which does not explicitly coordinate with the transport network but conforms its behavior to the air interface load and can be used both for the uplink and downlink communications.

4.2 Adaptive Fair Radio Interface Scheduling

The LTE packet scheduler is in charge of scheduling bandwidth resources among the users by following one of the specific policies to meet system performance targets. These targets may include, e.g., maximizing cell throughput, providing fairness of service among the users or guaranteeing the required QoS etc. Every packet scheduling scheme or policy has certain merits and demerits, e.g., if a scheme tries to maximize the cell throughput, it compromises the fairness among the users. Similarly if another scheme emphasizes fairness among users, it fails achieve the maximum cell throughput. Owing to the fact that the selection of the packet scheduling scheme belongs to those configurations which must be selected during the network planning phase, the network operator has to select one of the available schemes exclusively. In this study, an adaptive packet scheduling approach is introduced which dynamically changes its behavior based on system requirements. Though the downlink shared channel is taken as a reference, the proposed scheme is also valid for the uplink transmissions.

There are numerous research studies which focus on different aspects of the packet scheduling schemes in LTE. For example, QoS aware packet schedulers for OFDMA are discussed in [NH06] and [F.R04]. [A.07] and [P. 08] introduce a packet scheduling framework for LTE where the resource allocation procedure is decoupled into a time and a frequency domain scheduler. They also present a study where different scheduling schemes are used in both time and frequency domain to control fairness among users. In [Y. 11] authors have presented a modified proportional fair scheduler for an enhanced QoS service aware scheduling in LTE. However, they do not address the service fairness issues among the users in a certain QoS class. A study which is closely related to this work can be found in [G. 08]. It introduces a metric weighting based on the number of users in the cell to control resource assignment in the frequency domain scheduler.

In the following, background information of LTE scheduling is presented after which the proposed scheduling scheme is discussed in section 4.2.1 and 4.2.2.

4.2.1 LTE Packet Scheduler

As mentioned earlier, the LTE packet scheduling refers to the process of dividing and allocating bandwidth resources among users who have data to be transmitted at the air interface. In LTE the resource scheduling is performed every Transmission Time Interval (TTI) which amounts to 1ms. This work follows the decoupling principle of the LTE packet scheduling procedure as suggested in [P. 08]. In this way, every TTI the time domain scheduler performs user identification, QoS classification and bearer prioritization. Then at the frequency domain scheduler selected users are served with a certain number of resources. In the following further details about the two schedulers are given.

4.2.1.1 Time Domain Scheduler

The time domain (TD) scheduler is responsible for selecting a subset of active users in a cell for transmission in a TTI. The selection criteria involve a priority metric which may be based on service preferences, current user channel conditions (i.e., CQI) as well as other user throughput and packet delay constraints related to QoS requirements. In the following three basic schemes used for user prioritization and selection are introduced.

a) Blind Equal Throughput

The Blind Equal Throughput (BET) scheduler which is also called ‘Fair Scheduler’ provides throughput fairness among all active users regardless of their channel conditions. To achieve this property, the BET scheduler uses a priority metric which considers history values of average user throughput. This priority metric is calculated for a user *i* as follows

$$M_i^{BET} = \frac{1}{\overline{R_i}(t)} \tag{4.10}$$

here $\overline{R_i}(t)$ is the past average throughput of user *i*. The moothed value of $\overline{R_i}(t)$ is computed using a weighted moving average formula, e.g.,

$$\overline{R_i}(t) = \frac{1}{T} \cdot R_i(t) + (1 - \frac{1}{T}) \cdot \overline{R_i}(t-1)$$

here $R_i(t)$ is the instantaneous value of user received goodput at the MAC layer excluding any HARQ retransmissions. It should be noted that the value of $R_i(t)$

is taken as zero for that TTI during which user i is not scheduled despite being a candidate for scheduling. Moreover, the value of $\overline{R_i}(t)$ is only updated during the times when user i is active.

It is clear from equation 4.10 that the BET scheduler prioritizes those users whose average throughput has been lower in the past. Following this scheme the users with lower average throughput will be scheduled until they achieve the same throughput as the other users in the cell. This implies that users with bad channel conditions are allocated more resources compared to the users with good channel conditions. Consequently, throughput fairness among the users is achieved at the expense of spectral efficiency. That is why, the BET scheduler cannot attain as high cell throughput as achieved by other channel aware schedulers.

Practically, throughput fairness is not always required for all users in the system. Instead, fairness is desirable among a set of users belonging to a certain QoS class. Most commonly, the BET scheduler can be adapted for a system with multiple QoS classes by introducing QoS weight factors (Q) in the priority metric. This scheme is called weighted fair queuing. The priority metric now takes the following form

$$M_i^{BET_{qos}} = \frac{1}{\overline{R_i}(t)} \cdot Q_k \tag{4.11}$$

here Q_k is the priority weight associated to a certain QoS class k to which user i belongs. In this way, the resources are shared among the QoS classes in proportion to their associated weights while the users within a certain QoS class still experience the throughput fairness.

b) Maximum Throughput Scheduler

In LTE, the user channel quality can be estimated from the CQI (Channel Quality Indicator) reports which are periodically sent by the user equipment to the base station through control messages. With the help of CQI reports, the packet scheduler can predict the maximum achievable throughput for the respective user. This information can be used in the priority metric to prioritize users with good channel conditions over the users with bad channel conditions. This helps in achieving high spectral efficiency and hence high cell throughput. The packet scheduler which works according to this principle is called Maximum Throughput (MaxT) scheduler. The priority metric for the MaxT scheduler is given as follows

$$M_i^{MaxT} = \widehat{R}_i(t) \tag{4.12}$$

here $\widehat{R}_i(t)$ is the instantaneous achievable data rate based on channel quality or CQI value of user i. Though the MaxT scheduler is capable of delivering the highest possible cell throughput, it comes at the expense of fairness. This is because users with bad average channel conditions are selected less often and, therefore, they achieve lower throughput compared to the users with good channel conditions.

c) Proportional Fair Scheduler

Both the BET and the MaxT schedulers operate on two extremes of fairness and spectral efficiency. In practice an intermediate solution is required which lies between these extremes so that it exploits the good channel conditions while still providing a certain degree of fairness among the users. Such a trade-off behavior can be achieved with the help of the Proportional Fair (PF) scheduler. The priority metric for the PF scheduler is obtained by combining the priority metrics of BET and MaxT scheduler, i.e.,

$$M_i^{PF} = \frac{\widehat{R_i}(t)}{\overline{R_i}(t)} \tag{4.13}$$

The philosophy behind the PF scheduler is to weigh the MaxT priority metric with the inverse of past average throughput so that the users with bad channel conditions get a bit higher priority when they suffer from low throughput.

4.2.1.2 Frequency Domain Scheduler

The frequency domain (FD) scheduler allocates resources (number of PRBs) to the users provided by the TD scheduler. Theoretically, it is possible for the FD scheduler to use those scheduling schemes which have been discussed for the time domain scheduler. However, in order to avoid implementation complexities in real hardware a simple resource distribution scheme like Round Robin is employed in the frequency domain scheduling. As an outcome of the Round Robin scheme, the available PRBs are evenly distributed among the n selected users. Usually there exists an upper limit on the number of users N which can be served in a TTI. Therefore, if the time domain scheduler prepares a list of users with $n > N$, only the first N users in the prioritized list are served in a TTI leaving the rest of the users un-served.

4.2.2 Adaptive Fair Scheduler

In practice, if the radio interface is not congested, a network operator prefers the BET scheduler to achieve high fairness and enhance the cell coverage. On the other hand, in congestion situations the proportional fair scheduler is favored in order to increase spectral efficiency by compromising fairness. As far as system load in wireless networks is concerned, it changes dynamically with time due to several reasons like user mobility patterns, time varying user traffic profiles, and channel conditions etc. This may give rise to events when offered traffic load is low and the packet scheduler has more than enough radio bandwidth resources to serve that load. Such situations can also be encountered when a bottleneck appears in the transport network due to which small traffic load is observed at the radio interface. The proposed Adaptive Fair scheduler has been designed to assess such situations and behave accordingly. It evaluates the radio interface congestion level and acts as a BET scheduler if no or light congestion is detected and operates as a PF scheduler otherwise. The adaptive fair scheduler is also called coordinated radio interface scheduler because it coordinates with the transport and radio interface congestion to determine its behavior. The scheduler has been designed in a way that instead of making abrupt changes in its behavior between PF and BET schedulers, Adaptive Fair scheduler gradually shifts based on the perceived congestion level. An overview of such a behavior is shown in Figure 4.11. In this figure the fairness level is determined from the ratio of maximum value and minimum value among the active user throughput values in the cell. In this way, perfect fairness is achieved when this ratio has a value equal to 1.

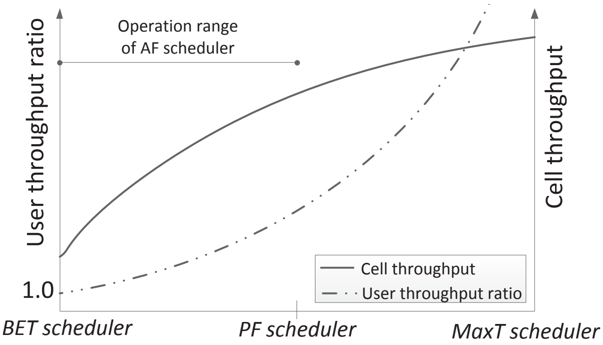


Figure 4.11: Operation range of the proposed adaptive fair scheduler. The solid curve represents the hypothetical cell throughput for different scheduler behaviors. The curve with the dotted line shows service fairness level among the users in terms of maximum user throughput to minimum user throughput ratio. Large values of this ratio indicate lower fairness.

The adaptive fair scheduler is proposed to work as a time domain scheduler for LTE. The priority metric for the AF scheduler is shown below:

$$M_i^{AF} = w_j(t) \cdot M_i^{BET} + (1 - w_j(t)) \cdot M_i^{PF} \tag{4.14}$$

substituting the values of M^{BET} and M^{PF} in the above equation:

$$M_i^{AF} = w_j(t) \cdot \frac{1}{\overline{R_i}(t)} + (1 - w_j(t)) \cdot \frac{\widehat{R_i}(t)}{\overline{R_i}(t)} \tag{4.15}$$

where

$$w_j(t) = \begin{cases} max(w_j(t-1) + \Delta w, 1) & \text{if } \Delta w \geq 0, \\ min(w_j(t-1) + \Delta w, 0) & \text{if } \Delta w < 0 \end{cases}$$

and

$$\Delta w = w_j(t) - w_j(t-1)$$

here $w_j(t)$ is the weight factor for a cell j which represents the congestion level at the radio interface of the respective cell. Its value range is from 0 to 1. For example, $w_j(t) = 0$ means that cell j is in a sever congestion and $w_j(t) = 1$ implies that the cell is undergoing a mild congestion or it is not congested at all. There are several possible ways to determine the weight factor based on the radio interface congestion level. This works follows the idea of estimating the radio congestion level in a cell by observing the PDCP buffer occupancy of all active users in the cell.

It has already been discussed in section 4.1.3.2 that buffer occupancy alone cannot always be used as a measure of the congestion level at a network interface. For example, consider a situation where several TCP users are being served in a cell. If the TCP window size is small there will be no large buffer occupancy at the eNode-B PDCP buffers despite actual radio congestion. Similarly, if users a have large TCP window size a large buffer occupancy can be observed regardless of the fact that the users are experiencing high throughput and there is no radio congestion.

An alternative approach of estimating radio interface congestion is to use the packet waiting time of the PDCP buffer queue instead of the bare queue occupancy level values. This packet waiting time is defined as Effective Buffer Filling Level (EBFL) which is calculated by dividing the total buffer occupancy with the total average cell throughput, in a way similar to equation 4.8, i.e.,

$$\text{EBFL} = \frac{\text{Total PDCP buffer occupancy}}{\text{Average cell throughput}} \tag{4.16}$$

Two threshold values of EBFL are used corresponding to the congestion level: The lower threshold value which maps to $w_j = 1$ represents no or a very light congestion level and the upper threshold value mapped to $w_j = 0$ indicates a severe

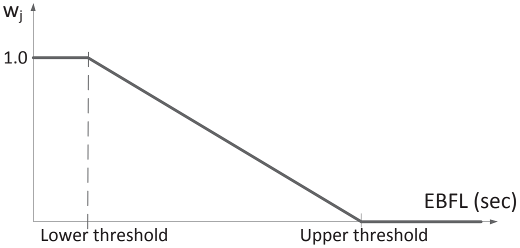


Figure 4.12: Linear mapping of EBFL values to the weight factor (w_j) values.

congestion. This can be seen in Figure 4.12 where the EBFL value is used to obtain the weight factor through the linear mapping.

It should be noted that the EBFL value is computed for each cell at the eNode-B. This means that the total buffer occupancy in the numerator of equation 4.16 is obtained by adding the PDCP buffer occupancy of all active users in the corresponding cell. Similarly the average throughput, shown in the denominator, is calculated by summing up the average throughput values of individual users, i.e.,

$$\text{Average throughput of cell } j = \sum_{i \in U_j} \overline{R_i}(t)$$

where U_j is a set of active users attached to cell j.

4.2.3 Simulation Scenario and Results

Figure 4.13 shows an overview of the simulation scenario in OPNET. The system is populated with 10 users running FTP download application. In order to highlight the effects of the Adaptive Fair scheduler on system performance, the users are classified into two groups. Each group consists of 5 users and the user mobility is disabled so that the users remain stationary during the whole simulation time. One group of the users is located near the eNode-B and hence has good channel quality. The second group of users is placed very far from the eNode-B at 200m distance and, therefore, suffers from bad channel quality. The simulation configuration parameters are shown in Table 4.5.

The user FTP application has been configured in a way that the users download files one after the other without any time gap. Therefore a large traffic load is observed in the system which brings the radio interface to the congested state. The transport network links (i.e., Ethernet links between serving-GW and eNode-B) are

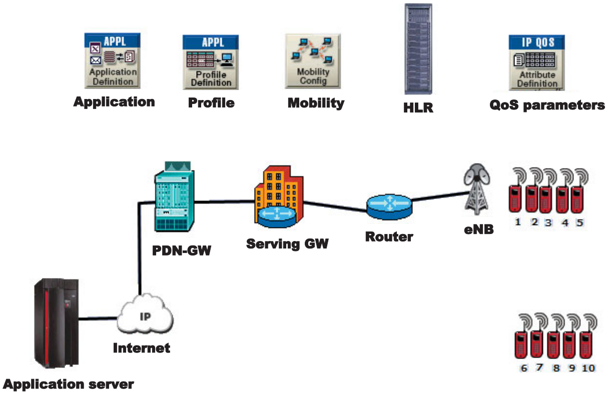


Figure 4.13: Overview of the considered simulation scenario in the OPNET simulator.

of 1Gbps capacity which is sufficient to carry the generated traffic load. Therefore, the only bottleneck in the system exists at the radio interface.

Table 4.5: Configurations for simulation scenario of Adaptive Fair scheduler.

Parameter	Configurations
Total Number of PRBs	50 PRBs (10MHz spectrum)
Number of users	5 users near the eNB at a distance of 50m and the other 5 users are placed near the cell boundary, i.e., at 200m distance from eNB.
MAC scheduler	Time domain: BET, PF and Adaptive Fair Frequency domain: Round Robin
Maximum number of scheduled user	6 users per TTI
EBFL thresholds	Lower threshold: 50msec Upper threshold: 200msec
User mobility model	none. Users are static.
User application	FTP file download
FTP traffic model	FTP File size: constant 10MByte. Continuous file downloads one after the other without time gap.
Simulation run time	2000 seconds

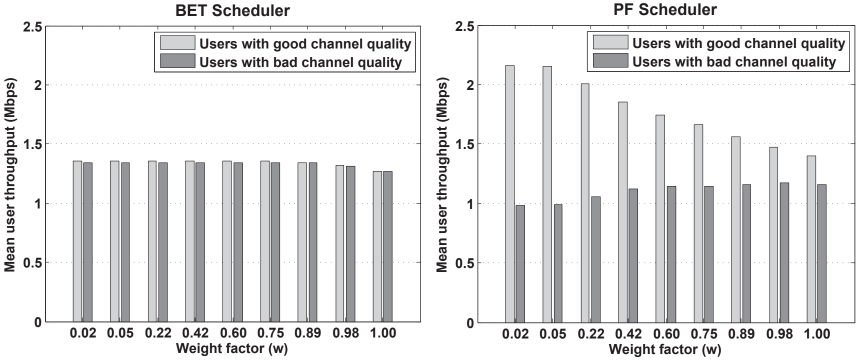
The congestion level at the radio interface is controlled by the throughput bandwidth shaping function at the last-mile router (i.e., Router in Figure 4.13). With the help of the bandwidth shaping function the amount of downlink traffic on the S1 link can be controlled in order to tune the radio interface congestion level. In this simulation setup the traffic shaping rate at the bandwidth shaping function has been varied to get a range of values (from 0 to 1) for the weight factor (w).

Figure 4.14(a) shows the mean downlink user throughput when the BET scheduler is used. The throughput values have been shown for user groups with good and bad channel conditions at different radio congestion levels represented by w. It should be remembered that the lower the value of w, the lower the radio interface congestion level. Moreover, as evident from equation 4.10 and 4.13, the w factor has no direct influence on the behavior of the BET and PF scheduler. Instead the w factor values on the x-axis of Figure 4.14 just represent different radio congestion levels at which performance of these two schedulers can be compared with the adaptive fair scheduler. It can be noticed in the Figure 4.14(a) that the the user throughput is not affected by the radio interface congestion which is an expected behavior of the BET scheduler. Hence, the users experience similar throughput regardless of their channel quality for all radio congestion levels.

Figure 4.14(b) shows the performance of the proportional fair scheduler in terms of the mean downlink user throughput. With the proportional fair scheduler users always manage to achieve high throughput when they have good channel conditions. Even during the transport network congestion, TCP tries to get higher data rates for users with better channel conditions until the packet drops happen due to congestion. TCP packet drops trigger the TCP timeout event and hence TCP has to undergo the slow-start phase. Such packet drops are avoided in this simulation setup by employing a large buffer for the bandwidth shaping function at the last-mile router. Due to this reason, users with good channel conditions can attain higher data rates compared to the users with bad channel condition despite the transport network bottleneck.

Figure 4.15 shows the behavior of the proposed Adaptive Fair scheduler. It can be seen that during the times when the radio interface has a high offered traffic load (i.e., small values of w), the adaptive fair scheduler behaves more like proportional fair scheduler. This helps users to exploit good channel quality to enhance overall cell throughput by compromising the fairness. As soon as the bottleneck shifts towards the transport network. it relieves the radio interface congestion and consequently the value of w increases more towards 1. This makes the adaptive fair schedule to act more like a BET scheduler to provide throughput fairness among the users.

The mean cell throughput values can be seen in Figure 4.16 for three types of time domain schedulers. It can be observed that the adaptive fair scheduler conforms with the radio interface congestion situation by providing high cell throughput when the radio interface is exposed to a high traffic volume. However, when there are access resources available at the radio interface, user throughput fairness



(a) (b)

Figure 4.14: Mean per user throughput values for BET and PF time domain scheduler.

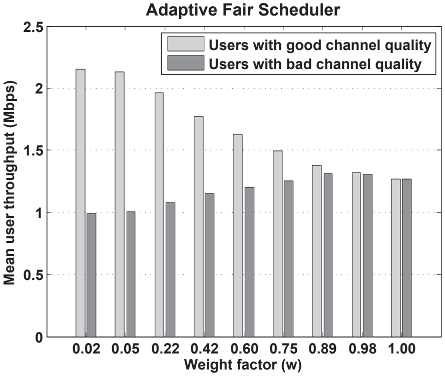


Figure 4.15: Mean per user throughput values for adaptive fair time domain scheduler.

is preferred by adapting to the BET scheduler behavior.

The discussion on simulation results concludes that the Adaptive Fair scheduler can dynamically change its behavior in response to system load conditions. The proposed scheduler is intelligent enough to precisely evaluate the radio interface congestion level and adapt the system demands. During the time when radio interface acts as a bottleneck due to scarce radio bandwidth resources, it enhances the spectral efficiency by tending towards Proportional Fair scheduler. At other times when radio interface is not congested, it provides throughput fairness among the

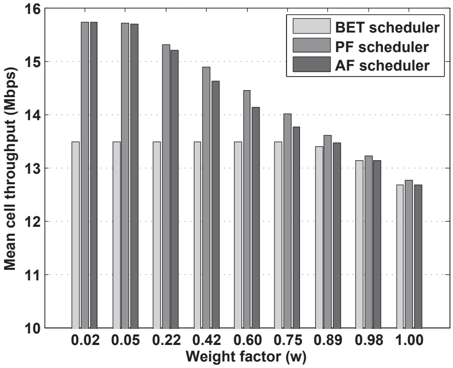


Figure 4.16: Mean per cell throughput values for three types of time domain schedulers.

users at the expense of radio resources through the use of Blind Equal Throughput scheme. The simplicity of Adaptive Fair scheduler in terms of implementation makes it suitable for deployment in real systems.

So far two enhancements related to the air interface scheduling has been discussed in this chapter. The next topic is about the packet queue management for the air interface scheduler. The following section discusses some popular queue management schemes which can be employed in this regard.

4.3 PDCP Buffer Management Schemes

Ever increasing data rate demands of the user applications can easily bring LTE air interface to a congested state despite its high efficiency. In other words, there could be situations when the instantaneous data rate available on the air interface is smaller than the data rate available on the transport network. This leads to buffering at the PDCP layer in the eNode-B, when referring to downlink communication. On the one hand, this buffering is a blessing that it provides flexibility to the MAC scheduler so that the instantaneous data rate at the air interface can be varied in order to adapt to user current radio channel conditions and get advantage of multi user diversity. On the other hand, if the data rate provided by the air interface fails to catch up with the data rate from transport network for a long period, it results in a large amount of buffered data accumulation. Too high PDCP buffer occupancy, in turn, causes longer queuing delays before the data can be transmitted over the air interface. Long queuing delays then lead to large packet end-to-end delays producing adverse effects for both the realtime and the non-realtime applications. This is because, realtime applications have strict requirements on end-to-end packet delay which must be fulfilled to achieve acceptable user Quality of Experience (QoE). For example, conversational VoIP demands mouth-to-ear delay to be less than 150ms in order to achieve transparent interactivity. Similarly, user QoE for TCP based non-realtime applications is also greatly influenced by the end-to-end packet delay as shown by the following equation

$$\text{TCP throughput} < \frac{MSS}{RTT} \cdot \frac{1}{\sqrt{PLR}} \tag{4.17}$$

where MSS is the maximum TCP segment size, RTT is the TCP segment round trip time and PLR is the packet loss rate [M. 97].

PDCP buffer occupancy also plays an important role in the handover process (see Section 2.4.6 for details about the LTE handover process). During inter eNode-B handovers, when the connection is interrupted from the source eNode-B and again made at target eNode-B, the data to UE is buffered at the source eNode-B and forwarded to target eNode-B over the S1/X2 interface. There are two constituents of this data: i) the contents of PDCP buffer at the beginning of handover event and, ii) new incoming data from S-GW until the destination of data delivery path is switched from source eNode-B to the target eNode-B. The larger the PDCP buffer contents, the higher will be the S1/X2 traffic volume. This large traffic volume which must be transported within a short period of time consumes expensive transport network bandwidths in addition to extending the handover delays [3GP08].

In the fixed internet, a typical action of a router is to drop packets when the data rate demand of an application exceeds the available data rate in a part of the network. This gives the application a hint about the network congestion in reaction to which it tries to adapt to the available network capacity by reducing the transmission rate. For example, TCP reduces its transmit window size on detecting the packet loss, thus adapting to the available rate. Similarly, other applications like VoIP or video streaming can detect the packet loss via RTCP (Real Time Transport Control Protocol) feedback, and can adjust to the network conditions accordingly.

In order to allow the above mechanisms to work for LTE and to avoid excessive delays, a buffer management scheme is required at the PDCP layer. This scheme should keep buffer occupancy to a minimum level needed to achieve the optimum end-to-end application performance. This work investigates two buffer management schemes which can be used to control PDCP buffer occupancy in an effective manner to optimize the end-to-end performance. The first scheme is a packet waiting time based discard mechanism which is also recommended by 3GPP standards. The second scheme is based on the well-known Random Early Detection (RED) mechanism. In addition, the performance of a simple tail drop scheme is also compared with that of the two aforementioned buffer management schemes.

4.3.1 Discard Timer

In this scheme, a maximum limit is imposed on the waiting time of a packet in a queue. Packets are time stamped upon their arrival in the queue. The waiting time of packets is continuously monitored and those packets, for which the maximum limit of waiting time is exceeded, are discarded. With the help of this scheme a precise upper bound on queuing delay can be achieved.

At the PDCP layer this scheme is implemented using a packet queue of large capacity. At the inlet, incoming IP packets from the higher layer are enqueued without any discard. However, when the data is requested by lower layers, typically, at the MAC scheduling events, the queuing delay of each forwarded packet is ensured to be less than the maximum threshold by discarding older packets. This mechanism is applied independently on each bearer's PDCP buffer. Figure 4.17 elaborates how the discard timer algorithm functions.

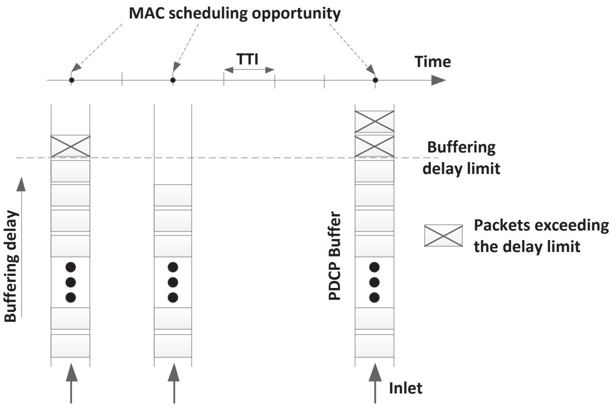
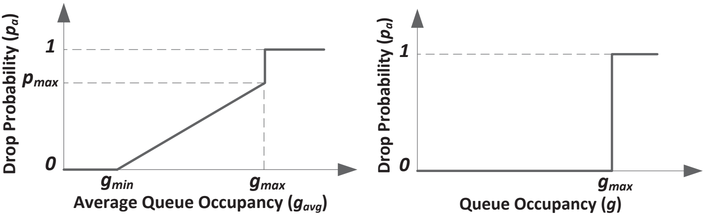


Figure 4.17: Discard timer based buffer management.

4.3.2 Random Early Detection (RED)

The basic idea behind random early detection is to detect incipient congestion and notify the end hosts, allowing them to reduce their transmission data rates before queues in the network overflow. For this purpose, an implementation of RED continuously monitors the average queue length; when it exceeds beyond a threshold, incoming packets are randomly dropped with a certain probability irrespective of the fact that there still exists room for more packets. With this strategy, the dropping of packets serves as an early notification conveyed to the source to reduce its transmission rate.

The RED algorithm itself consists of two main parts: average queue size estimation and the decision of whether an incoming packet should be dropped or buffered. The average queue occupancy is computed using a low pass filter with exponential weighted moving average. The average queue occupancy g_{avg} is then compared with two threshold values, a minimum threshold g_{min} and a maximum threshold g_{max}. As long as the average queue occupancy is less than the minimum threshold all incoming packets are simply enqueued and no drop takes place. When average queue occupancy grows beyond the minimum threshold but remains less than the maximum threshold, some of the incoming packets are dropped randomly following a certain probability p_a, termed 'drop probability'. If the congestion continues growing and the average queue occupancy exceeds the maximum threshold, all incoming packets are dropped to avoid persistently full queues. This behavior has been depicted in Figure 4.18(a).



(a) RED scheme

(b) Tail Drop scheme

Figure 4.18: Drop probability profile of RED and Tail Drop based schemes.

The drop probability p_a is calculated as a function of average queue occupancy g_{avg} as follows:

$$p_a = \begin{cases} 0 & \text{if } g_{avg} \leq g_{min} \\ p_{max} \cdot \dfrac{g_{avg} - g_{min}}{g_{max} - g_{min}} & \text{if } g_{min} < g_{avg} \leq g_{max} \\ 1 & \text{otherwise} \end{cases} \tag{4.18}$$

The advantage of the RED approach is that it prevents massive packet loss due to sudden bursts of traffic. This is because the occupancy of the queue stays closer to a moving average and not to the capacity. Therefore, space typically exists to accommodate the traffic bursts. Another advantage of RED is its ability to alleviate the TCP synchronization issue, a phenomenon associated with TCP sessions. In this phenomenon a massive traffic loss triggers the TCP back-off mechanism due to which all sessions enter the initial state of TCP slow start and then all start to ramp up their congestion windows simultaneously. As a result, for a sufficient amount of time the related link remains under-utilized.

The random nature of packet dropping in RED helps provide a fair resource allocation among the traffic flows to a certain extent. This is because RED drops packet randomly, the probability that a packet is dropped from a particular traffic flow is roughly proportional to that flow's share of bandwidth at that link. As high bandwidth flows send large number of packets to the queue, it provides more candidates for random dropping, thus penalizing them in proportion. A precise fairness, however, cannot be guaranteed with RED.

4.3.3 Tail Drop

The strategy behind 'Tail Drop' is very simple: if queue length is less than maximum threshold g_{max}, enqueue the incoming packet; otherwise drop it. Usually the maximum threshold g_{max} is set equal to the maximum queue capacity. In Figure 4.18(b) the drop probability profile has been shown for the tail drop scheme. It is illustrated that the drop probability p_a is always zero unless the queue occupancy $g$reaches the g_{max} value, in which case the probability jumps straight to 1.

Though this approach has been in use for many years, it has some drawbacks, e.g., TCP synchronization, lockouts, and full queue. The TCP synchronization issue has been explained under the RED discussion previously. The Lockout is a situation when a small fraction of traffic flows receives a large proportion of the bandwidth which leads to an unfair allocation of the link resources. The Full queue indicates very large queue occupancy which causes significantly high end-to-end packet delay and jitter.

4.3.4 Simulation Scenarios

Table 4.6 lists the LTE simulator configuration parameters and user traffic models. It can be noticed that the LTE transport network (S1 link) has been assigned with sufficient capacity to carry offered traffic volume. The only bottleneck for traffic exists at the radio interface in the downlink direction. The mean offered traffic volume at radio interface amounts to approximately 18Mbps. The mean radio interface capacity with random user movements in the simulations is observed to be ≈15Mbps/cell with the instantaneous value peaks going as high as 22Mbps. Higher offered traffic load than the available radio interface capacity gives rise to severe congestion at the radio interface which leads to high PDCP buffer occupancy. In contrast to downlink, the offered traffic load in uplink has been kept fairly small to avoid any congestion at uplink radio interface.

The VoIP traffic is carried over Guaranteed Bit Rate (GBR) bearers and FTP/HTTP traffic makes use of Non-Guaranteed Bit Rate (nGBR) bearers. Owing to the higher priority of GBR over nGBR traffic, the VoIP traffic will not be affected by the congestion. This way, only FTP/HTTP traffic has to suffer from the bottleneck at the radio interface.

In this study two types of traffic mixtures have been considered. Though the amount of offered traffic load is kept the same, the difference lies in the priority assignments of user traffic at the MAC scheduler. Both types of traffic mixtures will be used in analyzing the performance of buffer management schemes. The details of the set of simulation scenarios which use these traffic mixtures are given below.

(a) *Prioritizing HTTP users over FTP users*: Table 4.7 lists the user distribution with respect to the applications. The highest MAC priority (i.e., QCI 1) has been attributed to the VoIP traffic which is carried over GBR bearers. Moreover, HTTP users have been assigned higher MAC priority (QCI 8) compared to that of FTP users (QCI 9). This reflects one of the typical MAC priority assignment schemes for user traffic in the real world. It ensures in-time delivery of delay sensitive VoIP packets and low waiting time of web page downloads at the expense of high waiting time for FTP users.

(b) *Mixing HTTP and FTP users in the same priority class*: In this configuration HTTP and FTP users are mixed together in both priority classes, i.e., QCI 8 & QCI 9. This is another commonly found priority assignment scheme for the user traffic. It mimics the real world scenario where, for example, the premium users are always assigned to a higher priority class compared to the basic users. VoIP users are, however, still kept in the highest priority class of

Table 4.6: Configurations for simulation scenarios of PDCP buffer management.

User Profile Definition	
Number of active users	60 users per cell
Number of cells per eNB	1
FTP traffic model	File size: constant 5MByte Inter-request time: exp(45) sec
VoIP traffic model	GSM EFR codec (12.2kbps) Call length: 90sec Inter-arrival time: exp (50) sec
HTTP traffic model	Number of pages per session: 5 Average page size: constant 100KByte Inter-arrival time: exp (12) sec
User mobility model	Random direction (50km/h)
Network Configuration	
Total number of PRBs	50 (10MHz spectrum)
LTE MAC scheduler	Round Robin
Relative priority of QCI 8 to QCI 9	5:1
User handover	disabled
S1 link capacity	100Mbps (Ethernet 100BaseX)
Simulation run time	2000sec
Discard timer value	120–1300msec
RED parameters	p_{max}=5%, g_{min}=33%, g_{max}=100%
per bearer PDCP buffer capacity	30 – 100KByte

QCI 1 due the strict end-to-end packet delay requirements. Table 4.8 lists the user distribution for such a configuration of the user traffic.

Table 4.7: Distribution of users with respect to applications – Simulation scenario I.

Number of UE/cell – downlink			**Number of UE/cell – uplink**		
QCI 1	*QCI 8*	*QCI 9*	*QCI 1*	*QCI 8*	*QCI 9*
20 VoIP	20 HTTP	14 FTP	20 VoIP	none	6 FTP

Table 4.8: Distribution of users with respect to applications – Simulation scenario II.

Number of UE/cell – downlink			**Number of UE/cell – uplink**		
QCI 1	*QCI 8*	*QCI 9*	*QCI 1*	*QCI 8*	*QCI 9*
20 VoIP	12 HTTP + 6 FTP	8 HTTP + 8 FTP	20 VoIP	2FTP	4 FTP

4.3.5 Simulation Results

The analysis of simulation results can be divided into two parts to address the discussion of two configurations of the user traffic. For each configuration, the performance of the buffer management schemes namely RED, discard timer, and tail drop is studied. The effectiveness of each buffer management scheme is evaluated through various KPIs (Key Performance Indicators). These KPIs include PDCP buffer occupancy, number of packet discards, TCP one way delay, HTTP and FTP download time, number of successful FTP/HTTP sessions etc.

a) Prioritizing HTTP users over FTP user

Figure 4.19 gives an overview of shared PDCP buffer occupancy observed at the eNode-B for a few of the investigated configurations of the buffer management schemes. It can be noticed that when shared PDCP buffer size is virtually unlimited and no buffer management is performed, the occupancy can grow as large as 1.8MByte. This total occupancy is actually the sum of buffer occupancies by all active users. This suggests that maximum PDCP buffer capacity requirements for this particular scenario are 1.8MByte for the given TCP configuration and the limitation of buffer capacity below this value would cause packet drops. It is also evident from the figure that limiting the per bearer buffer capacity using any buffer management scheme, the buffer occupancy can be effectively controlled. It is important to mention that PDCP buffer occupancy value is actually determined by the TCP window size and the number of active parallel TCP sessions.

Table 4.9 shows several simulation statistics including PDCP packet drop ratios. In the tail drop scheme these packet discards are caused by buffer overflow, in RED these discards happen when buffer occupancy exceeds the minimum threshold (g_{min}) value, and in the discard timer these packets are dropped if their waiting time surpasses the configured threshold timer value. The packet drop ratio presented in the table mainly belongs to QCI 9 traffic class. This is because a majority of the discarded packets belong to QCI 9 priority class and a very few discards are seen for QCI 8 traffic. The reason for this behavior is five times higher priority of QCI 8 over the QCI 9 priority class as well as the low buffer capacity demand of QCI 8 traffic due to small sized HTTP pages. VoIP traffic which is mapped to the highest priority of QCI 1 experiences no packet loss in any simulation scenario.

When the PDCP shared buffer capacity is not limited, no packet drop is observed. As soon as the limitation on buffer capacities are imposed the packet drop ratio increases sharply. For example, reducing shared buffer capacity to half from 100 to 50KByte causes the packet drop ratio to increase up to two times, both for tail drop and RED schemes. For the discard timer scheme, 30% decrease in timer

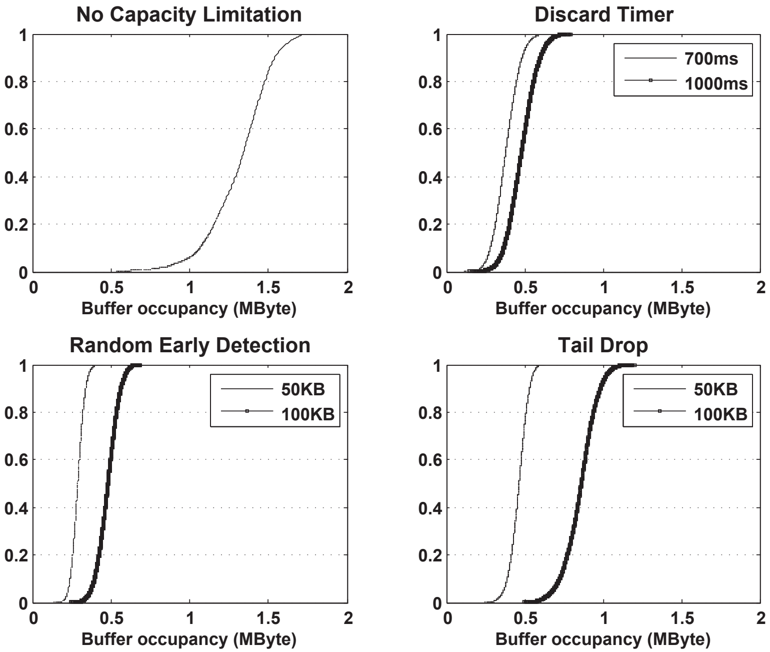


Figure 4.19: The shared PDCP buffer occupancy for different buffer management schemes. The figure shows CDF curves of shared buffer occupancy. In case of discard timer, buffer occupancy has been shown for 700ms and 1000ms timer values. For the RED and tail drop, the buffer occupancy has been shown for per bearer buffer capacity values set as 50KB and 100KB.

value brings about two folds increase in packet drop ratio.

Figure 4.19 shows that RED as the PDCP buffer management scheme can achieve lower PDCP buffer occupancy as compared to that of tail drop for an identical per bearer buffer capacity value. For example, with a per bearer buffer capacity of 100KByte 44% more reduction in total shared PDCP buffer occupancy is observed as compared to the tail drop case. For 50 and 30KByte buffer limitation cases, RED achieves respectively 78% and 85% reduction in total shared PDCP buffer occupancy as compared to the case with unlimited per bearer buffer capacity. On the other hand, the tail drop scheme could only achieve 65% and 80% reduction in total shared buffer occupancy for similar per bearer buffer limitation

Table 4.9: Statistic values of simulation results when prioritizing the HTTP users over the FTP users.

Per bearer PDCP buffer limitation	PDCP buffer occupancy (KB)	PDCP packet drop ratio	TCP one way delay – FTP DL (sec)	FTP file download time (sec)		HTTP page download time (sec)	
	Mean	Mean	Mean	Mean	SD.	Mean	SD.
NoLimit	1320	0.000%	1.497	89.93	57.88	0.66	0.08
Tail drop							
100KByte	857	0.315%	1.020	92.86	58.19	0.65	0.07
50KByte	462	0.701%	0.596	94.41	60.64	0.66	0.09
30KByte	264	1.007%	0.396	92.63	59.51	0.66	0.11
RED							
100KByte	480	0.247%	0.660	89.94	59.68	0.66	0.07
50KByte	290	0.489%	0.426	94.72	58.54	0.78	0.66
30KByte	196	0.867%	0.260	92.62	58.49	0.89	0.93
Discard timer							
1000 msec	481	0.797%	0.497	90.02	55.21	0.66	0.07
700 msec	381	1.289%	0.357	90.50	61.12	0.66	0.07
500 msec	316	1.577%	0.276	93.43	59.21	0.67	0.20
300 msec	210	1.888%	0.165	96.62	62.63	0.66	0.10
200 msec	149	2.341%	0.126	94.37	59.25	0.67	0.37
120 msec	90.1	2.250%	0.084	84.15	78.70	0.66	0.19

values. This is because the tail drop scheme discards packets when the buffer is fully occupied, but RED starts discarding PDCP packets with a certain probability, as soon as the buffer occupancy reaches the minimum threshold g_{min}, i.e., 33%. It is also interesting to note that the performance of RED with 100KByte per bearer buffer capacity is comparable to the tail drop case with 50KByte per bearer buffer capacity, and the discard timer case with 1000msec threshold value. These three cases attain the PDCP buffer occupancy and packet drop ratio in a similar range although their policies of packet discarding are quite different. For example, the tail drop scheme discards incoming packets when the queue is full, discard timer discards older packets in the buffer and RED discards randomly some of the incoming packets when the minimum threshold is reached. Application KPIs will help decide which of these policies is more friendlier to TCP.

Table 4.9 shows several configurations of the discard timer and the corresponding performance values. It is observed that reducing the discard timer value brings lower PDCP buffer occupancy but the corresponding magnitude of the packet drop ratio increases. This high packet drop ratio is expected to aggravate TCP performance. Owing to the fact that one PDCP packet carries one TCP segment in its payload, each PDCP packet discard will make TCP perform a retransmission. The higher the packet loss rate, the more the TCP retransmissions. For each TCP segment loss, TCP has to invoke its ARQ mechanism which reduces TCP throughput and hence increases file download completion time. Furthermore, in situations where the packet drop rate grows very high some TCP connections may not be able to recover leading to connection abort.

Though the limitation of the maximum shared PDCP buffer capacity causes packet drops, it also helps achieve the shorter TCP round trip time. This effect can be seen in Table 4.9 under 'TCP one way delay' statistic. From equation 4.18, it is evident that reducing TCP segment delay provides boost to TCP throughput. Therefore limiting the maximum shared buffer capacity, on the one hand, degrades TCP throughput due to packet losses, but on the other hand, it enhances TCP throughput performance by reducing TCP segment delay. The overall gain or loss in TCP performance is then decided by the combined impact of the two factors. In current simulation scenario, the resulting impact can be seen by the HTTP and FTP file download time as presented in Table 4.9. The best HTTP and FTP application performance is achieved when no limitation is imposed on maximum shared PDCP buffer size. Though buffer management schemes help achieve the low buffer occupancy, no considerable improvement in user QoE perception is observed. RED with 100KByte and discard timer with 1000msec provides the best performance, i.e., achieving 63% reduction in PDCP buffer occupancy without noticeable increase in HTTP/FTP file download time. It can be seen that a 120msec discard timer case attains ≈6 sec reduction in mean FTP download time compared to the baseline case. (Baseline case refers to the scenario with no buffer limitation and no buffer management). However, Figure 4.20 explains the rationale behind this achievement; the number of FTP file downloads has been decreased significantly compared to the baseline case. The reduced number of FTP file downloads is due to TCP connection aborts in response to huge packet discards. This phenomenon can also be seen for other buffer limitation cases, i.e., the number of FTP file downloads decrease along with increase in packet drop ratio. The number of HTTP page downloads are, however, almost identical for all scenarios which is because of high priority of QCI 8 traffic and the fewer associated packet discards.

Simulation results showed that VoIP users experience the best MOS value of 4.5 in all simulation scenarios. The reason behind this is the highest priority (QCI 1) assigned to VoIP traffic and no associated packet discards.

From the KPIs presented in this discussion, it can be concluded that RED with 100KByte buffer limitation and discard timer with 1000msec timer value provides a considerable reduction in PDCP buffer occupancy without significantly affecting

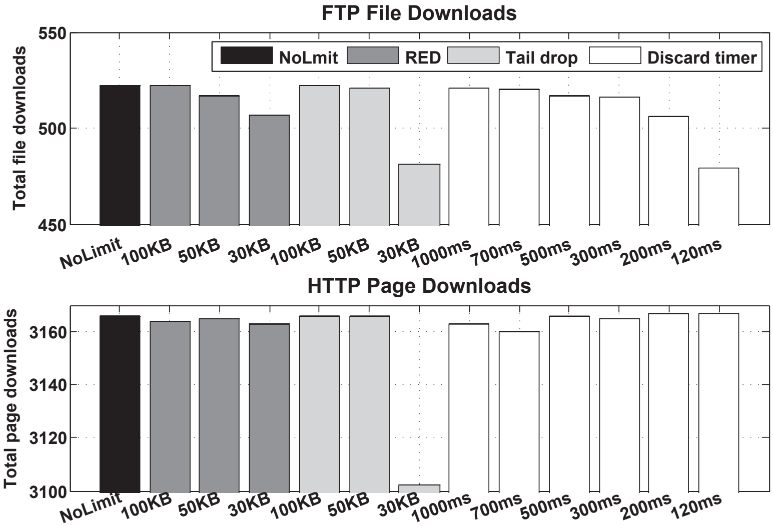


Figure 4.20: Number of successful sessions of FTP and HTTP users in downlink direction. The figure shows total number of completed HTTP and FTP file downloads for each scenario.

HTTP/FTP application performance. If further reduction in PDCP buffer occupancy is desired RED with 50KByte or a discard timer with 700msec & 500msec can be considered. They provide PDCP buffer occupancy reduction at the expense of marginal increase in HTTP/FTP file download time. However, the tail drop scheme failed to provide a good balance of PDCP buffer occupancy reduction and HTTP/FTP application performance as seen for the other two schemes.

b) Mixing HTTP and FTP users in the same priority class

Figure 4.21 shows CDF curves of PDCP buffer occupancy for some configurations of the investigated buffer management schemes. The maximum value of the PDCP buffer occupancy goes as high as 1.7MByte when the buffer capacity is not limited. Using buffer management schemes with appropriate configuration parameters reduces both the mean and maximum buffer usage. However, the effect of these buffer management schemes on end user QoE is to be determined. For this purpose statistical results of simulations have been presented in Table 4.11 & 4.10. Table 4.10 lists the important KPIs belonging to QCI 8 traffic while QCI 9 related simulation results have been shown in Table 4.11. The statistical value of "PDCP buffer occupancy" for all active users in the cell have been listed in Table 4.10 and have been reproduced in Table 4.11 for the ease of reference.

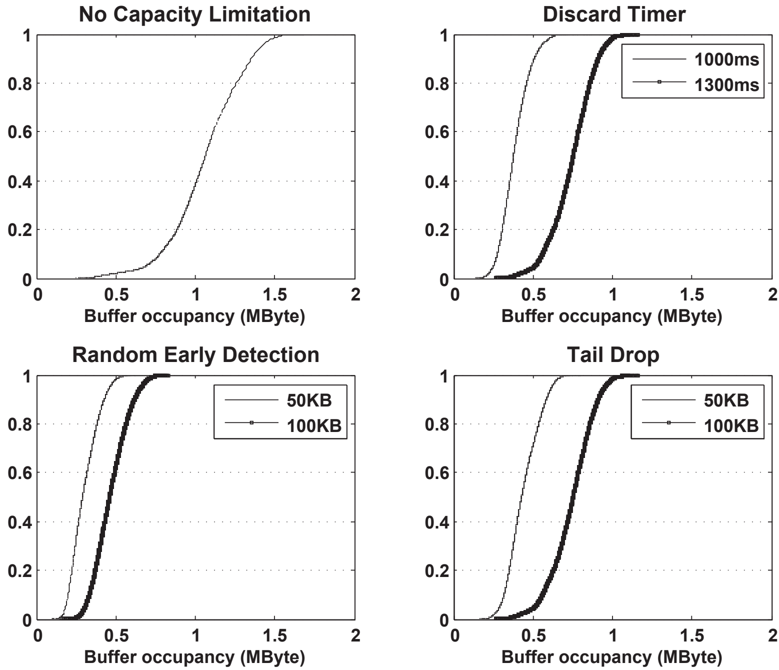


Figure 4.21: The shared PDCP buffer occupancy for different buffer management schemes. The figure shows CDF curves of shared buffer occupancy. In case of discard timer, buffer occupancy has been shown for 1000ms and 1300ms timer values. For the RED and tail drop, the buffer occupancy has been shown for per bearer buffer capacity values set as 50KB and 100KB.

Table 4.10 shows that the impact of buffer management schemes is trivial on the QoE of QCI 8 users. The packet discards are minor until per bearer PDCP buffer space is limited to a very small value. As explained earlier, this is because of higher priority of QCI 8 over QCI 9 traffic. Mean HTTP/FTP file download time values with buffer limitation are very close to that of baseline case. Although the limitation of buffer capacity makes TCP one-way delay shorter but this performance gain is nullified by the associated packet discards. Discard timer scheme

Table 4.10: Statistic values of simulation results – QCI 8.

Per bearer PDCP buffer limitation	PDCP buffer occupancy (KB)	PDCP packet drop ratio	TCP one way delay – FTP DL (sec)	FTP file download time (sec)		HTTP page download time (sec)	
	Mean	Mean	Mean	Mean	SD.	Mean	SD.
NoLimit	1060	0.000%	0.233	19.69	4.58	0.71	0.12
Tail drop							
100KByte	743	0.104%	0.228	19.70	4.22	0.70	0.12
50KByte	440	0.275%	0.156	20.54	4.43	0.72	0.15
30KByte	271	1.598%	0.095	20.89	4.18	0.76	0.54
RED							
100KByte	469	0.000%	0.138	19.70	4.25	0.69	0.10
50KByte	303	0.140%	0.087	20.47	4.24	0.72	0.68
30KByte	220	0.490%	0.064	21.76	3.90	1.02	1.06
Discard timer							
1300 msec	420	0.000%	0.232	19.74	4.41	0.71	0.12
1000 msec	385	0.000%	0.233	19.71	4.72	0.69	0.12
700 msec	317	0.003%	0.237	19.81	4.34	0.70	0.12
500 msec	276	0.008%	0.231	19.86	4.44	0.68	0.09
300 msec	200	0.105%	0.194	19.91	4.66	0.69	0.10
200 msec	147	0.272%	0.141	20.59	4.54	0.70	0.24
120 msec	82.1	0.690%	0.087	19.26	4.67	1.01	1.04

with 1300msec and 1000msec threshold values provides up to 64% reduction in buffer occupancy without tangible impact on HTTP/FTP file download time. RED buffer management scheme, with 100KByte buffer limitation, also performs well by providing 55% reduction in buffer occupancy. However, the tail drop scheme is seen to be the least efficient among the three schemes.

The number of successful HTTP/FTP file downloads for QCI 8 traffic have been shown in Figure 4.22. Excluding the discard timer scheme with 200msec & 120msec threshold configurations the number of file downloads for all cases are identical to one another. Very large packet drop ratios caused by the aforementioned discard timer threshold values make some TCP connections abort and hence the overall count of file downloads decreases.

It was observed in the previous part of the simulation result discussion that the effect of the PDCP buffer limitation is more severe on QCI 9 traffic than on QCI 8 traffic. Therefore, overall system performance must be judged by QCI 9 traffic KPIs. Table 4.11 shows the statistical results of simulations for QCI 9 traffic. As expected the magnitude of both packet drop ratio and “TCP one way delay” is

Table 4.11: Statistic values of simulation results – QCI 9.

Per bearer PDCP buffer limitation	PDCP buffer occupancy (KB)	PDCP packet drop ratio	TCP one way delay – FTP DL (sec)	FTP file download time (sec)		HTTP page download time (sec)	
	Mean	Mean	Mean	Mean	SD.	Mean	SD.
NoLimit	1060	0.000%	3.430	166.69	91.84	5.45	6.31
Tail drop							
100KByte	743	0.383%	2.323	170.01	84.81	5.35	6.11
50KByte	440	0.765%	1.384	175.03	79.24	5.90	7.10
30KByte	271	1.535%	1.003	186.55	83.30	4.81	7.30
RED							
100KByte	469	0.338%	1.560	168.74	72.03	5.64	6.68
50KByte	303	0.665%	1.210	172.09	85.87	5.98	6.33
30KByte	220	1.533%	0.770	193.04	91.16	6.35	7.55
Discard timer							
1300 msec	420	2.221%	0.610	170.16	83.39	5.98	6.07
1000 msec	385	2.439%	0.509	174.23	85.21	6.44	10.5
700 msec	317	2.589%	0.387	165.00	103.0	5.65	5.02
500 msec	276	2.571%	0.221	176.88	101.8	6.50	6.50
300 msec	200	3.090%	0.258	230.22	171.2	4.45	5.07
200 msec	147	3.615%	0.201	236.71	138.8	8.00	71.4
120 msec	82.1	3.248%	0.164	235.54	172.5	6.45	35.2

much higher than that for QCI 8 traffic. This phenomenon leads to long HTTP/FTP file download times. For example, considering the baseline case, the mean download time of HTTP/FTP file is almost 8 times higher for QCI 9 than that of QCI 8 traffic. Another important observation is the magnitude of standard deviation of HTTP/FTP file download time. Due to severe congestion and large number of packet discards these standard deviation values for QCI 9 are higher than that of QCI 8 traffic. In general, small values of the discard timer lead to very high values of standard deviation of file download time. This implies a non-uniform QoE will be perceived by users in the cell, i.e., some users will experience very short download time while others will have to wait longer for the file download. Such a system behavior is not preferred by network operators and, therefore, any parameter settings of buffer management schemes which lead to this consequence should be discouraged.

Comparing the performance of the three buffer management schemes in terms of HTTP/FTP KPIs reveals that RED with 100KByte and discard timer with 1300msec threshold perform optimal in providing significant reduction of PDCP buffer occupancy at the cost of minor increase in HTTP/FTP download time. The tail drop scheme with 100KByte buffer limitation provides identical user QoE as delivered by the other two schemes however it requires up to 43% higher PDCP buffer occupancy. Further reduction in PDCP buffer occupancy can be achieved by RED with 50KByte and discard timer with 1000msec threshold at the expense of up to 4.4% performance degradation in FTP performance compared to baseline case. The number of HTTP/FTP file downloads for RED with 100KByte & 50KByte as well as for discard timer with 1300msec & 1000msec thresholds are seen in the same range as exhibited by the baseline case.

As far as the VoIP users are concerned, they again experience no packet loss in any of the simulation scenarios. Therefore, all VoIP users enjoy the best score of 4.5 as a perceived MOS value in all simulation scenarios.

As conclusion it can be claimed that LTE system having arbitrarily large memory space for PDCP buffers produces very high buffer occupancy which could make system perform suboptimal during the inter eNode-B handovers. On limiting the PDCP buffer capacity without proper buffer management schemes leads to packet drops, i.e., the tail drop phenomenon. According to simulation results, though lower PDCP buffer occupancy is achieved by the tail drop scheme, it severely degrades user application performance. On the other hand, it is observed that when the RED & discard timer buffer management schemes are used, a significant reduction in buffer occupancy is achieved without tangible effect on user QoE.

The optimal configuration parameters of RED and the discard timer schemes depend on several factors, e.g., the user application type, traffic mixture and congestion level in the system. This fact has also been observed in other investigations related to this study [U. 11b] [U. 11a] [U. 12a]. In such circumstances although a single optimal configuration for a buffer management scheme cannot be provided but a range of feasible values can be specified to facilitate the tuning of these parameters in achieving close-to-optimal system performance. The study of simulation results with different offered traffic loads and traffic mixtures suggests two ranges for discard timer thresholds, i.e., 700–1300msec and 300–700msec for scenarios with high radio interface congestion and slight radio interface congestion, respectively. The optimal performance for RED buffer management schemes can be realized by a capacity limitation in the range of 50–100KByte.

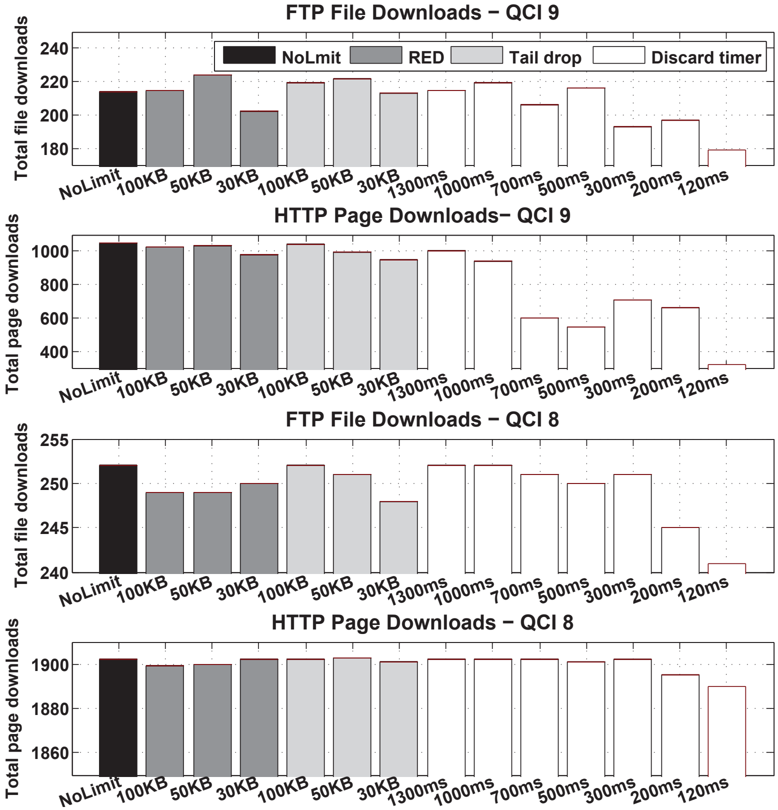


Figure 4.22: Number of successful sessions of FTP and HTTP users in the downlink direction. The figure shows total number of completed file downloads for each scenario.

5 User QoE Enhancement using Multihoming

This chapter highlights the importance of multihoming in wireless heterogeneous network to enhance user QoE and improve the network performance. It discusses state-of-the-art multihoming solutions and stresses the need of a traffic flow management mechanism to control bandwidth resources of the integrated networks. Based on the requirements of multihoming in heterogenous networks, a comprehensive flow management architecture is also developed which is compatible with 3GPP proposed SAE architecture. In addition, several mechanisms are proposed to facilitate the bandwidth resource management of multihomed users.

5.1 Multihoming

The term ‘Multihoming’ refers to a node with more than one attachment point to the network. Multihoming is realized either through the configuration of multiple IP addresses on a single network interface of a node, or more commonly, by installing multiple network interfaces on a single node each assigned with an IP address. Traditionally, the use of multihoming was desired to add reliability and redundancy to the network connection to ensure continuous operation during connectivity outages or other network failures. While increased resilience and availability still remains the primary objective of multihoming, an increasing interest is being observed in exploiting other benefits from multiple network connections. In particular, multihoming can be leveraged for improving the performance and capacity of wide-area networks, lowering bandwidth costs, and optimizing end user QoE.

Multihoming can be implemented at host or site level. A host with two or more independent connections to the Internet is called a multihomed host. These connections may or may not belong to the same Internet Service Provider (ISP). Typically, a multihomed host is capable of detecting connection failures and moves the established communications from the failed path to one of the other working paths. In addition, the multihomed host can also make use of available network paths based on a certain policy, e.g., to perform load balancing. This be will discussed in further detail in Section 5.2. In a manner similar to multihomed hosts, a site can also maintain two or more independent connections to the Internet and this called a multihomed site. A 'site' in this context is an entity autonomously operating a network using IP, and in particular, determining the addressing plan and routing policy for that network [ABG03]. A multihomed site makes use of multihoming service to guarantee fault tolerant and reliable connectivity to its host.

5.1.1 State-of-the-art

A measurement based analysis to quantify benefits of multihoming by Akella et. al. [A. 03] reveals that a potential performance gain beyond 40% can be achieved by employing properly planned multihoming. This multitude of potential benefits has kept multihoming a research subject during the past years. The main hurdles in realizing multihoming are proper routing, load balancing across multiple paths, and to maintain TCP/UDP sessions through cut-overs. Though a wide range of solutions were proposed, scalability and avoiding huge routing table have been the main concerns in this area. The identifier-locator separation techniques are widely assumed to be a solution to such problems and have been greatly explored in the proposed solutions.

Identifier-locator separation can be implemented in several different ways, e.g., splitting the IP address space into two portions where one portion represents end-host identifiers and the other portion is used as wide-area locators. Hosts use identifiers as source and destination addresses in the packets, and the border routers encapsulate these packets with an outer header which contains locators. This approach is generically called 'map-and-encapsulate'. Other proposed schemes include geographically based address prefixes, transport protocols with multihoming support (e.g., Stream Control Protocol (SCTP) [R. 07], Multipath TCP (MPTCP) [FRHB12] etc.), and introducing an additional level of identifier above the IP address like HIP (Host Identity Protocol) [MN06].

As far as standardization is concerned, a large number of proposals have been under discussion to cover different classes of solutions. After a long review process, SHIM6 (Site Multihoming by IPv6 Intermediation) [NB09] came out as a standard solution for multihoming in IPv6 networks. Other mentionable proposals which are still active include LISP (Locator/ID Separation Protocol) [FFML12], ILNP (Identifier Locator Network Protocol) [AB12], MPTCP, NAT66 (IPv6-to-IPv6 NAT) [WB11], and HIP. Over the next paragraphs some of the aforementioned multihoming proposals are described briefly.

SHIM6 is a host-centric solution which inserts a ‘shim’ on top of the IP routing sub-layer and beneath the IP endpoint sub-layer. In this scheme IPv6 addresses are used both as identifier and locator. The IPv6 address which is used to initialize the connection, plays the role of identifer during the whole communication life. This identifier is called ULID (Upper Layer ID) and is used by the shim layer in performing mapping to locators. The failure detection and recovery process of SHIM6 remains independent and transparent to higher protocol layers.

HIP is another host-based solution which makes use of the identity/locator split approach and offers end-to-end mobility and multihoming. HIP inherits some security features by employing the public key component of the private-public key pair as the host identifier. In networks that implement HIP, all occurrences of IP addresses in applications are replaced with the host identifier. This results in decoupling of the transport layer from the Internet layer in TCP/IP that allows a mobile host to preserve its transport layer connections upon movement. The host identity can also be used for looking up the current location of a host because it is supposed to be a long-term identifier.

MPTCP improves resource utilization and failure tolerance by using multiple simultaneous paths between multihomed peers while still maintaining the backward compatibility with the traditional TCP. MPTCP can be considered as an add-on set of features on top of TCP which starts like regular TCP but if extra paths exist, additional TCP connections are created. Though MPTCP distributes the traffic load between working paths using TCP-like mechanisms, to an application layer it appears to be a single TCP connection.

SCTP is a message oriented, reliable transport protocol with inherent support for multihoming. In SCTP one of the paths is selected as the primary path and the rest become secondary paths. In case the primary path fails for whatever reason, a secondary path is chosen and utilized. When the primary path becomes available again, the communication can be moved back without the application being aware of any issue. In addition, SCTP also allows multiple simultaneous data streams within a single connection or association. For example, web page images can be transmitted together with the web page text.

5.1.2 Relation to Mobility Management

Multihoming and mobility management are closely related and can be used in a complementary fashion. Mobility management refers to a functionality which allows a mobile user terminal to maintain the same IP address as the terminal changes its attachment from one network to another network. In other words, mobility concerns redirection to previously unknown IP addresses while multihoming describes a node's ability to redirect packets between multiple IP addresses that it has configured simultaneously [ORCTV].

There are two mobility management approaches: 'reactive mobility management' and 'proactive mobility management'. Reactive mobility management is considered as a response to link layer handover. Though the sophisticated protocols based on the reactive approach claim to reduce performance degradation during the handover, there is an inherent minimum latency for the user traffic to be redirected to the new network attachment point. This is because it always takes one round trip propagation time to register a new IP address with the mobility management entity in the network and get the first redirected packets at the new IP address. During this process, the packets in flight toward the old IP address are lost.

The proactive mobility management approach substantially improves the handover performance by anticipating the imminent handover and preparing for it at the right time. It requires a user terminal to monitor related link layer characteristics of a network connection to foresee impending handover, obtain a new IP address from the target access network, and register it with the mobility management entity in the network before initiating a link layer handover. In this way, a 'make-before-break' strategy is followed which prevents excessive delays and packet losses during the handover. The cross-layer interaction and network information retrieval requirements of both proactive and reactive approaches can be satisfied using the IEEE 802.21 standard for Media Independent Handovers or other mechanisms like [U. 07a] developed by the author.

5.1.3 Selected Multihoming Solution

A number of research studies can be found making use of cross-layer techniques and soft handover to optimize handover cost in terms of packet delay and loss in heterogeneous networks. For example, Song and Jamalipour [SJ05] describes an intelligent scheme of vertical handover decisions in selecting the best handover target from the several candidate heterogeneous networks. Several other proposals have been made to improve the performance of cellular and 802.11 networks. Song et. al. [W. 07],[SJZS06] has discussed admission control schemes to improve the performance of integrated networks. Fei and Vikram [YK07] proposes a service differentiated admission control scheme based on semi-Markov chain which is although very accurate but has high computational complexity. [SZ05] provides an efficient alternative based on moment generating function but at the price of accuracy. Similarly, Zhai et. al. [H. 05] has shown that by controlling the collision probability with the help of input traffic rate of users, the maximum throughput can be achieved by keeping 802.11 network in non-saturated state. Other studies are focused on developing solutions for load balancing in the integrated heterogenous networks. Such a proposal can be found in [LS03], [SZC07] where policy based load balancing framework has been presented to effectively utilize the aggregated resources of loosely coupled cellular/WLAN network. In contrast of these studies, the goal of this work to explore the practical limits of achievable performance in a heterogeneous network scenario. For this purpose, cross-layer techniques are employed in order to go down to the MAC layer functionalities of involved access technologies. Through a coordination of IP and MAC layers, this work aims is to maximize the spectral efficiency of network bandwidth resources and fulfill the application QoS requirements at the same time. The proposed solution not only adapts to dynamic load conditions of the access networks but also conforms to time varying channel conditions of the mobile users. Considering the aforementioned factors into account, here the capabilities of user multihoming are exploited in order to achieve system wide optimized performance and improved user QoE.

In the context of this research work where user multihoming should be assisted by the network entities and the correspondent hosts are oblivious of user mobility, a simplified solution is desirable. Preferably, the proposed solution should make use of already existing mobility management functions and provide the multihoming capabilities as an add-on. For example, consider the reference network architecture presented in Chapter 3 which is based on SAE standards for integration of LTE and WLAN networks. In such a heterogeneous network, Mobile IP has already been chosen as a standard mobility management protocol. Though 3GPP has not yet standardized multihoming support in heterogeneous networks, a solution based on a natural extension of Mobile IP already exists.

Mobile IP, in its pure form, delivers mobility service to a user terminal moving from one visited network to another. Even if the terminal has multiple active network interfaces, only one of them has to be chosen to work with Mobile IP. This restriction is lifted by the Multiple Care-of Address (MCoA) [WDT$^+$09] extension to Mobile IPv6 which enables a user terminal to register all of its active network interface addresses as care-of addresses with its home agent. However, for communication purposes only one of these care-of addresses is used and the rest are considered as backup. In this way, the MCoA extension introduces a limited multihoming support on top of Mobile IPv6 while maintaining backward compatibility. This support is further enhanced by another IETF proposed standard and extension to MCoA [G. 10], which allows a mobile terminal to use all of its network associations simultaneously. This level of multihoming support fulfills the prerequisite to perform adaptive flow management operation as discussed in Section 5.2.

Based on the scenario configuration in host-based mobility management, an implementation of Mobile IP and the aforementioned extensions may be required both at the mobile terminal and at its correspondent host (e.g., when using the 'route optimization' option of mobile IPv6). However, in this work the route optimization option is not the focus, therefore, it is mandatory only for the home agent and mobile terminals to implement the mobility management and multihoming functionality.

At this point, it is worth mentioning that the use of other multihoming approaches (discussed in Section 5.1.1) in conjunction with Mobile IP should not be ruled out. There are several proposals about complementing Mobile IP with well-known multihoming approaches, e.g., see [BGMA07]. On concept level these proposals are completely in line with the discussions and findings made in this work.

5.2 Flow Management

A multihomed user is assumed to make efficient use of aggregated bandwidth resources available from its multiple network attachments. This particular task is discussed under the subject of Flow Management. In flow management, different traffic flows are directed to different network interfaces based on a certain set of policies. Such a policy has generally a wide scope encompassing, e.g., QoS requirements of user applications, service costs of access networks, traffic load balancing, network path security etc. The execution of these flow management policies relies on a set of traffic flow handling options which are described as follows:

- **Flow distribution**: In this option a particular traffic flow is assigned to a certain network attachment or path based on the associated policy. For example, TCP based traffic flows, like FTP file or email download do not have stringent QoS demands and, therefore, may be directed to a WLAN path while QoS sensitive realtime applications like, VoIP may use the LTE network.

- **Flow splitting**: In flow splitting, the packets belonging to one large traffic flow are distributed among the different network paths in order to speed up the transmission using aggregated bandwidth. In this case, the receiver of traffic flows or some other entity in the network must be responsible for reordering the packets received over multiple paths. An example could be a user watching a HD video stream of a football match who distributes the traffic flow over the WLAN and HSDPA network paths.

- **Flow multi-casting**: In order to add redundancy and reduce overall transmission errors, a single traffic flow may be duplicated over multiple network paths so that each path carries the complete traffic flow. This is performed by multi-casting a traffic flow to multiple global IP addresses of the receiving host. This option is useful, to reduce overall transmission errors through added redundancy, in a scenario where a mobile terminal has multiple attachments to wireless access networks with high bit error rate.
- **Flow dropping**: In a certain situation where a mobile terminal lacks sufficient bandwidth resources, a less important traffic flow may be discarded in the access network instead of forwarding it to the terminal. For example, a video call can be transformed into a audio call when the access link quality is not good enough.

In the context of aforementioned reference architecture for heterogeneous networks where multihoming is realized using Mobile IPv6 and its extensions, the execution entity of the flow management function should be logically located at the home agent for downlink traffic and at the user terminal for uplink traffic. This execution entity translates the flow management policy into requests supported by Mobile IPv6 and its extensions. In turn, the respected functions of Mobile IPv6 enforce the required traffic handling to achieve the desired effects. Figure 5.1 depicts the above described traffic handling options of flow management in the context of Mobile IPv6.

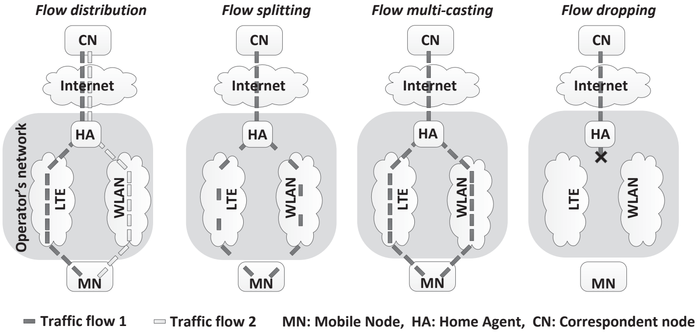


Figure 5.1: Different options of traffic flow handling when performing flow management in heterogeneous networks.

Flow management has the potential to achieve a multitude of network performance and management gains, a few of which are described in the following.

- Flow management allows a mobile terminal to exploit the bandwidth of available network paths. More importantly, it can utilize the specific access technology links in accordance with their characteristics to get maximum benefit out of them. This feature is the main topic of discussion in this work.
- Flow management can be oriented to add reliability and redundancy to a traffic flow by duplicating the data packets over two or more network paths so that better QoE can be achieved even in bad channel conditions.
- Flow management can cut the monetary costs of access network usage for a mobile user, e.g., the users can devise a policy to download email attachments or podcasts only through a free hotspot access in order to reduce the usage of expensive wide-area network access.
- Flow management can also help enforce security measures both for the users and network operators. A simple policy can direct security sensitive data over a trusted link so that additional encryption (e.g., through the IPSec protocol) is not required to secure the transmission.
- Today, the Internet cloud applications offer high quality contents (e.g., HD movie rentals through on-demand video streaming, apps with large databases, cloud data storage etc.) which, in turn, demand high QoS specifications. Sometimes a mobile terminal cannot meet these QoS requirements for different reasons. In such situations, flow management can bundle together the available bandwidth resources from several network paths to provide an application with a service of required QoS.
- Flow management offers an effective tool to perform network offloading or load balancing of the network traffic.
- Flow management allows network operators to optimally assign bandwidth resources to their users in order to improve their QoE as well as to enhance overall network capacity.

It is common practice to categorize flow management functions based on the fact whether the policies/decisions are made by the network operator or by the end user. In ‘network-centric flow management’, the network operator alone is in-charge of bandwidth resource management operations performed using the policies of flow management. Though the end users can inform the network operator about their preferences, the final decision has to come from the network operator. Usually, such flow management policies or decisions are dynamically derived based on user and network operator’s preferences as well as the measurements collected at different metering points in the network and at the user terminal. Such measurements may include the traffic load in a certain part of a network, uplink/downlink channel conditions for a user, number of users attached to the base station or access point, buffer occupancy of the router queue, available battery power of a user terminal, geographical location of a user etc.

It is also possible that an end user solely controls the flow management operations. This is called ‘user-centric flow management’. In this case, the user terminal may request certain information from the associated networks which may facilitate dynamic flow management policies. It is completely up to the network operator’s disposal which kind of information can be offered to the end user. Although user-centric flow management gives end users complete freedom to manage their bandwidth resources but due to the lack of network information this might not be an optimal solution. For example, consider users who want to switch their ongoing video call from the LTE network to a freely available WLAN network. But they are not sure whether the WLAN can deliver the QoS required for this call due to the fact that the load on that network and its capacity is unknown. In contrast to this, network-centric flow management can exploit wider information including that which a network operator does not want to publicize. Therefore, network-centric flow management has the potential to perform optimal resource utilization of bandwidth resources available to an end user. This would create a win-win situation both for end users whose preferences are considered in flow management decisions, and for the operators who manage to satisfy their customer’s requirements with improved resource utilization. Owing to these advantages, this work focuses on enhancing user QoE by using network-centric flow management. The next section lays the foundations of a comprehensive system architecture which can be overlaid on the SAE architecture to support both user-centric and network-centric flow management.

5.3 Flow Management System Architecture

An extension to the SAE architecture for the integration of heterogeneous access technologies has been discussed in Section 5.1.3. The extension was intended to enable multihoming support for users. The evolution is continued and another add-on is being presented in this section. The proposed architecture has been developed under the framework of the SAIL project[Sp13][Ser13] and brings flow management capabilities to the network as explained in Section 5.2. Once the flow management is realized, this work will make effective use of its features, e.g., to offload a congested network, to improve network resource utilization by exploiting the user and channel diversity, and to enhance the end user QoE. Considering the time-varying channel quality of access links and user traffic demands, the proposed system is designed to react and adapt to these variations in order to continuously deliver optimized performance. In general, most of the actions envisaged within the scope of this system architecture can be characterized in three basic phases, i.e.,

1. Collection of information from network entities and the user terminal.
2. Taking appropriate decisions based on the collected information.
3. Enforcing the decisions by instantiating suitable mechanisms.

Although these three phases, most of the time, are invoked iteratively, but it is also possible for them to follow a different pattern depending on the collected information and the outcome of decision processes.

5.3.1 Functional Entities

Based on the above identified three phases, this system architecture defines three functional entities each of which is dedicated to an individual action phase. These entities are assumed to be independent of, and abstracted from OSI layers or any protocol. This component-based system architecture is easy to develop and offers great flexibility in terms of deployment and integration with most of the existing network architectures. The functional entities are described as follows:

- **Information Management Entity (IE)**: This entity is used to collect useful information which is required to make important decisions. This information may include traffic load on a certain network link, signal strength at the user-terminal or base station, buffer occupancy of a router queue etc. In addition to technical parameters, this entity can also offer other dynamic information, e.g., the user preferences, user's feedback about QoE etc. In order to keep the implementation of IEs simple, they are not assumed to be intelligent; rather they can only perform straightforward processing tasks on the gathered information like filtering, aggregation, abstraction etc. The IEs can be either implemented in dedicated devices such as meters, or they can be hosted on existing network elements like routers, access points, base-stations, gateways, user-terminals etc. Decision making entities can retrieve the required piece of information from an IE by sending a direct request or by subscribing to it. In case of subscription, the IE sends the particular information automatically to the subscribed decision making entity based on the type of subscription, e.g., on the occurrence of an event, periodical transmission of certain pieces of information etc.

- **Decision Making Entity (DE)**: A DE is the most intelligent part of this system architecture. It makes use of information available from the IEs to take a decision in accordance with pre-defined policies. Examples of such decisions are: association to a certain access network, vertical handover hints, change in a service treatment, grant or deny user access to a service/network etc. Typically, in 3GPP networks, the decisions are taken in one centralized location in the network. However the possibility of a distributed decision mechanism cannot be ruled out. A DE undertakes policies from user/operator preferences about the QoE, security, costs, QoS guarantees, network resource allocation etc.

- **Execution and Enforcement Entity (EE)**: The decisions made by a decision making entity are conveyed to the relevant EE for execution and enforcement. Although in some cases DE and EE may be hosted on the same device, but generally EEs are more distributed in the network to facilitate the execution of a decision involving several network elements, e.g., a handover has to be performed in collaboration with access points, routers, database servers, user-terminal etc.

The provision of allowing each of these functional entities to be placed on one or distributed over serval network elements enables to support different configurations, topologies, and scenarios. Moreover, as some of the functionalities have to be implemented at different layers, the architecture facilitates the use of modern cross-layer optimization techniques.

5.3.2 Inter-Entity Communication

Table C.1 lists a number of control interfaces designed for inter-entity communication purposes. With help of these bi-directional interfaces an entity can communicate with any other entity of the same or different type. However, a direct interface between IE and EE entities has not been foreseen which is because of their different functional designations. Moreover, in order to enable communication with external functional entities (e.g., different OSI layers, mobility management function, user interface etc.) a control interface (O_{EXT}) has been proposed.

Pertaining to system performance demands, ease-of-deployment objectives, or requirements to implement certain functionality, it is possible to distribute functional entities across several network elements so that their inter-communication

Table 5.1: Control interfaces for inter-entity communication.

Interface	Description
O_{DD}	Interface between two Decision Making entities
O_{II}	Interface between two Information Management entities
O_{EE}	Interface between two Execution & Enforcement entities
O_{DI}	Interface between Decision making entity and Information Management entity
O_{DE}	Interface between Decision Making entity and Execution & Enforcement entity
O_{EXT}	Control interface to an external functional entity

is realized using the above mentioned interfaces. In this case, the information exchange between different network elements hosting these functional entities must be coordinated using an Inter-Node Communication (INC) function. Some functionalities of the INC comprise the compatibility identification of the communicating entities, the possible conversion of inter-node messages, generic security/authentication services etc.

The O_{DI} interface between DE and IE entities comprises the following messages:

- **Configure IE request**: This message type will be used by a DE to configure the operation of an IE as well as to subscribe to necessary pieces of information. This includes various performance parameter notifications and the procedures to collect that information, e.g., mean value of a user's channel quality indicator or peak traffic load on a certain link in core network, etc. As part of the subscription request, this message contains either the time period value during which the requested piece of information is collected and sent to the subscriber or the description of an event (e.g., a threshold value) which triggers the transmission of information. This is a proactive way of gathering information. This message may also be used to configure an IE to send certain information to another IE.

- **Configure IE response**: This is response of the IE to a 'Configure IE request' message which contains the status of the requested operation.

- **Information request**: This is a reactive way of requesting a certain piece of information to which IE should respond immediately.

- **Information notification**: This message is used by an IE to send the requested information to the DE or another IE. This information may come as a response to an 'Information request' or subscription performed by 'Configure IE request'.

- **Notification response**: This is an optional acknowledgement message which may be sent by the receiver of the 'Information notification' message.

The interface between DE and EE (O_{DE}) is used to exchange two types of messages described in the following.

- **Execution request**: A DE uses this message type to covey its decision to an EE in order to execute and enforce it. Typical examples of these decisions are: modifications in routing table, connection attempt to an access network etc.

- **Execution response**: This message is sent by the EE to inform the DE about the status of the requested operation.

The interface between peer IEs comprises **Information Exchange** messages which may be used in special situations where a coordination is required between various IEs in order to collect and send the requested information to a DE. Similarly, another message which is exchanged by peer EEs over the (O_{EE}) interface is called **Execution Exchange**. This message is intended to help EEs enforce a particular decision of the DE which requires performing certain actions at more than one point in the network.

5.3.2.1 Information required by DE

In general information needed by a DE to make a decision can be structured as follows:

- **Resources**: There is two types of resources, i.e., network resources and user terminal resources. Network resources are essentially described in term of attributes which reflect the capacities and capabilities of the nodes and links. Such attributes of a network are specific to its technology and composition. Examples of such attributes are bandwidth capacity of link, noise rise and HS-DSCH codes in an access HSPA network, number of PRBs in LTE access network, data rate available from a WLAN access point etc. Likewise, user terminal resources may be described by attributes, e.g., battery power, available network interfaces etc.

- **Context**: It refers to relevant constraints for a decision process, e.g., geographical location, type of service, time of day etc. A context-aware decision is expected to help optimized system performance.

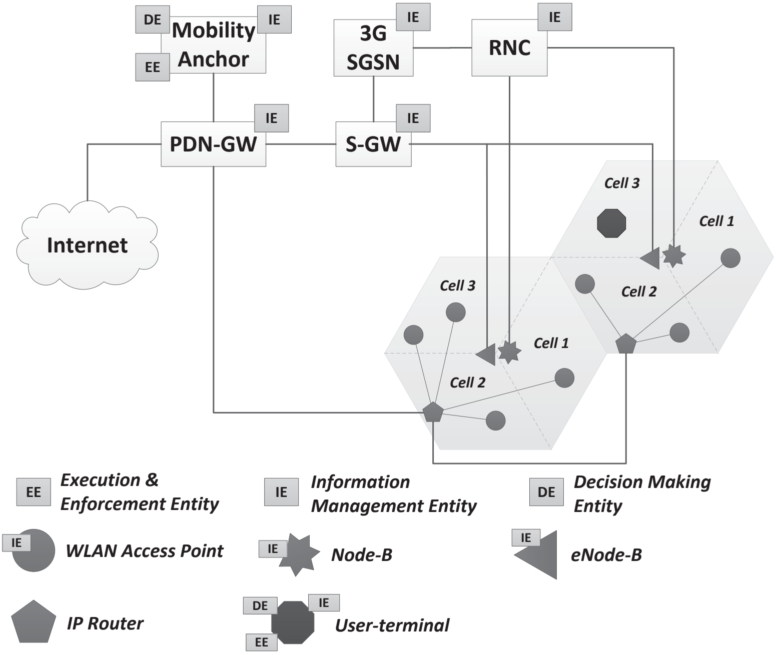


Figure 5.2: An example of flow management architecture overlaid on 3GPP SAE architecture.

- **Requirements**: The requirements can be from the perspective of a user or an application. From an application point of view such requirements can be related QoS demands like minimum throughput, maximum end-to-end packet delay, jitter, losses etc. Likewise, from user perspective these requirements can be related to service cost or pricing and expected QoE.

- **Policies**: The policies driving a decision can either be from a user's or an operator's viewpoint. The users describe their preferences in terms of policies, e.g., save battery power, minimize the monetary cost of a service, preferred interface for an application etc. From the operator's perspective policies can be used to perform load balancing in the networks, to select the suitable network access for user to fulfill application QoS demands, to minimize the network operation cost or energy consumption etc.

5.3.3 Overlaying on SAE Architecture

Figure 5.2 shows an example, how flow management functional entities can be hosted on network elements of the SAE architecture. The network shown in the figure integrates heterogeneous access technologies both from 3GPP (i.e., UMTS/HSPA and LTE) and non-3GPP (i.e., IEEE 802.11 or WLAN) standards. The complete network is owned by one network operator who controls the operation of all access technologies. The base stations of the aforementioned two 3GPP access technologies co-exist at a certain site and individually serve three cells of that site. The access coverage of WLAN APs is distributed in the area and overlaps with the coverage of 3GPP access technologies. The geographical location of these WLAN APs is decided by the network operator as a part of the network planning task and is beyond the scope of this work. However, the main purpose of these WLAN APs is to provide a means for traffic offload. Owing to the fact that today's mobile devices can simultaneously connect to WLAN as well as to one of the 3GPP access technologies; this creates a scenario for users to benefit from multihoming when being in the coverage of WLAN APs.

The Mobility Anchor (MA) is responsible for mobility management of the users using one of the 3GPP supported protocols such as DSMIPv6, Proxy MIPv6 etc. As a part of mobility management, the MA is required to act as the intermediate destination through which all multihomed user traffic has to pass. This provides MA with a possibility to control downlink traffic on each network path of a multihomed user. This also makes the MA the most suitable network element to execute downlink flow management decisions by hosting an EE entity on it. A DE entity is also being hosted on the MA which will be called DE_n for ease of reference. All network-centric flow management decisions are mainly taken by the DE_n entity. In principle, DE_n can be located anywhere in the network, but in order to avoid additional signaling traffic and involved delays, hosting it on the MA is highly recommended. In addition to making decisions for network-centric flow management, DE_n may also assist user-centric flow management operations.

Another DE entity is hosted on the user-terminal that is in charge of user-centric flow management in addition to assisting network-centric flow management operations. This DE entity will be referred to as DE_u hereafter. As all uplink flows originate from the user-terminal, it is an ideal place to control uplink traffic on various network paths of a multihomed user. This also justifies the hosting of an EE entity on the user terminal with the help of which flow management decisions related to uplink traffic are enforced.

Other than the above described DE and EE entities, almost all network elements host one IE entity, e.g., at base stations, WLAN APs, gateways, and RNCs. The user terminal also hosts an IE entity. All of these IE entities gather the pieces of information required by the DEs. Typically, DEs subscribe to IEs for certain information which is transmitted periodically.

An important aspect of network-centric flow management is the network resource grouping. As 3GPP access technologies manage their network resources per cell basis, therefore flow management also aggregates all network bandwidth resources available in a cell both from WLAN and 3GPP access technologies when performing resource allocation. The resource allocation process will be further explained later in this chapter and also in Chapter 6.

5.3.4 Flow Management Architecture Implementation for Simulator

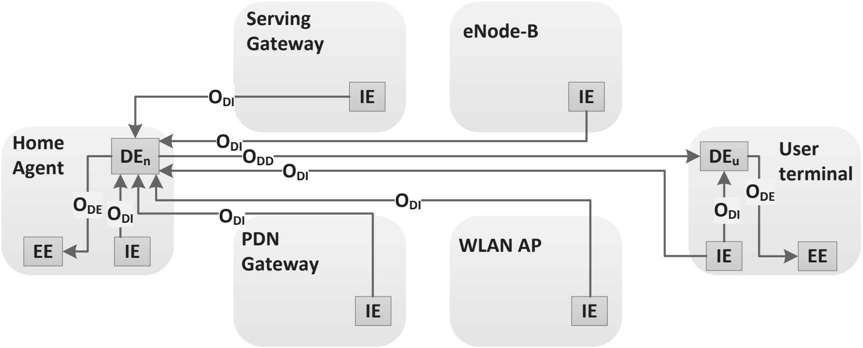


Figure 5.3: Network-centric flow management architecture for the heterogenous network simulator developed in Chapter 3.

Figure 5.3 presents the network-centric flow management architecture to be used in conjunction with the heterogeneous network simulator where LTE and WLAN access networks are integrated together as per 3GPP standard. The DE_n entity takes all flow management decisions. These decisions are executed locally using the O_{DE} interface to EE entity. The decisions to be executed at user terminal are propagated via the O_{DD} interface to the DE_u which enforce them using the local EE entity.

Within the focus of the heterogeneous network simulator, the IE entity at the user terminal is basically used to obtain user preferences and QoS related information. The IE entities in the network are mainly used to retrieve the measurements of packet delays and losses in the transport network, i.e., on the Ethernet links which connect the PDN-GW (or S-GW) with the eNode-B and WLAN APs. An accurate measurement of packet loss in the network is performed with the help of the GTP protocol [3GP08]. The GTP protocol is used to tunnel traffic from PDN-GW to eNode-B and WLAN APs. GTP packet header has a 16-bit field that uniquely identifies this packet and allows detection of loss. Moreover, packet delays for both uplink and downlink traffic are measured using the One Way Active Measurement Protocol (OWAMP) [STK$^+$06].

OWAMP has been designed as a high precision mechanism to measure one-way delay in networks. In OWAMP, small test packets are sent from the sender to the receiver. The test packet carries a sequence number and a time-stamp to reflect packet sent time. At the receiving end, one-way delay is computed from the difference of sent time and the receive time of the test packet. It is clear that the operation of OWAMP requires that clocks of both the sender and the receiver to be synchronized. A very accurate time source can be made available to hosts participating in OWAMP operation using the Global Positioning System (GPS) (accurate to approx. 10ns), CDMA-based time sources (accurate to approx. 10μs), or through the Network Time Protocol (NTP) primary time servers (accurate to approx. 1ms). The inaccuracy within OWAMP measurements itself is estimated to be in the range of 55–60μs.

OWAMP is also called ‘one-way ping’ in contrast to standard ping which provides round trip delay. However, the use of OWAMP is more favored as it provides more insights by measuring uplink and downlink packet delays separately. Such information can be used to tune performance of applications which rely on round trip time (e.g., TCP) and also those rely on one-way delay (e.g., video streaming).

In addition to delay and loss assessments, per bearer PDCP buffer occupancy is measured by the IE entity hosted on eNode-B and an overall MAC buffer occupancy by the IE entity hosted on the WLAN AP. This provides an indication of downlink radio interface congestion for the users of both access technologies. In the core network, the IE entities hosted on the S-GW and the PDN-GW provides router queue buffer occupancies to help estimate downlink transport network congestion. Similar measurements are also performed at the uplink transport network interfaces of the eNode-B and WLAN AP.

The algorithms and policies used by the DE_n in making decisions will be discussed in more detail in the next sections of this chapter. Typical examples of these decisions are: when a particular user terminal should attach or de-attach to WLAN access network and how much traffic should be directed to each network path for uplink and downlink communication of a multihomed user. The decisions related to the association with the WLAN access network are executed at the user terminal via the DE_u entity. The decisions regarding traffic distribution have to be executed both at home agent and at user terminal.

5.4 Downlink Flow Management

The most important piece of information in deciding appropriate network path(s) for a multihomed user is the knowledge of user application demands and available capacity of access links. Based on this knowledge, the DE_n can make efficient use of available resources following any policy of resource utilization. The estimation of user access link capacity is therefore an important task which should be executed with the greatest possible accuracy. The higher the precision of capacity estimation, the more efficient will be the resource utilization. Alternatively, without estimation of link capacity, loading it with an arbitrary amount of user traffic would either lead to link under-utilization causing wastage of resources or over-utilization which causes excessive queuing delays and buffer overflows. Both of these situations will result in user QoE degradation. In this section, a few methods for access link capacity are devised both for WLAN and LTE access technologies in downlink.

5.4.1 Capacity Estimation of WLAN Access Link

Legacy WLAN (IEEE 802.11 a/b/g) provides no QoS when scheduling user traffic. Essentially, there is only a single queue in a WLAN AP where all incoming traffic is received, held, and then transmitted over the air to the users in a “First Come First Serve” (FCFS) manner. That is why overall throughput of a WLAN AP and that of the users being served is highly variable based on the number of active users in the system, their offered traffic load as well as their channel conditions. In order to estimate the link capacity of the user, the most commonly used techniques require data traffic to flow between the user terminal and WLAN AP. Using test data flows for this purpose causes bandwidth overheads. And if this measurement has to be based on actual user traffic flows, it brings two disadvantages. First, this cannot be employed for a user terminal that has just attached to a WLAN AP and has not yet received any data. Second, the variations in link capacity cannot be captured unless user traffic floods the link, e.g., with TCP based FTP file downloads.

This work proposes a novel way of managing WLAN bandwidth resources in an efficient way which also provides an accurate estimation of downlink user capacity. This approach relies on the following pieces of information to operate: number of active users attached to the WLAN access point and their PHY data rate at a particular time instance. This information is always available at a WLAN AP and can be accessed via a hosted IE entity. Assume there are N active users attached to a WLAN AP which are being served in a round-robin manner in downlink. Consider t_i as the time required to transmit one complete IP packet of size d_i bits to a user i. The value of t_i is computed based on the user's current PHY data rate, IP packet size, and MAC/PHY protocol overhead bits. In such a scenario, the throughput of user i denoted by $\omega^{\mathrm{rr}}_{\mathrm{user}i}$ can be estimated as follows.

$$\omega^{\mathrm{rr}}_{\mathrm{user}i} = \frac{d_i}{\sum_{i=1}^{N} t_i}. \tag{5.1}$$

Similarly the access point throughput $\omega^{\mathrm{rr}}_{\mathrm{AP}}$ is given by

$$\omega^{\mathrm{rr}}_{\mathrm{AP}} = \frac{\sum_{i=1}^{N} d_i}{\sum_{i=1}^{N} t_i}, \tag{5.2}$$

The assumption that the access point serves users in a round-robin manner can be realized by controlling downlink user traffic sent to a WLAN AP by home agent. It has been discussed earlier that home agent acting as a mobility anchor, receives all downlink user traffic from the application server and then tunnels it to the users over their network paths. In other words, the EE entity at the home agent is capable of distributing user traffic over their available network paths through a kind of traffic shaping. If DE_n decides to send an equal amount of downlink traffic to all active users of a WLAN AP, the MAC queue will hold an equal amount of data from all users. Owing to the fact that the WLAN AP transmits data in a FCFS manner, in the long run users will receive an equal amount of data. Hence, this can be seen as if WLAN AP is scheduling users in round-robin manner.

The round robin way of scheduling WLAN resources, however, does not make an optimum use of the resources. This point can be elaborated with following example. Consider a single active user attached to a WLAN access point who is receiving a UDP flow comprises a fixed IP packet size of d bit. Assuming 54Mbps PHY data rate, the user experiences a throughput of $\frac{d}{t_{54\mathrm{Mbps}}}$ where $t_{54\mathrm{Mbps}}$ is the time to transmit one packet. As soon as another user with 6Mbps PHY data rate (who is also receiving a similar UDP flow) associates to the same access point, the overall access point throughput now amounts to $\frac{2d}{t_{54\mathrm{Mbps}}+t_{6\mathrm{Mbps}}}$. Considering a basic channel access mechanism of 802.11a $t_{6\mathrm{Mbps}} \simeq 5.6 \cdot t_{54\mathrm{Mbps}}$ which implies that joining of the second user reduces the overall access point throughput by 70%. This is because round-robin is a fair scheme which gives equal chance of medium access to all active users irrespective of their channel conditions.

One way to help in this situation is by performing resource management in such a manner that it provides users with medium access time in proportion to their

PHY data rate values. In other words, the users are given equal shares of the time slices so that the users with the higher PHY data rate can transmit more packets as compared to the users with the lower PHY data rate. This scheduling effect can be achieved in the above example if 6 packets from the first and 1 packet from the second user traffic flow are sent to WLAN access point. This is because the first user can receive 5.6 packets in a time period required by the second user to receive one packet. It implies that the DE_n should assign a traffic shaping rate for the first user which is 5.6 times higher than that of the second user. This will enhance the overall system throughput by 196% compared to simple round-robin scheme. This has been illustrated graphically in Figure 5.4 where the round-robin scheme is compared with the currently proposed 'channel aware' scheme.

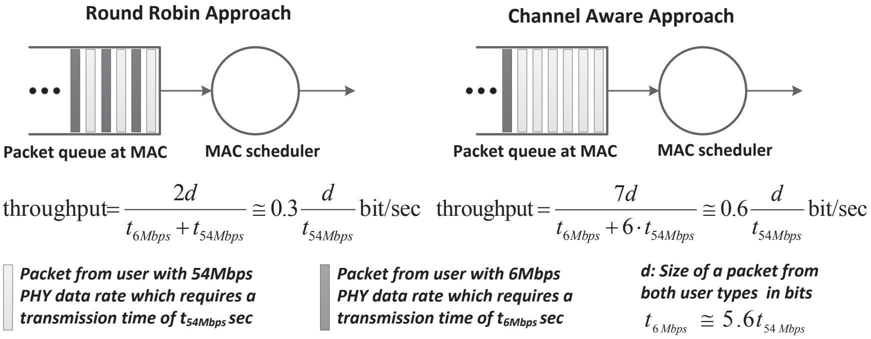

Round Robin Approach
Channel Aware Approach
Packet queue at MAC
MAC scheduler
Packet queue at MAC
MAC scheduler
$\text{throughput} = \frac{2d}{t_{6Mbps} + t_{54Mbps}} \cong 0.3 \frac{d}{t_{54Mbps}}$ bit/sec
$\text{throughput} = \frac{7d}{t_{6Mbps} + 6 \cdot t_{54Mbps}} \cong 0.6 \frac{d}{t_{54Mbps}}$ bit/sec
Packet from user with 54Mbps PHY data rate which requires a transmission time of t_{54Mbps} sec
Packet from user with 6Mbps PHY data rate which requires a transmission time of t_{6Mbps} sec
d: Size of a packet from both user types in bits
$t_{6\,Mbps} \cong 5.6 t_{54\,Mbps}$

Figure 5.4: Quasi-packet-scheduling of downlink user traffic at WLAN access point. Throughput is computed for WLAN access point in downlink.

It should be clear that the overall system performance gain in the channel aware scheme comes at the cost of reduction in throughput of the second user. However, the scheme is fair enough to give users their system throughput share in proportion to their PHY data rate while considerably improving the overall system throughput. The achievable system throughput gain strictly depends on the PHY data rates of the active users. For example, in a scenario where the users have the same PHY data rate, the channel aware scheme cannot bring any additional gain over the round-robin scheme. Nevertheless, the performance of the channel aware scheme will always be equal to or greater than that of the round-robin scheme.

In order to compute the throughput of a system following the channel aware scheme, assume r_i as the achievable data rate for a user who is the only active user associated to the access point. The r_i actually reflects the channel conditions or PHY data rate of the user i. This is because $r_i = \frac{d_i}{t_i}$, where t_i is the actual time taken by user i to transmit a packet of size d_i bits with a certain PHY data rate. Now consider that more users join this access point so that the total number of active users becomes N and all users are receiving a similar traffic flow comprising the same packet size of d bit. In this case, a user i's the share from the overall throughput should be in proportion to his achievable data rate r_i. The fraction of the share which has a range $(0,1]$, is denoted by e_i such that

$$e_i = \frac{r_i}{\sum_{i=1}^{N} r_i}. \tag{5.3}$$

In this way, the overall system throughput $\omega_{\mathrm{AP}}^{\mathrm{ch}}$ will be computed as follows

$$\omega_{\mathrm{AP}}^{\mathrm{ch}} = \frac{\sum_{i=1}^{N} e_i \cdot d_i}{\sum_{i=1}^{N} e_i \cdot t_i} \tag{5.4}$$

and the throughput of user i is given by

$$\omega_{\mathrm{user}_i}^{\mathrm{ch}} = \frac{e_i \cdot d_i}{\sum_{i=1}^{N} e_i \cdot t_i} \tag{5.5}$$

The above described schemes of scheduling WLAN bandwidth resources are just two examples. In general other scheduling schemes can also be developed and imposed as a policy at the DE_n entity.

5.4.2 Capacity Estimation of LTE Access Link

The LTE MAC scheduler relies on very complex algorithms and mechanisms to make efficient use of the available bandwidth resources while fulfilling the QoS demands of different services. Therefore, the individual user's throughput and overall cell throughput in LTE continuously varies due to several time variable factors like, channel conditions of all users, QoS requirements of the user traffic, behavior of congestion control and resource allocation algorithms, user traffic pattern, cell load level etc.

This work introduces a simple but effective way to estimate available user downlink capacity for the LTE access link. It has been discussed in Chapter 2 and 4 that downlink user traffic is mainly buffered in the PDCP buffers of the eNode-B before transmission over the air interface. Each bearer of a user has a dedicated queue at the PDCP layer to buffer the incoming traffic data. When scheduled by MAC scheduler for transmission, this data flows down to the RLC layer. After doing the required processing like segmentation, encapsulation etc., the RLC and MAC layers forward the data to the physical layer for transmission over the air interface. Considering this process, if the IE entity hosted on the eNode-B reports the mean throughput of a bearer's data flowing from PDCP to RLC layer along with its mean PDCP buffer occupancy, an accurate estimation of time varying link capacity can be carried out. This process is explained in the following.

A target is set for the maximum packet queuing delay $\breve{\gamma}$ for the PDCP buffer of a bearer. This value is multiplied by the bearer throughput h reported by the IE entity to obtain a target occupancy μ for the PDCP buffer of that bearer. To start this process, the DE_n initially decides to send a small amount of traffic load ε for the user bearer and monitors the reported throughput h as well as the PDCP buffer occupancy value ψ. This helps DE_n adjust the target PDCP buffer occupancy μ so that

$$\mu = h \cdot \breve{\gamma} \tag{5.6}$$

At the next time instant, if the IE entity reports the PDCP buffer occupancy as ψ, the additional data required to achieve target buffer occupancy μ is given as

$$\widehat{\mu} = \mu - \psi = h \cdot \breve{\gamma} - \psi \tag{5.7}$$

Assuming that the IE entity periodically sends reports of ψ and h every t seconds, the deficiency in buffer occupancy should be equalized within t seconds. This will require a stepwise increase in the existing traffic load ε by an amount of $\frac{\widehat{\mu}}{t}$ so that the adjusted value of traffic load $\widehat{\varepsilon}$ to be sent over the LTE path will be

$$\widehat{\varepsilon} = \varepsilon + \frac{\widehat{\mu}}{t} \tag{5.8}$$

This cycle continues and the LTE link capacity of a user's bearer is adaptively adjusted every t seconds using equations 5.6 to 5.8. The basic principle of this mechanism lies in the fact that the PDCP buffer occupancy of a bearer reflects congestion level at the air interface. The DE_n entity always tries to maintain a buffer occupancy of μ for a bearer and any change in that occupancy indicates the tendency of either an increase or a decrease in the bearer throughput. For example, when bearer throughput reduces due to some reason (e.g., cell overload or bad channel conditions) the egress data rate from PDCP buffer becomes lower than the ingress data rate which makes the PDCP buffer occupancy grow. The opposite is true when the bearer throughput increases. Such a change is reflected in $\widehat{\mu}$ through ψ which, in turn, adjusts the traffic load $\widehat{\varepsilon}$ on the link.

In case of realtime traffic, the value $\breve{\gamma}$ is determined by the de-jitter length $\gamma_{\text{de-jitter buffer}}$ which is the length of de-jitter buffer in units of seconds. When dealing with TCP based non-realtime traffic, the TCP re-order timer $\gamma_{\text{tcp reorder buffer}}$

value should be employed as explained in Section 3.4.5. As far as, the value of t is concerned, the simulation study shows that a value in the range of 10–50 ms serves the purpose.

5.4.3 Simulation Scenarios and Results

In order to evaluate the performance of heterogeneous networks which support simultaneous use of multiple interfaces of the user terminal, this section relies on simulation based studies of the system. Owing to the fact that 3GPP has standardized the integration of WLAN & LTE access networks without multi-homing support and that the 3GPP standard has been extended in this work to enable user terminals to exploit multi-homing and flow management features, there are two main scenarios to be compared against each other. The user terminals in 3GPP defined heterogeneous network can perform seamless vertical handovers (HO) between the WLAN and LTE access networks but they cannot use the two network paths simultaneously. This will be referred to as "3GPP HO" scenario. The default resource utilization policy in "3GPP HO" scenario is that the user terminals communicate through the LTE network when they are away from WLAN AP and execute a handover to the WLAN network as soon as they are found to be in its coverage. Whereas, the second scenario will be referred to as "Multi-P" where user terminals can exploit the flow management features. In this scenario, the user terminals which are in the overlapped coverage of the LTE and WLAN access networks may simultaneously make use of two network paths. The distribution of user traffic over the two network paths is mainly managed by the DE_n entity of the flow management architecture. During the time when the user terminals are not in the access coverage of WLAN, they have only the LTE access link to communicate.

As the 3GPP standard does not allow the simultaneous use of multiple accesses, this shortcoming poses a serious implication during the vertical handover. This is because when a user terminal executes a handover from one access network to another, the user data buffered in different elements (e.g., router, base station etc.) of the previous network could be lost. For example, LTE keeps the received IP packets mainly at the PDCP layer while WLAN keeps the data buffering at the MAC layer queue before transmission over the air interface. When a user executes a vertical handover, Mobile IPv6 registers the new care-of address with the home agent and, therefore, the data arriving at the old care-of address is discarded. IP packets lost in this way have to be recovered by the upper layers through retransmissions. This behavior leads to application performance degradation both for TCP and UDP based services.

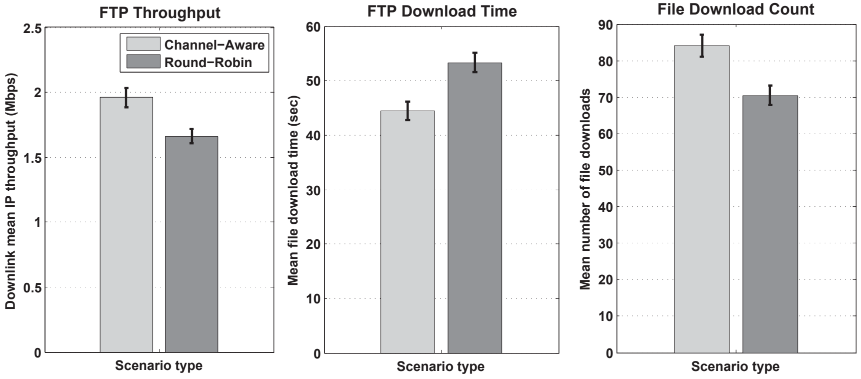
In the "Multi-P" scenario, through the multi-homing support, the user terminals are enabled to use the WLAN access when being in its coverage and also keep the LTE connection alive at the same time. This avoids any potential data loss in the LTE network. However, there could be a problem when a mobile user's connectivity to the WLAN access network is lost all of a sudden by moving out of the coverage. This could leave some packets buffered in the WLAN access network which will be eventually lost. In order to minimize such losses, the DE_n entity considers a user terminal's WLAN connectivity active only when its PHY data rate is 9 Mbps or higher. This is because when a mobile user terminal's PHY data rate happens to be 6 Mbps, it implies that the user has walked to the edge of the WLAN access coverage which could be an indication that loss of WLAN access link is imminent. As the DE_n entity does not send new traffic data over the WLAN network for such users, it gives them a chance to receive the buffered data at the WLAN AP before the actual loss of the link happens. Moreover, in contrast to the "3GPP HO" scenario, the "Multi-P" approach can control the buffered data at the base station with the help of $\gamma_{\text{de-jitter buffer}}$ and $\gamma_{\text{tcp reorder buffer}}$ parameters. Hence, keeping the buffered data to a minimal level helps minimize data loss.

In order to conduct this simulation base study, two simulation setups are investigated. The first simulation setup is intended to show the advantages of using the "Channel Aware" approach over the "Round-Robin" approach of the WLAN resource management. This setup is composed of five FTP users who are downloading files one after the other using only the WLAN access link. In the absence of any resource management function a TCP connection would buffer data at the access point equal to its windows size. This implies that all users will have the same number of IP packets buffered at the MAC queue of the access point which resembles to a situation created by the "Round-Robin" approach. In this way, it can be claimed that "Round-Robin" approach also represents the "3GPP HO" scenario where user traffic is not shaped according to any resource allocation function like flow management.

The second setup represents a mixed user traffic case where 23 users are accessing a number of services commonly found in daily life, e.g., VoIP, FTP, HTTP, News video streaming as well as Skype video calls. It can be noticed that this section focuses only on downlink communication and therefore uplink applications are not considered here. The users move within one LTE eNode-B cell which has access coverage overlap with two WLAN access points as shown in Figure 5.6. Other simulation configurations can be seen found in Table 5.2.

Simulation Setup 1: FTP User Traffic

Figure 5.5 shows the FTP downlink performance as experienced by users in the



(a) (b) (c)

Figure 5.5: FTP downlink performance comparison between "Round-Robin" and "Channel Aware" resource management approaches in WLAN.

first simulation setup. There are five FTP downlink users who are moving within the coverage of a WLAN access point. The users perform FTP file download through the WLAN access without using LTE connectivity. Figure 5.5(a) shows the mean value of per user downlink throughput at the IP layer, Figure 5.5(b) shows the mean file download time, and Figure 5.5(c) compares the mean number of successful FTP file downloads by a user in one simulation run. The error bars on bar plot represent 95% confidence interval values.

It can be seen that by using the "Channel Aware" approach for resource management, a throughput gain of ≈20% can be achieved. This gain is not as much as shown in Figure 5.4 where one of the ideal situations was presented. As explained earlier, the gain is realized by exploiting the good channel conditions of users in the presence of users with bad channel conditions. However, if the users have similar channel conditions then the "Channel Aware" scheme cannot bring much additional gain. As the channel conditions of mobile users are changing continuously in the simulation, the users appear to have similar channel conditions. As a result, the "Channel Aware" approach was able to improve the access point throughput by ≈20% which is still a substantial pay-off.

Simulation Setup 2: Mixed User Traffic

In this setup, the system is populated with users who generate a rich traffic mixture as shown in Table 5.2. Figure 5.7 shows a comparison of FTP downlink ap-

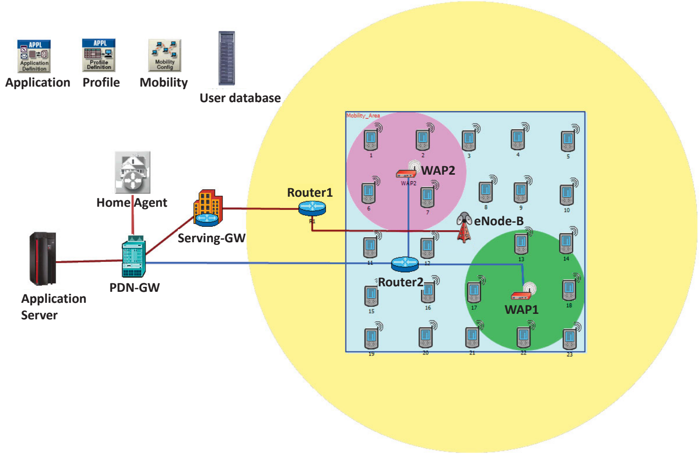


Figure 5.6: Simulation scenario setup in the OPNET simulator. The large circular area shows the coverage of LTE and two smaller circular areas represent the WLAN network coverage. The user movement is restricted to the rectangular area.

plication performance for two scenarios, i.e., without multihoming (“3GPP-HO”) and with multihoming (“Channel Aware”). Mean downlink throughput measured at the IP layer has been depicted for each user in Figure 5.7(a). The error bars represent the 98% confidence interval value. It is evident that users in the “Channel Aware” scenario manage to achieve ≈6% higher throughput compared to that of users in “3GPP-HO” scenario. This is much less than the value seen in previously discussed setup of FTP users in WLAN network. This can be attributed to the presence of realtime traffic which is a hurdle in achieving higher performance for “Channel Aware” scheme. The reason is that “Channel Aware” relies on exploiting good channel conditions of the users. A user with good channel conditions is offered an opportunity to receive a large amount of data to increase spectral efficiency. If this user is receiving TCP based flow like FTP, then the TCP data rate can adapt itself to the available link capacity. However, if the user is receiving realtime traffic flow with fixed data rate which is much less than the offered link capacity, the provided resources remain under-utilized. This causes the “Channel Aware” scheme to exhibit lower efficiency.

Figure 5.7(b) shows the mean time to download a file of 10MByte size as experienced by the users in two comparison scenarios. Similarly, the mean number of files downloaded by each user during the whole simulation run time are shown

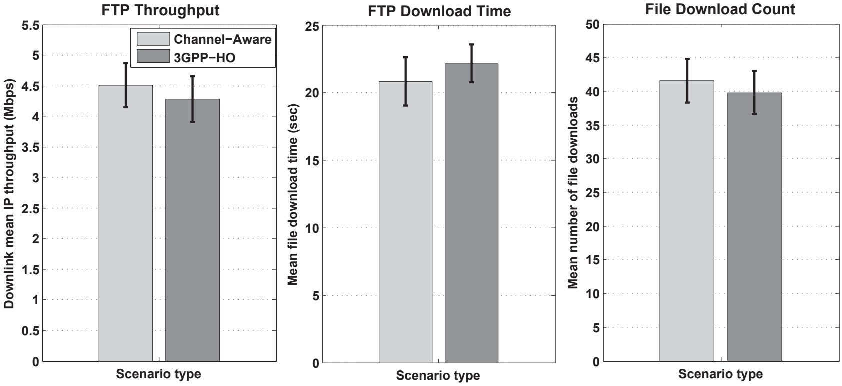
Table 5.2: Simulation configurations for evaluation of the downlink flow management scheme.

Parameter	Configurations
Total number of PRBs	25 PRBs (5 MHz specturm)
Mobility model	Random Direction (RD) with 6 km/h
Number of users	5 VoIP, 3 live News video streams, 7 Skype video calls, 3 HTTP and 5 FTP downlink users
LTE channel model	Macroscopic pathloss model [25.06], Correlated Slow Fading.
LTE MAC scheduler	Time domain: Optimized Service Aware, Frequency domain: Iterative RR approach [S. 12]
WLAN access technology	802.11a, RTS/CTS enabled, coverage ≈ 100 m, operation in non-overlapping channels
Transport network	1Gbps Ethernet links, no link congestion
VoIP traffic model	G.722.2 wideband codec, 23.05kbps data rate and 50frame/s
Skype video model	MPEG-4 codec, 512kbps, 30frame/s, 640x480 resolution, play-out delay: 250ms
Live News video model	MPEG-4 codec, 1Mbps, 30frame/s, 720x480 resolution, play-out delay: 250ms
HTTP traffic model	100 bytes html page with 5 objects each of 100Kbytes, page reading time: 12s
FTP traffic model	FTP File size: constant 10MByte, as soon as one file download finishes, the next FTP file starts immediately.
TCP configurations	TCP new Reno, Receiver buffer: 1Mbyte, Window scaling: enabled, Maximum segment size: 1300Byte, TCP reorder timer: 250ms
DE_n decision interval	Every 20ms
Simulation run time	1000 seconds, 13 random seeds, 95% Confidence interval

in Figure 5.7(c). These two figures also show the performance gain of the same magnitude as observed in Figure 5.7(a).

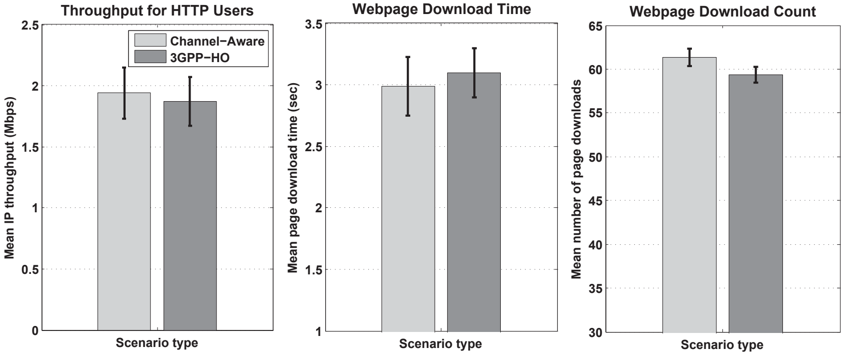
A performance comparison of HTTP application can be seen in Figure 5.8. HTTP users are seen to achieve less throughputs compared to downlink FTP users. The rationale behind this is the smaller sized (100KByte) HTTP objects compared to large FTP files (10MByte). Due to its small size, an embedded object download finishes during the “slow start” phase of TCP. This prevents the users from obtaining higher steady throughput achievable only in the post “slow start” phase of TCP. This is also a hurdle for the “Channel Aware” approach to show its full performance.

Figure 5.9 shows the box plots (also known as box-and-whisker plots) to represent the user perceived MOS scores of VoIP and video applications. A box-and-whisker plot graphically depicts the groups of numerical data through their five



(a) (b) (c)

Figure 5.7: FTP downlink performance comparison between "3GPP-HO" and "Channel Aware" approaches.



(a) (b) (c)

Figure 5.8: HTTP downlink performance comparison between "3GPP-HO" and "Channel Aware" approaches.

number summary, i.e., (1) minimum, (2) maximum, (3) median (or second quartile), (4) the first quartile, and (5) the third quartile. The bottom and top of the box are the first and third quartiles, respectively. The band near the middle of the box is the median. The whiskers represent the maximum and minimum of all the data values. Moreover, any data not included between the whiskers is plotted as an outlier with a cross '+' sign. Further explanation about the box plot has been given in Appendix B.

The MOS values of VoIP call are computed during the simulation run every 1 sec using E-model as discussed in Section A.3.1.2 of Appendix Chapter. The Evalvid tool discussed in Section A.3.2.2 is used to estimate user QoE as MOS values for each user call. These computations are performed offline after the completion of a simulation run. It can be seen that "Channel Aware" approach succeeds in delivering the best user QoE to VoIP, Skype video, and News video users. For this purpose, the "Channel Aware" approach utilizes its capabilities to estimate user access link capacities and to manage the resources in a way such that the QoS demands of realtime applications are always fulfilled. As long as the user stays outside the WLAN access network coverage, the realtime traffic is served over the LTE access link which offers QoS aware service by giving higher priority to VoIP/video traffic over FTP/HTTP traffic. Therefore, during this time both "Channel Aware" as well as "3GPP-HO" can attain good MOS values. However, when the user is in the coverage of the WLAN access point, the users in "3GPP-HO" execute a complete handover to WLAN access network which fails to offer QoS differentiation to the delay sensitive realtime applications. This leads to poor user QoE during these times.

In contrast to "3GPP-HO", the proposed "Channel Aware" approach makes an accurate estimation of available WLAN access link capacity and utilizes it accordingly. Moreover, any deficit in the required bandwidth demands of the application is fulfilled from the LTE access link. In this way, the users are always served with the necessary bandwidth resources required to achieve user satisfaction irrespective of the network congestion.

The VoIP MOS value is affected by both end-to-end delay and packet loss rate while the model used to compute video MOS is indifferent to the packet delays as long as they are less than the de-jitter buffer length (250ms). Any video packet delayed beyond the de-jitter buffer length is assumed to be lost. These lost packets drag the third quartile of video MOS values beyond 2. On the other hand, the VoIP QoE evaluation model considers a continuous effect of packet delays which leads the third quartile to stay above a MOS value of 3. Moreover, both applications suffer from the packet losses during the vertical handover which is reflected by the minimum values and outliers in the MOS plot.

Another closer look at realtime application performance is provided by Figure 5.10 which represents the end-to-end packet delay of video applications. It is obvious from the figure that the "Channel Aware" approach manages to keep delays under control by making use of intelligent flow management. However,

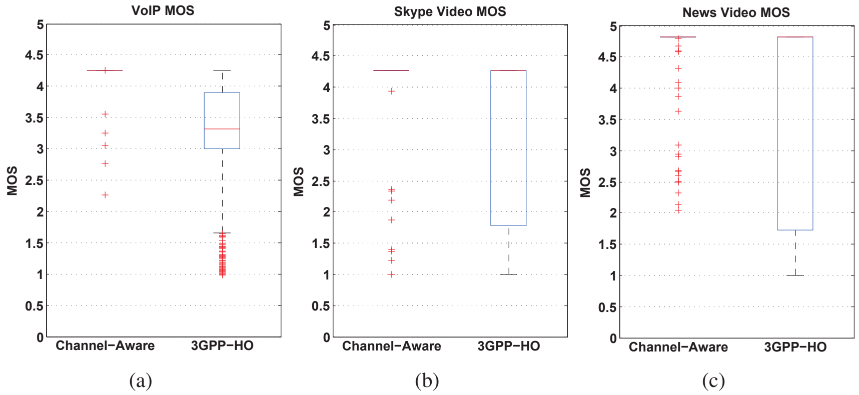


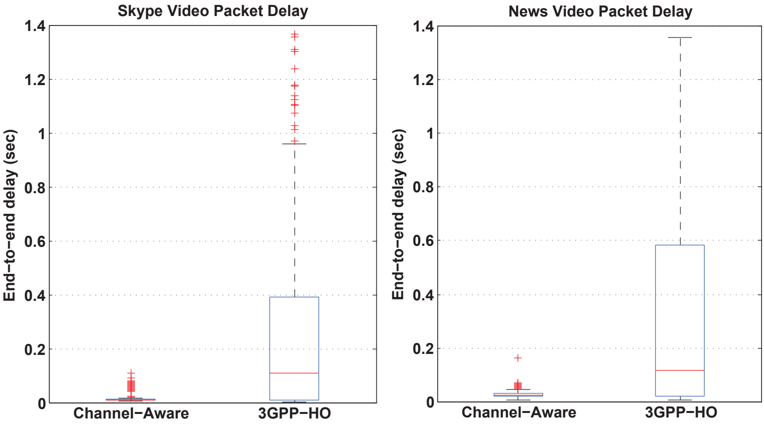
Figure 5.9: Downlink performance comparison of realtime applications between "3GPP-HO" and "Channel Aware" approaches.

the users in the "3GPP-HO" scenario suffer from excessive delays for extended periods of time. These are the times when the users are being served by WLAN access networks. The reason for large packet delays is the presence of FTP users in the WLAN access networks. Owing to the fact that the radio interface is the only bottleneck for users in the "3GPP-HO" scenario, the TCP connections buffer a significant amount of data (which is equal to the TCP window size of 1MByte) at the single MAC queue of the WLAN access point. Due to this high buffer occupancy and 'First Come, First Serve' strategy of the WLAN MAC, the realtime users suffer from large queuing delays.

In contrast to this, the "Channel Aware" approach manages the WLAN MAC buffer occupancy so that it does not grows excessively high. This is because here the downlink data is kept at the home agent in individual user buffers. The DE_n entity sanctions only that amount of user data to flow to the WLAN access point which can be handled by that particular user's access link. This alleviates the head-of-line blocking situation observed in the "3GPP-HO" scenario.

5.5 Uplink Flow Management

In the previous section, it has been shown that multihoming and flow management are capable of delivering substantial improvement in user QoE. However, that section focused only on the downlink communication. This section extends



(a) (b)

Figure 5.10: Video packet delay comparison for downlink communication.

the investigations to the uplink communication. The mechanisms devised to estimate downlink access link capacity will also be used here with some essential modifications.

5.5.1 Estimation of WLAN Link Capacity

In the WLAN network where a number of users are contending for medium access to perform uplink transmission, the network capacity and the individual user throughput is highly variable. In this way, another variable factor which affects the throughput, in addition to those mentioned in Section 5.4.1, comes from the collisions of contending stations. This makes it rather complex to mathematically compute individual user throughput in such a network. In this work two approaches are proposed which can help estimate the user throughput over time without analytical modeling of the network.

5.5.1.1 Approach 1 - Random Access:

This approach is based on the same principle as explained in Section 5.4.2 for the LTE access link capacity estimations. It requires a metering function, as part of the information management entity (IE), which is introduced in the WLAN MAC layer of the user terminal. The metering function is intended to measure the outgoing data rate from the buffer, as well as, the buffer occupancy level. These two values are obtained periodically and sent to the DE_u entity residing at the UE. In the beginning, the DE_u entity directs a sufficient amount of user traffic to the WLAN MAC for transmission. This data stays at the MAC layer buffer before transmission over the radio interface. The DE_u entity now continuously receives the buffer occupancy level reports and adjusts the size of traffic flow to the WLAN network path to keep the buffer occupancy at the target level. In this way, if the buffer occupancy level increases, it hints at reduction in available WLAN path capacity due to some reason, e.g., congestion, poor channel conditions, more collisions etc. In this case, the DE_u entity accordingly reduces the traffic flow amount directed towards the WLAN path. The opposite is true if a reduction in MAC buffer occupancy is observed which suggests an improvement in the path capacity. The DE_u entity then takes advantage of this by sending more traffic towards the WLAN path. Following this approach time varying WLAN network path capacity can be tracked and used by the DE_u entity.

In this approach, the DE_n entity informs the DE_u entity to act autonomously in estimation and utilization of WLAN access link resources. As a part of this decision, the DE_n entity may also inform the DE_u entity to take care of a maximum or minimum traffic flow size to be directed to the WLAN access link.

The practical implementation of the suggested approach is straightforward as it does not require any modification in the UE hardware or WLAN MAC protocol. As far as the dynamic size of the target MAC buffer occupancy is concerned, it is computed as suggested in equation 5.6.

5.5.1.2 Approach 2 - Time Round-Robin:

This approach involves a quasi-scheduling of the WLAN network resources. The main idea behind this approach is to save the network resources which are otherwise wasted due to contention in channel medium access and packet collisions when multiple users transmit simultaneously. By saving these network resources the user throughput can be improved and the network capacity can be increased. This approach is realized by the active cooperation of both DE_n and DE_u entities. In this approach the DE_n entity receives the information about the active users associated with a WLAN AP. This information is periodically sent by the IE entity residing at the WLAN AP. The DE_n entity builds a list of active users and assigns time slots to them to be transmitted in a round-robin manner. In order to minimize the signalling traffic and associated delays, the scheduling decision is communicated to the DE_u entities in the form of ON-OFF periods. A user terminal transmits only during the indicated ON period of self-repeating scheduling intervals and halts the transmissions otherwise. For example, if two users are associated to an access point, they will share the network resources in a round-robin way so that one user transmits only for a fixed time period allowed by its allocated time slot. During the transmission time of the first user, the second user halts transmissions and waits for the first user's time slot to end. When the time slot of the first user elapses, the channel access time for the second user starts and lasts for duration equal to its time slot. During this time, the first user has to stop its transmissions over the WLAN. This cycle is followed until any change happens in the network topology, e.g., a new user joins or an existing user leaves the WLAN network etc. In response to such events, the DE_n entity reschedules the users and conveys the updated schedule to the DE_u entities of the active users.

Assume that each user i is assigned with a time slot of length θ_i during which it is allowed to transmit exclusively. Referring to Section 5.4.1 for the definition of r_i as the achievable data rate for a user i, the mean user throughput $\omega^{\text{trr}}_{\text{user}_i}$ can be computed using the following equation, i.e.,

$$\omega^{\text{trr}}_{\text{user}_i} = r_i \cdot \frac{\theta_i}{\sum_{i=1}^{N} \theta_i}. \tag{5.9}$$

where N is the total number of active users associated to the WLAN access point.

This approach assumes that the participating users are precisely time synchronized which is realistic to be achieved in the real world. Section 5.3.4 mentioned a few techniques to do synchronization with an accuracy in the range of 10ns to 1ms. Without precise time synchronization, there could be some contention for medium access during the overlapped period of the two adjacent time slots. However, if the length of the time slot is much larger than the overlapping period, the users will still be able to enjoy sufficient contention free time periods for their transmission. Moreover, during the overlapping period only two users will be competing for the medium access which is still better than the situation where all active users are contending for the medium access and hence degrades the network performance.

Another problem associated with this approach is related to the WLAN MAC buffer contents at the user terminal. Owing to the fact that the WLAN MAC function of the user terminal is unaware of the proposed approach, it will try to transmit the existing buffered data even when the user time slot has elapsed. A possible solution could be that the WLAN radio transmitter is switched off after transmission and switched on again when the next time slot for transmission approaches. This can prevent the users to contend for the channel during the time slot of the intended user. However, in this case the packets already waiting in the MAC buffer could be lost. To avoid this packet loss and minimize the contention period, the DE_u entity sends only two IP packets to the WLAN MAC for transmission. When the

WLAN MAC function de-queue a packet for transmission, a software interrupt is sent to the DE_u entity which responds by delivering another IP packet to the MAC buffer. This way, at any time instant during the transmission time slot of a user there are at most two packets lying in the buffer. As soon as the time slot of the user elapses, the DE_u entity stops sending new packets to the WLAN MAC and therefore this user contends for the medium access with the user of the adjacent time slot to transmit only two packets. Another possible way to circumvent this problem could be to use guard times between two adjacent time slots.

The quasi-scheduling decisions are conveyed to the users using LTE signalling in order to avoid any excessive delays. Due to the fact that each user follows its time slot for transmission, there are very few events of packet collision or medium access contention involved. As a consequence, an improvement in the network capacity is expected. This mechanism resembles the famous token ring protocol, however, it does not require any modification on the WLAN MAC protocol. Though in this work simple round-robin scheme is used, the proposed quasi-scheduling approach is general enough to accommodate other resource sharing schemes.

It should be noted that the first approach for the WLAN link capacity estimation can be used by both network-centric and user-centric flow management schemes. However, the second approach must be used with the help of network functions and is therefore only suitable for the network-centric flow management scheme.

5.5.2 Estimation of LTE access link capacity

The estimation of the LTE access link capacity available to a user terminal is performed by using the same solution as discussed in Section 5.4.2. For this purpose, the IE entity hosted on user the terminal periodically provides the PDCP buffer occupancy and the LTE uplink throughput information to let the DE_u entity manage the uplink bandwidth resources. The relevant parameters for resource management (e.g., whether to use the LTE path, upper and lower thresholds of data traffic to be sent on the LTE path etc.) are provided by the DE_n entity as a part of the decision and policy. Based on the decision parameters and link capacity estimation algorithm, DE_u autonomously judges the amount of available bandwidth on the LTE uplink and accordingly sends the user data traffic over this path. This approach of link capacity estimation can be used by the user-centric as well as the network-centric flow management.

5.5.3 Simulation Scenarios and Results

In this section simulation results are used to evaluate the performance of the proposed approaches in multihoming scenarios. Similar to the study of downlink approaches, mainly the performance of the two scenarios is compared. “3GPP-HO” once again refers to the scenario which does not support multihoming or flow management. The other scenario is called “Multi-P” where the users are enabled to exploit multihoming and flow management features. Both scenarios follow similar policies of access network association as observed in the study of downlink approaches. That is, the users in the “3GPP-HO” scenario are preferably served over the WLAN access link as long as they are in WLAN coverage. On leaving that coverage, they are seamlessly handed over to the LTE network. In “Multi-P” scenario, the users can make use of WLAN and LTE access links simultaneously. More specifically, realtime users prefer to utilize the WLAN access link capacity as much as possible and may get additional required capacity from the LTE access link in order fulfill a fixed application data rate demand. On the other hand, the FTP users try to utilize the capacities available from both access links simultaneously.

As the ‘proactive mobility management’ or make-before-break strategy is not followed during vertical handovers in the “3GPP-HO” scenario, the users have to suffer from packet losses in both uplink and downlink directions. “Multi-P” avoids such losses by supporting multihoming when entering the WLAN access network coverage. Moreover, in the “Multi-P” scenario the DE_u entity does not send new data packets for transmission to the WLAN interface when the PHY data rate is 6Mbps. This is due to the fact that when a mobile user terminal achieves the lowest PHY data rate of 6Mbps, it is an indication that the user has moved near to the coverage boundary and loss of WLAN access link is imminent. At this point, the user is given an opportunity to transmit already buffered data at the MAC layer and hence minimize the packet loss rate. This is the same strategy which is followed by the DE_n entity in the downlink direction as discussed in Section 5.4.3.

The achieved user QoE is compared in “Multi-P” and “3GPP-HO” scenarios for popular uplink applications, i.e., Skype video calls, FTP uplink, and VoIP. For this purpose, a simulation setup is created which is composed of 20 users moving in an area where the coverage of one LTE cell and two WLAN access point is available (similar to Figure 5.6). The other simulation configuration details are the same as listed in Table 5.2.

Within the “Multi-P” scenario two approaches for the WLAN access link capacity estimation are employed as discussed in Section 5.5.1. In this way, there are three scenarios to be discussed in this simulation setup, i.e., “3GPP-HO”, “Random Access”, and “Time Round-Robin”.

In addition to the aforementioned simulation setup of mixed user traffic, another simulation setup is created with FTP users moving within the coverage of one WLAN access point without having LTE connectivity. The motivation behind this simulation setup is to study the performance comparison of the two “Multi-P” approaches, i.e., “Random Access”, and “Time Round-Robin”.

Simulation Setup 1: Mixed User Traffic

In this simulation an uplink traffic mixture is generated by 5 VoIP calls, 4 Skype video call, and 11 FTP uplink users. Figure 5.11 shows the uplink FTP application performance. For example, the mean IP uplink throughput as experienced by each FTP user is shown in Figure 5.11(a). It is evident that the highest mean uplink throughput has been achieved by the “Time Round-Robin” approach of “Multi-P” and the lowest performance is shown by the users in the “3GPP-HO” scenario. Quantifying the performance figures, “Time Round-Robin” approach manages to provide approximately 35% higher IP throughput compared to the “3GPP-HO” scenario. Moreover, an approximately 27% gain in uplink IP throughput is observed for the users following the “Random Access” approach compared to the “3GPP HO” scenario. The reason for “Time Round-Robin” to attain higher performance compared to the “Random Access” approach is due to its ability to minimize medium access contention in WLAN by allocating each associated user terminal a 10ms time slot for exclusive transmissions. All performance metrics shown in Figure. 5.11 indicate that “Multi-P” algorithms outperform the “3GPP HO” approach by making better use of aggregated bandwidth resources through network path capacity estimation.

As far as the realtime applications (i.e., VoIP and video) are concerned, their performance is evaluated by comparing the Mean Opinion Score (MOS) values as shown in Fig. 5.12. These results indicate that the “Multi-P” approaches provide an excellent performance for the VoIP calls, as well as, for the video conference application type. The reason for “3GPP-HO” users to achieve low MOS value can be attributed to packet losses during the vertical handover as well as to the lack of QoS support in 802.11a networks. Thanks to algorithms of “Multi-P” scenario which estimate and manage 802.11a resources in a way that not only the required QoS for realtime traffic is provided but also an enhanced throughput performance for non-realtime users is accomplished.

A comparison of Figures 5.9 and 5.12 reveals that the user in “3GPP-HO” scenarios are exposed to be inferior to the QoE in downlink compared to that of uplink. The reason for this behavior can be explained by comparing the end-to-end packet delay figures, i.e., Figure 5.10 and 5.13. It is noticed that end-to-end packet

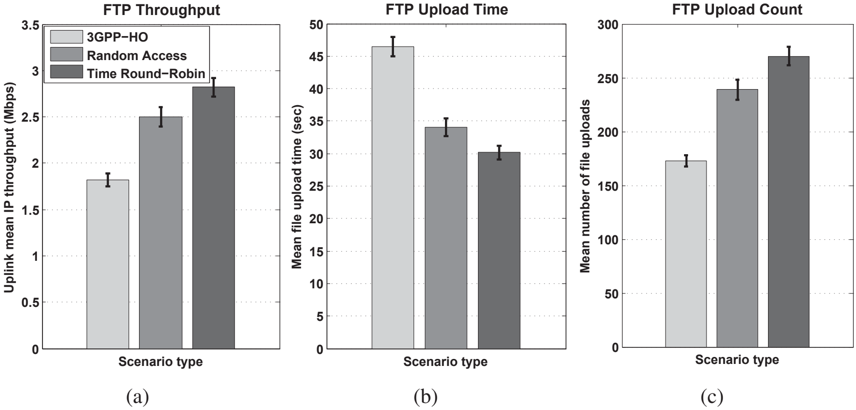


Figure 5.11: FTP uplink performance comparison among "Time Round-Robin", "Random Access", and "3GPP-HO" approaches.

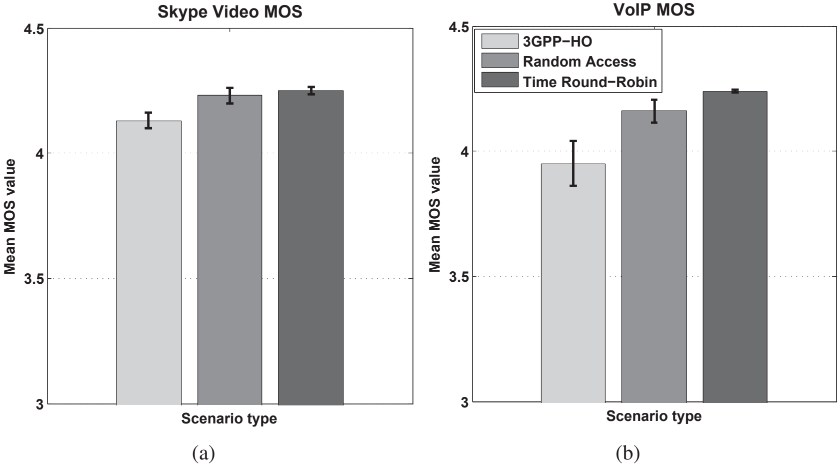
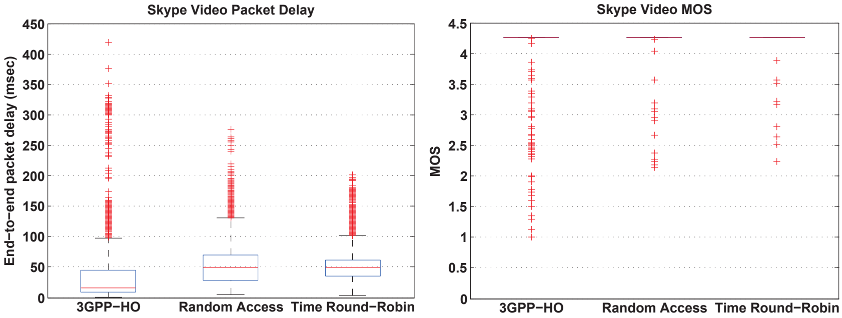


Figure 5.12: Video and VoIP uplink application performance comparison among "Random Access", "Time Round-Robin", and "3GPP-HO" WLAN resource management approaches.

delay for uplink transmission has been confined to a much lower value range compared to that of downlink transmission. This is because in downlink, the data packets from all user traffic are buffered in a single MAC queue at the WLAN access point which causes head-of-line blocking for realtime application traffic. As a result, the queuing delay for these packets grows excessively high leading to poor user QoE. In contrast to this, in uplink the users have individual MAC queues in their terminals and they have to compete only for the medium access to transmit. As long as the number of active users is relatively small the medium access delay remains restrained causing less adverse effects on user QoE.



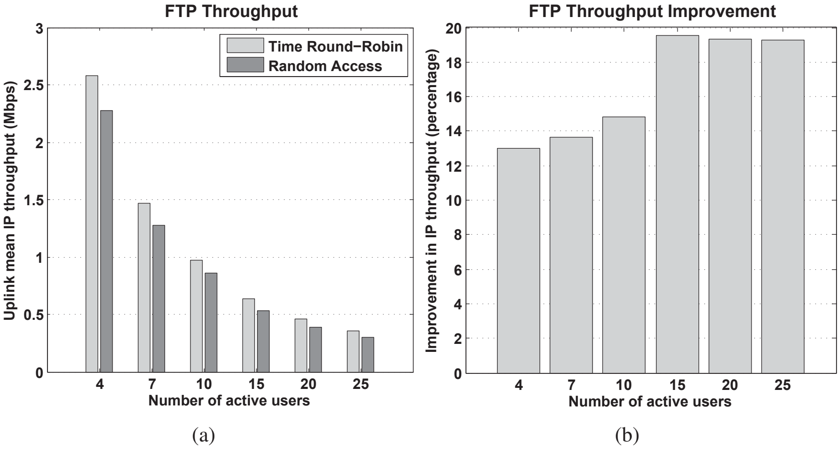
(a) (b)

Figure 5.13: Video packet delay comparison for uplink communication.

Simulation Setup 2: FTP User Traffic

In this simulation setup only FTP uplink users are considered. In order to compare the performance of "Time Round-Robin" and "Random Access" approaches at various load levels, the number of active users are varied from 4 to 25. Figure 5.14(a) shows the mean IP throughput achieved by each user as a performance comparison metric. It is illustrated that as the number of active users increases the throughput share of each user reduces accordingly. Furthermore, the "Time Round-Robin" approach is always seen to outperform the "Random Access" approach by providing a throughput gain above 12%. It is also evident from Figure 5.14(b) that the throughput gain increases with the number of active users in WLAN access network. This is due to the contention in medium access and associated transmission collisions which grow along with the user population. Owing to the fact that the "Time Round-Robin" approach circumvents such collisions by giving each user 10ms dedicated time slot for transmission, a noticeable improvement in the user throughput is observed.

A drawback of “Time Round-Robin” approach is the OFF period during which the users let a single user transmit exclusively without any contention. As the number of participant users increases, this OFF period also increases proportionately. In the absence of multihoming, when a user has only WLAN access link to transmit, these long periods tend to produce adverse effects on application performance because of the associated delays. Such an effect will be more unfavorable for realtime applications that are very sensitive to packet delays. Therefore, employing the “Time Round-Robin” approach is not recommended in the absence of multihoming support when a large number of users are to be served.



(a)

(b)

Figure 5.14: Per user mean IP uplink throughput comparison among the “Time Round-Robin” and “Random Access” approaches.

This chapter introduced overlay architecture for flow management which complements 3GPP compliant heterogeneous network architecture in achieving enhanced user QoE. A prerequisite of intelligent flow management was fulfilled by introducing novel approaches for access link capacity estimation. The effectiveness of flow management and the proposed mechanisms to estimate user access link capacity was validated using simulation results. The next chapter will discuss further cutting-edge approaches based on analytical modeling of access technologies. This will let flow management optimally utilize network bandwidth resources by overriding the default packet scheduling schemes of the access technologies.

6 Analytical Solution for Optimized Resource Allocation

This chapter discusses analytical solutions in order to optimize network resource allocation for multihomed users. Chapter 5 has extensively discussed how the resource allocations can be made efficient using multihoming in conjunction with the Flow Management. That study is further extended in exploring the upper limits of achievable network performance. For this purpose, the techniques of operations research or mathematical optimization are employed. A prerequisite in exploiting these techniques is the analytical modeling of network access links which describe how the network (spectrum) resources are translated to the achievable user data rates. Such functions along with other conditions of serving users derive the modeling of the optimized resource allocation process.

The organization of this chapter is as follows. Section 6.1 serves as an introduction to Linear Programming which is a technique of mathematical optimization. The development of analytical models for network access links is discussed in Section 6.2. After that, the resource allocation problem is formulated in Section 6.3 along with a simulation based study of the proposed approach. Finally, the computational complexity of the developed analytical solution is discussed in Section6.4 and the chapter is concluded by devising alternative approaches based on heuristic algorithms which offer matched performance and exhibit less computational complexity.

6.1 Linear Programming

The idea about linear programming can be traced back to Fourier's work in 1826. However, it was George B. Dantzig who introduced it as a discipline to solve a large class of optimization problems in 1947. The term 'programming' should not be confused here with the act of developing computer code. In fact, in the 1940s this term was synonymous with 'planning'. This way, linear programming is a subset of mathematical programming which is itself a field of operations research. Linear programming is renowned as one of the most commonly employed operations research methods. Other operations research techniques include simulation, queuing theory, regression analysis, stochastic processes, network analysis, game theory, etc.

The Linear Programming deals with the efficient allocation of limited resources with the objective of maximizing profit or minimizing cost. In formal words, a linear program is concerned with the problem of maximizing or minimizing a linear function, subject to linear constraints. Some models naturally exhibit linearity based on physical properties of the problem. Others may be linearized by employing mathematical transformation techniques.

The following points must be true of a problem to be solved using linear programming:

1. The variables whose values are to be decided in an optimal fashion must be non-negative. These unknown variables are called 'decision variables'. They can be written as

$$x_j \geq 0, \quad j = 1,2,3,\ldots,n \tag{6.1}$$

2. The criterion for choosing the optimal values of decision variables must be a linear function of the decision variables. This criterion is referred to as objective function and can be written as

$$O = c_1x_1 + c_2x_2 + \ldots + c_nx_n = \sum_{j=1}^{n} c_jx_j \tag{6.2}$$

   where $c_1,c_2,c_3,\ldots,c_n$ are constant values.

3. The optimal value of decision variables must abide by the operating rules, called 'constraints'. Each constraint must consist of either an equality or an inequality associated with some linear combination of the decision variables, e.g.,

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n \gtreqless b \tag{6.3}$$

   where $a_1,a_2,a_3,\ldots,a_n$ are constant values.

The requirements for the objective function and constraints as linear functions of the decision variables justify the term 'linear' in linear programming. A combination of certain values for the decision variables is called a 'solution'. A solution is considered as a 'feasible solution' if it satisfies all constraints. Furthermore, any feasible solution is called 'optimal solution' if it also achieves the desired maximum / minimum.

In some cases, no feasible solution exists for a problem. This could be mainly because of contradictions among the constraints. Such a problem is called ‘infeasible problem’. Another form of problems is referred to as ‘unbounded problems’. A problem is unbounded if it has a feasible solution with arbitrarily large objective values, e.g. consider the following simple problem,

$$\begin{aligned} \text{maximize} \quad & x_1 - x_2 \\ \text{subject to} \quad & x_1, x_2 \geq 0. \end{aligned}$$

In this case, setting the value of x_2 equal to zero, x_1 can be assigned an arbitrarily large value in maximizing the objective function. This makes it an unbounded problem. In the process of problem formulation and finding optimal values, it is important to detect when a problem turns out to be infeasible or unbounded.

A prerequisite for linear programming theory and algorithms is that the problem variables must be real. This assumption is fulfilled in many real-life applications of linear programming. However, sometimes, meaningful values of the problem variables can be only Integer. For example, consider an example where the optimal value of the production of items has to be computed. In such a problem, the output should represent an integer number of items to be produced. In these situations, imposing integrality requirements on some of the variables makes the problem belong to ’Mixed Integer Programming (MxIP)’ class.

MxIP problems are categorized under the class of ‘NP-complete’ problems which exhibit an extremely high computational complexity. In solving MxIP problems, sometimes a work-around is employed by relaxing the integrality requirements. This process is known as ‘relaxing’ of the corresponding MxIP problem. With the help of this relaxation a near-optimal solution is achieved by judiciously rounding off the fractional values of integer variables in the optimal solution. Such an approach introduces relatively a small rounding off error provided the typical values for integral variables are in the order of tens or above.

6.1.1 Advantages of Linear Programming

The main reasons for wide use and recognition of linear programming to analyze numerous operations research problems are as follows. First, a large class of problems from many areas can either be represented or approximated as linear programming models. Second, a number of well-established and efficient methods are available to solve linear programming problems. Especially, with the recent evolution of computer hardware and sophisticated linear programming software, the solution of even very large problems is fast and inexpensive. Finally, a linear programming problem can be easily extended or limited by manipulating associated constraints. This feature is particularly helpful in carrying out sensitivity analysis through the variations of problem data.

6.1.2 An Illustrative Example

The general process for a simple linear programming problem is to graph the constraints to create a walled-off area on the x- & y-plane, termed as 'feasibility region'. Then identify the coordinates of the corners of that feasibility region and evaluate the objective function for these coordinate values to find out the maximum / minimum value. This process is illustrated here for a simple example of linear programming problem, described as follows: Find the values of x_1 and x_2 which maximize the sum term $x_1 + x_2$, subject to constraints $x_1, x_2 \geq 0$ and

$$3x_1 + x_2 \leq 0$$

$$x_1 - x_2 \leq 1$$

$$x_1 + 2x_2 \leq 7$$

Figure 6.1 shows the plot of the above described constraint set and the resulting feasibility region (x_1 is plotted in x-plane and x_2 is plotted on y-plane). It has been proved that a linear objective function always takes on its maximum / minimum value at a corner point of the feasibility region, provided the constraint set is bounded. Although in some cases the maximum / minimum may occur along an entire edge but it still occurs at the corner point of that edge as well.

In the above described example, the coordinates of corner points of the four-sided feasibility region are as follows: (0,0), (1,0), (3,2), and (1,3). The corresponding value of the objective function ($x_1 + x_2$) for these coordinates are as follows: 0, 1, 5, and 4, respectively. Hence, the desired optimal or maximum value is 5 which is obtained when $x_1 = 3$ and $x_2 = 2$. Moreover, the minimum value for the objective function would be 0 when $x_1 = 0, x_2 = 0$.

6.1.3 Simplex Method

A two variable linear programming problem can be solved easily using the graphical method outlined in the previous subsection. A bit more complex problems can be addressed by using algebraic techniques to solve systems of linear equations. However, the real world problems may comprise hundreds of variables which cannot be solved efficiently using the aforementioned techniques. This is where the Simplex Method comes into the picture. This method was introduced by George Dantzig which tests the adjacent vertices of the feasibility region in sequence so

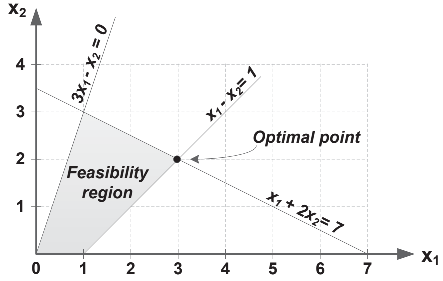


Figure 6.1: Graph of a system of linear equations. Identifying the feasibility region and optimal point.

that at each new vertex the value of the objective function improves or remains unaltered. The simplex method is generally very efficient and takes $2n$ to $3n$ iterations in solving a problem with n equality constraints. The worst case complexity of simplex method has been shown as exponential in certain rare cases.

In addition to the simplex algorithm, interior-point methods are also widely employed to solve linear programming problems. As explained earlier, the simplex methods reach the optimum by traversing around the boundary of the feasibility region in search for the extreme points of the feasibility region. In contrast to this, interior-point methods approach the optimal solution from the strict interior of the feasibility region. Generally, these methods consist of a self-concordant barrier function used to encode the convex set and exhibit polynomial complexity for both average and worst cases. The interior-point method was first invented by John von Neumann and the later enhanced by Narendra Karmarkar[Kar84]. Though various forms of interior-point method exist today, Mehrotra's predictor-corrector [Meh92] method is considered among the best interior-point methods which is competitive with the simplex method, especially for large scale problems.

6.1.4 Software Tools

A computer software which employs certain algorithms to solve linear programming problems is called 'Linear Programming Solver' or LP solver. In the following, a few of the popular LP solvers are listed.

- **lp_solve** [lRG13]: This a free linear (integer) programming solver provided under the 'GNU lesser general public license'. Basically, it is a library (a set of software routines) which can be invoked from almost any computer programming language like C/C++, C#, Java, .NET, Delphi, VisualBasic etc.

lp_solve's capabilities are not restricted to model size, i.e., number of variables and constraints. The software package also includes IDE (Integrated Development Interface) to access lp_solve functionality using a graphical user interface.

- **CPLEX** [IBM13a]: The CPLEX solver has been named after the sim***plex*** method implemented in the ***C*** computer programming language. However, today it also provides additional methods to solve operations research problems. In fact, it is a state-of-the-art LP solver for which the size of problem is merely limited by the computer capacity. This commercial tool has been designed to solve large scale, complex problems where other LP solvers either fail or become unacceptably slow. A free version of CPLEX is available for students but is limited to 300 variables and constraints.

- **GLPK** [GLP13]: The name stands for GNU Linear Programming Kit. This is another freely available tool licensed under the 'GNU General Public License' and is organized in the form of a callable library. The GLPK is a simplex-based solver which is capable of solving large scale linear, MIP, and other related problems. The packet includes a stand-alone LP solver as well as an Application Programming Interface (API) component.

Other mentionable LP solvers include FortMP [For13], Gurobi [GUR13], MINOS [MIN13], and KNITRO [KNI13].

An LP solver accepts a linear programming problem as an input only if described according to a certain modeling language for mathematical programming. There are numbers of options available in this regard, e.g., AMPL (A Mathematical Programming Language) [A M13], OPL (Optimization Programming Language) [OPL13], GAMS (General Algebraic Modeling System) [The13] etc. In addition, there are various file formats which can be used to present the linear programming problem and archive it, e.g., MPS (Mathematical Programming System), .nl format, .sol format, etc.

In this work, the AMPL modeling language has been used in conjunction with CPLEX and lp_solve during the development and test phase. However, for the purpose of integration with the OPNET simulator an educational version of 'IBM ILOG CPLEX Optimization Studio' [IBM13b] was employed. The OPNET simulator uses the CPLEX API for the C-language to interface with the LP solver.

6.2 Analytical Modeling of Network Access Links

In a wireless access network, frequency spectrum and its usage time are the main network resources which are shared by its users. Considering the common practice where the access networks are assigned with a fixed amount of frequency spectrum, each access network has a given number of network resources. For example, an LTE access network may be installed with 5MHz spectrum bandwidth, or an IEEE 802.11a network typically operates with 16.6MHz bandwidth. Having the fixed network resources, a network's capacity and its performance, in turn, depends on the fact how efficiently these spectrum resources are utilized. For example, if the users have good channel conditions, they can employ higher modulation schemes to transmit more data bits for a given amount of network resources. In other words, good channel conditions allow to achieve higher spectral efficiency. The opposite is true for the users who are suffering from bad channel conditions.

Owing to the above facts, in multihoming scenarios, a strategy to increase network capacity is to serve a user over that particular network path which costs less network resources. This strategy is an advanced extension of the MaxT scheduling technique discussed in Chapter 4. The original MaxT exploits only the user diversity, but the proposed strategy, additionally, exploits the diversity of multiple access networks in order to attain high spectral efficiency. This work adapts the term "network path cost" to represent the required network resources to offer a certain data rate. The network path cost is described in different units for different access networks. For example, it is described in the units of [second] in WLAN networks and in term of [PRB] in LTE networks. This will be further elaborated over the next subsections.

The network path cost for a user can be computed using the pertaining channel condition information. Such information is accessible through cross-layer communication from the MAC layer of the corresponding access technology. This information along with the other knowledge about the design and operation of the access technology paves the way to an accurate estimation of the network path cost. Over the next subsections, it is discussed how the network path cost can be computed for the users in LTE and WLAN networks.

6.2.1 LTE Access Network

LTE performs a managed scheduling of available bandwidth resources. These resources are assigned to users in terms of PRBs (Physical Resource Blocks). A PRB is the minimum resource unit which can be allocated to a user in LTE. Based on the allocated frequency spectrum size, LTE has a given number of PRBs. The LTE

MAC scheduler residing at the eNodeB schedules these PRBs using 1ms Transmission Time Interval (TTI). In Section 4.2.1 it has been discussed with details how the LTE MAC scheduler allocates resources to the users. In the following, a short summary of the process is outlined.

- In first step, the buffer status report is received for all users. The buffer status report indicates which users are candidates for transmission in a particular TTI. In downlink, the occupancy of buffers at RLC/PDCP layer at eNodeB is used to indicate whether there exists user data to be transmitted. In uplink, user terminals periodically send their buffer status reports to the MAC scheduler at the eNodeB.
- With the help of buffer status reports, the MAC scheduler compiles a list of users which are candidates for resources in this particular TTI. The next step is to categorize them into GBR and non-GBR users.
- The GBR users usually have the highest priority. Therefore, they are served right away with the required resources.
- The remaining resources will be assigned to non-GBR users. For this purpose, these users are sorted according to their assigned bearer priority. The users with the higher priority are allocated with the available resources using a scheduling scheme like Round Robin, Proportional Fair, etc. In the resource allocation process, the MAC scheduler may also consider the user channel conditions which can be described in terms of the MCS (Modulation & Coding Scheme) index. There is a predefined range of MCS indices which can be found in 3GPP standards [36.12].
- The assigned PRB count and MCS value of a user are used to lookup the Transport Block Size (TBS) from a table defined in the 3GPP specifications [36.12]. This is a two dimensional table where each row lists TBS sizes corresponding to the number of PRBs for a particular MCS index.
- The looked-up TBS value defines the MAC frame size to be transmitted in that TTI for a particular user. This information is conveyed to the RLC layer which delivers an RLC PDU of the indicated size to MAC layer. After attaching the necessary header, the MAC layer forwards the MAC PDU to the Physical layer for transmission.

The above discussion implies that the knowledge of TBS values of a user in a time window can help calculate that user's throughput as also illustrated in Figure 6.2. In addition, the figure also depicts that for a given TBS index, the achievable LTE throughput value has almost a linear relationship with the number of

PRBs. When described mathematically, this relationship can be used to determine the required number of PRBs (say q_{lte}) per TTI to achieve a certain data rate R_i [Mbps] for a user having TBS index i. That is,

$$q_{lte} = \alpha_i \cdot R_i + \beta_i \tag{6.4}$$

The above is a linear equation which can be employed to represent any of the TBS curves in Figure 6.2. Although these curves are not the straight lines, but can be approximate to straight lines without significant loss of accuracy. The α in the above equation is the slope of the approximated straight line described in units of PRB/Mbps. The β is that line's intercept at the y-axis and represents number of PRBs. Both α and β together determine the network path cost of a user's LTE access link. The valid value range for q_{lte} is $1 < q_{lte} \leq PRB_{max}$, where PRB_{max} is the maximum number of available PRBs.

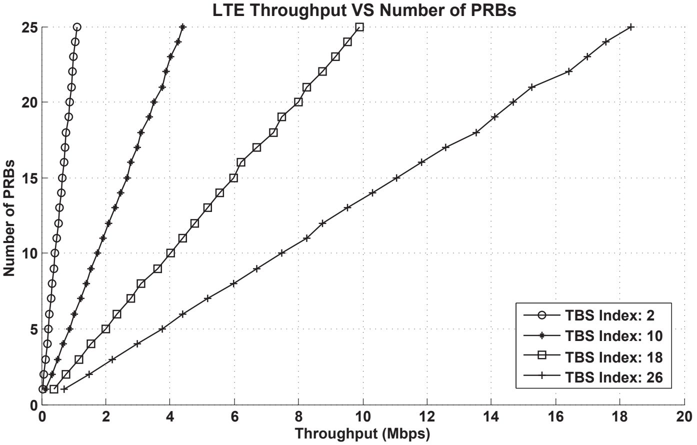


Figure 6.2: Relationship of LTE air interface throughput and number of PRBs for different TBS index values [36.12]. Each curve represents one TBS index.

Appendix Chapter C provides the details of curve fitting data used to define a highly accurate linear relation between number of PRBs and achievable throughput at different TBS indices.

6.2.2 WLAN Access Network

Section 2.6 described the operation of WLAN MAC protocol in great detail. It was highlighted there that the 802.11 MAC uses one of the following three techniques to provide channel access control mechanisms: (i) Point coordination function (PCF) (ii) Distributed coordination function (DCF), and (iii) Hybrid coordination function (HCF). PCF is not part of the Wi-Fi Alliance interoperability and, therefore, is rarely found implemented on any portable device. As far as HCF is concerned, it was originally introduced for the IEEE 802.11e standard, but it is hard to find any compliant hardware due to higher implementation costs. In recent times, although the 802.11n standard has incorporated the HCF mechanism and is becoming increasingly popular, but to-date it is available only on a limited number of portable devices. The statistics show that a dominant percentage of today's portable Wi-Fi capable devices operate in the DCF mode of 802.11a/b/g. Therefore, this work focuses on the DCF mode of operation when modeling WLAN access link. Moreover, owing to the fact that three flavors of 802.11, i.e., a, b and g, follow very similar procedures in medium access mechanism, 802.11a will be considered as an example in following discussions.

DCF has two channel access mechanisms (i.e., the basic access and RTS/CTS) as explained in the following description of for packet transmission. In WLAN with several active users, the sender must sense the medium activity before transmitting a packet to ensure that there is no transmission from other stations. If the medium is sensed idle, the sender can transmit. If the medium is busy the sender continues monitoring the medium until it is sensed idle for DIFS time period. After this the sender has to wait for another time period determined by the random back-off interval. This is done to minimize the collisions from other stations. At this stage, if basic scheme of DFC is used then the sender simply transmits the data packet. However, if RTS/CTS scheme of DFC is being followed, the short RTS and CTS packets are exchanged to reserve the medium before the transmission of data packet. The use of RTS/CTS scheme reduces the probability of collisions as well as fixes the notorious 'hidden terminal' problem.

After providing an understanding of the packet transmission procedure in WLAN, the process of network path cost estimation is described over the next subsections. First, a 'pure' downlink scenario is considered where the users just receive data from the access point and do not transmit anything over the WLAN access network. In other words, there is only one transmitter (i.e., the access point) and therefore no medium contention exists. Second, an uplink scenario is taken into the consideration where all users have data to transmit and, therefore, a medium contention situation is created.

6.2.2.1 Downlink Communication

Consider a WLAN access network comprising a station associated with an access point. Assuming that the station is just receiving a downlink traffic flow from the access point, there is no contention for medium access. In such a case, the transmission of a single data packet of average size $E[P]$ [bit] size requires T_s seconds. Considering the RTS/CTS scheme of DCF, this time also includes the transmission of control frames, i.e., RTS, CTS, SIFS, DIFS, and ACK frames. Figure 2.23 graphically depicts the transmission of data packets in such a scenario. The value of T_s is straightforward to compute, i.e.,

$$T_s = T_{\text{backoff}} + T_{\text{DIFS}} + T_{\text{RTS}} + T_{\text{CTS}} + T_{\text{data}} + 3 \cdot T_{\text{SIFS}} \tag{6.5}$$

$$T_{\text{backoff}} = \frac{W_{min} - 1}{2} \cdot \sigma \tag{6.6}$$

$$T_{\text{DIFS}} = T_{\text{SIFS}} + 2 \cdot \sigma \tag{6.7}$$

Here T_{DIFS} and T_{SIFS} represent the duration of DIFS and SIFS frame space, respectively. The T_{RTS}, T_{CTS}, and T_{ACK} are the duration of RTS, CTS, and ACK control frames, respectively. σ is the duration of the 'slot time' as defined in 802.11a standard and T_{backoff} represents the time spent in back-off phase. The T_{data} is the time required to transmit a single IP data packet including the MAC & PHY headers. Its value is calculated considering the transmission rate at the physical layer. The W_{min} is the minimum value of the contention window in 802.11a. The numerical values of these parameters are defined in 802.11a specifications and have been listed in Table 6.1 for reference. The duration of control frames (i.e., RTS, CTS, and ACK) varies along with the physical layer data rate of the involved users. The computed values of these durations is shown are Table 6.2.

Table 6.1: MAC/Physical layer parameters of 802.11a.

SIFS	**SlotTime**	**RTS**	**CTS**	**ACK**	**W$_{\text{min}}$**	**W$_{\text{max}}$**
16 μs	9 μs	160 bit	116 bit	116 bit	16	1024

It is clear that the 802.11 MAC follows the Time Division Multiple Access (TDMA) scheme where the users share the wireless access medium for short periods of time. Considering resource allocation time intervals of 1 second, a user needs an exclusive medium access for a T_{alloc} fraction of that interval to achieve a unitary data rate of 1 bit/sec. The value of T_{alloc} actually determines the network path cost and has its direct dependence on the transmission time of a packet, i.e.,

Table 6.2: Duration of control frames in 802.11a for different physical layer data.

Data Rate (Mbps)	Modulation	Bits per symbol	RTS Duration		CTS/ACK Duration	
			Symbols	μs	Symbols	μs
6	BPSK	24	7	28	5	20
9	BPSK	36	5	20	4	16
12	QPSK	48	4	16	3	12
18	QPSK	72	3	12	2	8
24	16-QAM	96	2	8	2	8
36	16-QAM	144	2	8	1	4
48	64-QAM	192	1	4	1	4
54	64-QAM	216	1	4	1	4

T_s. The value T_{alloc} amounts to

$$T_{alloc} = \frac{T_s}{E[P]} \tag{6.8}$$

Moreover, the network path cost to support a data rate of R [bps] over this WLAN access link will be as follows

$$q_{wlan} = \frac{T_S}{E[P]} \cdot R \tag{6.9}$$

The valid value range for the network path cost is $0 < q_{wlan} \leq 1$ second. Any computed value falling outside this range indicates that the access link cannot support the desired data rate.

6.2.2.2 Uplink Communication

In uplink communication, the users have to compete for medium access in order to perform transmissions. A network path cost can also be computed here if the packet transmission delay is known for the access link. However, due to the random back-off time and packet collisions, the time required to transmit a packet is highly variable. In [Bia00], Bianchi has used a two dimensional Markov chain model to compute the achievable throughput of 802.11b network. The analysis considered a number of active stations having the same channel conditions and traffic pattern. Chatzimisios et. al. [P. 02] has extended that model to calculate the average packet delay. This work further extends the analysis of Chatzimisios et. al. in order to support 802.11a PHY data rates as well as to make it applicable to user mixtures operating at different PHY data rates. In the following, a summary of the analysis performed by Chatzimisios et. al. is described.

The mathematical analysis assumes a network of n contending stations where each station always has a packet to transmit. The analysis has been divided into two distinct parts. In first part, the behavior of a single station with the help of a Markov model is analyzed. The outcome of this analysis is the stationary probability τ with which a station transmits packet in a randomly chosen slot time. This probability is independent of the access mechanism (i.e., Basic or RTS/CTS). Then, by studying the events that can occur within a generic slot time, the average packet delay for both Basic and RTS/CTS access methods is expressed as a function of the computed value τ, i.e.,

$$\tau = \frac{2 \cdot (1-2p)}{(1-2p) \cdot (W+1) + pW \cdot (1-(2p)^m)} \tag{6.10}$$

In Equation (6.10) W_{min} is the minimum contention window size and m is the "maximum back-off stage" so that the maximum contention window is $W_{max} = W_{min} \cdot 2^m$. The p is the conditional collision probability, i.e., it is the probability of a collision seen by a packet being transmitted on the channel or access medium.

$$p = 1 - (1-\tau)^{n-1} \tag{6.11}$$

With the help of numerical techniques the system of nonlinear equations, i.e., (6.10) and (6.11) is solved. This provides a value of τ in terms of three parameters, i.e., m, n, and W_{min}. The value of τ is then used to compute the probability (P_{tr}) that there is at least one transmission in the considered slot time. Moreover, it is also used to compute the probability (P_s) that an occurring packet transmission is successful.

$$P_{tr} = 1 - (1-\tau)^n, \quad \text{and} \tag{6.12}$$

$$P_s = \frac{n \cdot \tau \cdot (1-\tau)^{n-1}}{1-(1-\tau)^n} \tag{6.13}$$

The relationships of P_{tr} and P_s are used to calculate $E[slot]$ which is the average length of a slot time. It should be marked, unless ambiguity occurs, that the term slot time refers to either the (constant) value σ, or the (variable) time interval between two consecutive back-off time counter decrements, i.e., $E[slot]$. The average length of a slot time is obtained considering that, with probability $1-P_{tr}$ the slot time is empty; with probability $P_{tr}P_s$ it contains a successful transmission, and with probability $P_{tr}(1-P_s)$ it contains a collision, i.e.,

$$E[slot] = (1-P_{tr})\sigma + P_{tr} \cdot P_s \cdot T_s + P_{tr}(1-P_s) \cdot T_c \tag{6.14}$$

here σ is the duration of an empty slot time, the T_s is average time during which medium is sensed busy because of a successful transmission, and T_c is the average time period for which the medium is sensed busy by each station during a collision. Assume that $E[X]$ is the average number of slot times for a successful packet transmission. The value $E[X]$ is found by multiplying the number of slot times the packet is delayed in each back-off stage by the probability to reach this back-off stage. The final form of $E[X]$ is as follows

$$E[X] = \frac{(1-2p)\cdot(W+1)+pW\cdot(1-(2p)^m)}{2\cdot(1-2p)\cdot(1-p)} \tag{6.15}$$

Finally, the average delay of a successfully transmitted packet $E[D]$ is given as follows

$$E[D] = E[X]\cdot E[slot] \tag{6.16}$$

In Equation (6.16) the values of σ, m, and W_{min} can be obtained from Table 6.1. The values of the other two unknown parameters, i.e., T_s and T_c depend on the fact whether the basic or RTS/CTS scheme is chosen.

$$T_s^{bas} = T_H + T_{E[P]} + T_{SIFS} + \delta + T_{ACK} + T_{DIFS} + \delta \tag{6.17a}$$

$$T_c^{bas} = T_H + T_{E[P]} + T_{DIFS} + \delta \tag{6.17b}$$

$$T_s^{rts} = T_{RTS} + T_{SIFS} + \delta + T_{CTS} + T_{SIFS} + \delta + T_H + T_{E[P]} + T_{SIFS} + \delta + T_{ACK} + T_{DIFS} + \delta \tag{6.18a}$$

$$T_c^{rts} = T_{RTS} + T_{DIFS} + \delta \tag{6.18b}$$

where δ is the propagation delay, $T_H = T_{PHYhdr} + T_{MAChdr}$ is the time to transmit the header data associated with the PHY and MAC protocols, and $T_{E[P]}$ is the time to transmit a data packet of mean size $E[P]$.

In the above described analysis of Chatzimisios et. al., it has been assumed that all stations have the same channel conditions and therefore transmit with the same PHY data rate. In a realistic scenario this assumption cannot always be fulfilled. In order to make Equation (6.16) valid for a scenario where the users have different channel conditions and PHY data rates, it must be extended as follows. The direct influence of the PHY data rate on the average packet delay estimation can be observed in the computation of T_s and T_c (see Equation (6.17) and (6.18)). This is where the user PHY data rate determines the value of $T_{E[P]}$ which is the time to transmit the data packet. Therefore, an 802.11a network of users with different PHY data rates can be seen as a queuing system with a single server and a single buffer queue. In this system, the effect of the PHY data rate is directly incorporated into the service rate so that a job is served at the PHY data rate of the corresponding user terminal. The mean service time of such a system can be computed as follows.

$$\widehat{T}_{E[P]} = E\left[\frac{E[P]}{\text{PHY data rate}}\right] \tag{6.19}$$

Furthermore, when stations are transmitting at different PHY data rates, the transmission speed of control frames (i.e., T_H, T_{RTS}, T_{CTS} and T_{ACK}) in the network is limited by the station having the lowest PHY data rate. This implies that the users must transmit control frames at a PHY data rate which can be decoded or understood by all users. However, the data packet is still transmitted by the user's own current PHY data rate.

The value of $\widehat{T}_{E[P]}$ replaces $T_{E[P]}$ in Equation 6.17 and 6.18 to compute the extended values of $\widehat{T}_s$ and $\widehat{T}_c$. They are, in turn, plugged into Equation (6.14) which produces $\widehat{E[slot]}$ so that

$$\widehat{E[slot]} = (1-P_{tr})\sigma + P_{tr}P_s\widehat{T}_s + P_{tr}(1-P_s)\widehat{T}_c \tag{6.20}$$

Owing to the fact that the value of $E[X]$ is independent of the user PHY data rate and therefore does not require any modification, Equation (6.16) takes the following form

$$\widehat{E[D]} = \widehat{E[slot]} \cdot E[X] \tag{6.21}$$

Figure 6.3 shows the average packet delay experienced by the users transmitting in the 802.11a network with RTS/CTS enabled. The solid lines show the estimated values computed using Equation (6.21). The markers on a solid line represent the delay values obtained from the simulation results. It is evident from the figure that the modified model can precisely estimate the mean packet delay when all users have same PHY data rate as well as when users with different PHY data rate are mixed together. The mean packet delay computed with the help of Equation (6.21) can also be used to estimate the average user throughput Y in the network, i.e.,

$$Y = \frac{E[P]}{\widehat{E[D]}} = \frac{E[P]}{E[X] \cdot \widehat{E[slot]}} \tag{6.22}$$

It should be noted that Y is the single average throughput value which will be experienced by each user in the network. The value of Y will decrease sharply when more users will join the network and it will also be influenced by the PHY

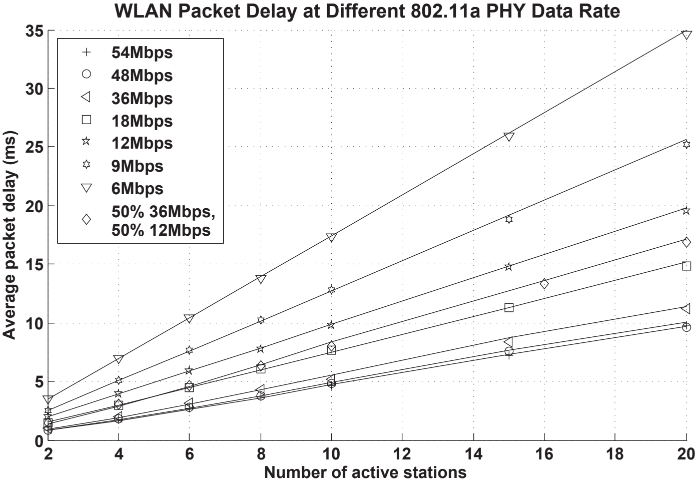


Figure 6.3: Average packet transmission delay experienced by users operating at different PHY data rates in an 802.11a network. The solid lines show the estimated values computed using analytical approach. The markers on lines represent the delay values obtained from the simulation results.

data rates of transmitting users. But, all users will experience the same average throughput Y irrespective of their individual PHY data rates. This can be elaborated with the help of aforementioned queuing model where during the service time of a job other jobs have to wait in the queue. This leads to similar queuing delays for jobs from all stations and hence results in the same average throughput for each of them.

6.3 Problem Formulation in Linear Programming

The development of analytical relations to compute the network path cost for LTE and WLAN paves the way for a problem formulation of optimized resource allocation using linear programming. In the following, a generic and abstract resource allocation problem is formulated in linear programming as a starting point. Later on, the access network specific parameters will be incorporated into it to generate a full-fledged model applicable to the heterogeneous network of LTE and WLAN.

Table 6.3 shows the generic mathematical model in algebraic form. The model works with two data sets, i.e., U which represents the set of multihomed users and

Table 6.3: Generic mathematical model for the resource allocation in algebraic form.

Given

U a set of multihomed users

L a set of access network links

$f(R_{j,l})$ Linear function of network path cost which maps user data rate $R_{j,l}$ to the access networks resources, $\forall j \in U, \forall l \in L$

λ_j Minimum aggregated data rate demand of a traffic flow destined to user j, $\forall j \in U$

Λ_j Maximum aggregated data rate allocation for a traffic flow destined to user j, $\forall j \in U$

Ω_l Available resources on access network l, $\forall l \in L$

Defined variables

$R_{j,l}$ Data rate carried over the access link l to user j, $\forall j \in U, \forall l \in L$

Maximize

$$\sum_{j \in U} \sum_{l \in L} R_{j,l}$$

Subject to

1. $\lambda_j \leq \sum_{l \in L} R_{j,l} \leq \Lambda_j, \quad \forall j \in U$
2. $\sum_{j \in U} f(R_{j,l}) \leq \Omega_l, \quad \forall l \in L$

L which represents a set of access network links available to the user. A linear function $f(R_{j,l})$ of network path cost is used to compute the required network resources in order to support a data rate $R_{j,l}$ for a certain user j over its network access link l. This function makes use of analytical relations of network path cost developed over the previous section. λ and Λ are the parameters which represent the possible range of data rate assignment for a certain user. Specifically, λ is the minimum data rate demand of the user which can be imposed to offer a certain mean data rate for realtime services, like VoIP, video, etc. And Λ enforces an upper limit on data rate assignment which is helpful in restricting data rates of TCP flows as well as to achieve service fairness. Ω_l is the amount of total available resources for an access network l which can be utilized by this model.

The variable $R_{j,l}$ is defined by the model to represent the data rate of a user j over an available access link l. This is the only output of this model when solved using linear programming. The goal of this model is to maximize $R_{j,l}$ by minimizing the associated network path costs so that an optimized network performance is realized.

The data rate assignment by the model is subject to two constraints. First, the assigned data rate must fall within the range defined by λ and Λ. Second, the required network resources to achieve the assigned data rates should not exceed the available resource of the access networks.

In case the available resources are not sufficient to satisfy the minimum data rate requirements of all users, this will result in an infeasible problem. Otherwise, the solution will provide an optimized allocation of network resources for the given users channel conditions. These channel conditions are implicitly incorporated in the network path cost function $f(R_{j,l})$. As the users move, their channel conditions vary due to which the parameters of the problem changes as well. This implies that the problem should be reevaluated after a period of time which is short enough to adapt to variations in user channel conditions. This way, an optimized network performance can be guaranteed for mobile users over time.

6.3.1 Download Communication

This section develops a mathematical model to allocate network resources in a 'pure' downlink communication scenario. In this scenario, the users do not transmit over the WLAN access link, instead they use it to receive traffic from the access point. Hence, there is no contention for the access medium in the WLAN network which allows the use of Equation 6.9 as network path cost function. Luckily, this is a linear function which can be readily used in conjunction with linear programming. As far as the LTE access network is concerned, the linear function described by Equation 6.4 is used as the network path cost function.

Table 6.4 shows the algebraic form of the mathematical model developed for 'pure' downlink scenario through an implementation of generic model. The input parameters α and β are the components of the path cost function for the LTE access link. ϕ represents the path cost for the per unit data rate carried over the WLAN access link. λ and Λ are the limitations on data rate assignments which are set according to user application types. The total number of available resources for LTE access network are represented by Ω. Moreover, for WLAN access network, medium access time of 1 second is considered as the available network resource.

The model defines two variables (R^{lte} & R^{wlan}) to represent the data rate carried over two access links of each multihomed user. In addition, a binary variable E is defined to represent the use of the LTE access link for each user. It acts as an auxiliary variable in the formulation of the path cost function for the LTE access link as shown in the first constraint. This is because β should be added to the network resource consumption only when a user receives data over LTE access links, i.e., the corresponding R^{lte} is non-zero. The fourth constraint in the model determines the value of E as 1 if $R^{lte} > 0$ and as 0 otherwise.

The objective of this model is to maximize the user data rates over both LTE and WLAN access networks. This objective has to be achieved under the six constraints shown in Table 6.4. The first constraint ensures that the total number of PRBs required to serve users over their LTE access link should not exceed the available number of PRBs Ω. The second constraint indicates that the aggregated network path cost of WLAN access links from all users should not rise beyond 1 second limit. The third constraint is imposed to keep the aggregated assigned user data rate in the range defined by λ and Λ. The model can flexibly assign the data rate to the network access links of the users, e.g., the user data rate demands can either be fulfilled by transmission over a single access link that has the lowest network path cost or over both access links simultaneously in order to bundle up their bandwidths. This is indicated by constraint 5 and 6.

6.3.1.1 Simulation Scenarios and Results

In this section, the performance of the proposed scheme for optimized resource allocation is evaluated with the help of a simulation scenario. Figure 6.4 shows an overview of the scenario in OPNET. The system is populated with 12 users generating a rich traffic mixture of: Voice over IP (VoIP), downlink File Transfer Protocol (FTP), Hyper Text Transfer Protocol (HTTP), video conference (i.e., Skype video call), and video streaming. The users move within one LTE eNodeB cell, and within this cell one wireless access point is present. Table 6.5 shows the parameter configuration for this scenario.

The network performance achieved by the linear programming approach will be compared with the other two approaches discussed in Chapter 5, i.e., “3GPP-HO” and “Channel Aware”. It can be recalled that in the “3GPP-HO” approach multihoming is not supported, instead the policy is to serve a user preferably over WLAN access network in the overlapped coverage of WLAN & LTE access networks. In contrast to this, the “Channel Aware” approach makes use of multihoming and flow management to serve users efficiently. In this approach the capacity of each of the user access links is precisely estimated and all available bandwidth resources are bundled together in achieving the best user QoE. Now with the help of the linear programming approach, data rates are assigned to the users in a way that network capacity is maximized as well as the minimum data rate demands are met for all users. For this purpose, the DE_n employs the resource allocation

Table 6.4: Mathematical model for the resource allocation in algebraic form for downlink communication.

Given

U a set of multihomed users

α_j Data rate dependent part of the LTE link cost in [PRB/kbps] for user $j, \forall j \in U$

β_j Data rate independent part of the LTE link cost in [PRB] for user $j, \forall j \in U$

ϕ_j Path cost of WLAN access link in [sec/kbps] for user $j, \forall j \in U$

λ_j Minimum data rate [kbps] demand of a traffic flow destined to user $j, \forall j \in U$

Λ_j Maximum data rate [kbps] allocation for a traffic flow destined to user $j, \forall j \in U$

Ω Number of available PRBs for the LTE access network

Defined variables

R_j^{lte} Data rate in [kbps] carried over the LTE access link to user j, $R_j^{lte} \geq 0, \quad \forall j \in U$

R_j^{wlan} Data rate in [kbps] carried over WLAN access link to user j, $R_j^{wlan} \geq 0, \quad \forall j \in U$

E_j Auxiliary binary variable which hints the use of LTE access link by user j; its value for a user j is 1 if $R_j^{lte} > 0$ and 0 otherwise, $\forall j \in U$

Maximize

$$\sum_{j \in U} \left(R_j^{lte} + R_j^{wlan} \right)$$

Subject to

1. $\sum_{j \in U} \left(\alpha_j \cdot R_j^{lte} + \beta_j \cdot E_j \right) \leq \Omega$
2. $\sum_{j \in U} \left(\phi_j \cdot R_j^{wlan} \right) \leq 1$
3. $\lambda_j \leq \left(R_j^{lte} + R_j^{wlan} \right) \leq \Lambda_j, \quad \forall j \in U$
4. $\left(R_j^{lte} / \Lambda_j \right) \leq E_j \leq \left(R_j^{lte} \cdot 10^{20} \right), \quad \forall j \in U$
5. $0 \leq R_j^{lte} \leq \Lambda_j, \quad \forall j \in U$
6. $0 \leq R_j^{wlan} \leq \Lambda_j, \quad \forall j \in U$

model shown in Table 6.4. At each decision instant, the model is solved using updated parameters of user channel conditions and QoS demands and the resulted data rates are then imposed on the user access links by EE entities.

Figure 6.5 shows the performance of the FTP downlink application in terms

Table 6.5: Simulation configurations for evaluation of the downlink flow management scheme using linear programming.

Parameter	Configurations
Total number of PRBs	25 PRBs (5 MHz specturm)
Mobility model	Random Direction (RD) with 6 Km/h
Number of users	2 VoIP calls, 1 video streaming, 3 Skype video calls, 2 HTTP and 4 FTP downlink users
LTE channel model	Macroscopic pathloss model [25.06], Correlated Slow Fading.
LTE MAC scheduler	Time domain: Optimized Service Aware, Frequency domain: Iterative RR approach [S. 12]
WLAN access technology	802.11a, RTS/CTS enabled, coverage $\approx$ 100 m, operation in non-overlapping channels
Transport network	1Gbps Ethernet links, no link congestion
VoIP traffic model	G.722.2 wideband codec, 23.05kbps data rate and 50frame/s
Skype video model	MPEG-4 codec, 512kbps, 30frame/s, 640x480 resolution, play-out delay: 250ms
Streaming video model	MPEG-4 codec, 1Mbps, 30frame/s, 720x480 resolution, play-out delay: 250ms
HTTP traffic model	100 bytes html page with 5 objects each of 100Kbytes, page reading time: 12s
FTP traffic model	FTP File size: constant 10MByte, as soon as one file download finishes, the next FTP file starts immediately.
TCP configurations	TCP new Reno, Receiver buffer: 1Mbyte, Window scaling: enabled, Maximum segment size: 1300Byte, TCP reorder timer: 50ms
DE_n decision interval	Every 100ms
Data rate demands $[\lambda, \Lambda]$	[200kbps, 25Mbps] for FTP and HTTP users
Simulation run time	1000 seconds, 10 random seeds, 95% Confidence interval

of IP throughput and file download time. It is evident that the “Linear Programming” approach achieves the highest performance among the three competing approaches. The optimized resource allocation strategy of the “Linear Programming” approach helps increase user QoE experience by 25% in terms of file download time compared to “3GPP-HO” approach. The “Linear Programming” approach also outperforms the “Channel Aware” approach by reducing the file download time up to 13%.

Similar conclusions can also be drawn for the HTTP application whose performance has been shown in Figure 6.6. It can be noticed that the HTTP application could attain much less IP throughput compared to the FTP application. This is due to small sized embedded objects of web page. The download of these objects finishes before the TCP connection could achieve the maximum possible through-

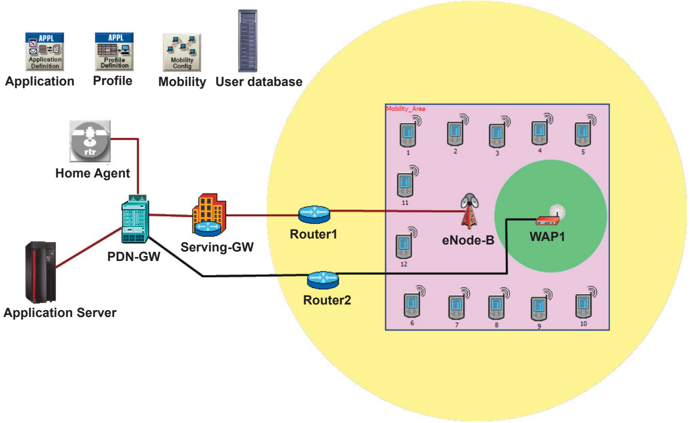
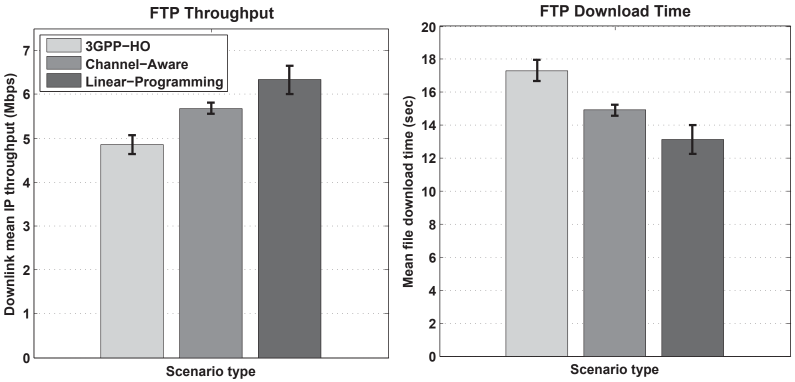


Figure 6.4: Simulation scenario setup in the OPNET simulator. The large circular area shows the coverage of LTE and the smaller circular area represents the WLAN network coverage. The user movement is restricted to the rectangular area.



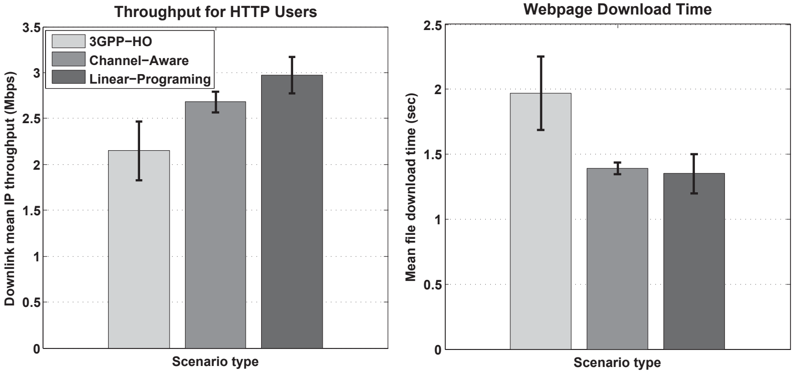
(a) (b)

Figure 6.5: FTP downlink performance comparison for “3GPP-HO”, “Channel Aware”, and “Linear Programming” approaches.

put. Owing to this fact, even the “Linear Programming” approach could not significantly improve user QoE over the “Channel Aware” approach. However, a substantial gain is observed compared to default “3GPP-HO” approach. A large variation in web page download can also be noticed for the “3GPP-HO” approach, the reason of which is as follows. If an HTTP user is found in WLAN coverage, “3GPP-HO” approach serves it solely over the WLAN access link. Now if the WLAN access network is not heavily loaded because there are no FTP users at that instant, then the HTTP users get high throughput and finish page download fast. In other situations, they have to share bandwidth resources with the demanding FTP users and hence page download time elongates.

The performance gain of “Channel Aware” and “Linear Programming” over “3GPP-HO” approach can be attributed to manifold factors. For example, both of them are capable of aggregating bandwidth resources from multiple access links, they can accurately estimate the capacity of an access link and utilize it accordingly, they can periodically reevaluate their assessment about the network conditions and the capacity of user access links, etc. In addition, “Linear Programming” is capable of performing optimized resource allocations. Among all of them, the ability to estimate user access link capacity is the feature which helps these approaches to establish a definite superiority over the default “3GPP-HO” approach. Without access link capacity estimation, the buffers at the air interface could have very large occupancy. On the one hand, employing large buffers leads to long queuing delays which adversely affects realtime applications in particular. On the other hand, keeping buffer capacity small causes numerous packet drops which degrades, especially, the performance of TCP based applications. By exploiting the knowledge of user access link capacity, “Channel Aware” and “Linear Programming” approaches just send the sufficient amount of data to the air interface schedulers which avoids both the large queuing delays and packet drops. In addition, it also minimizes the risk of losing the buffered packets at the access point during the instants of link failure or vertical handover.

The user QoE for VoIP application has been depicted in Figure 6.7. The Box plot shows the sample values of MOS computed for a user employed wideband codec. It can be seen that both “Channel Aware” and “Linear Programming” deliver excellent performance by keeping MOS values at the maximum level for most of the time. However, the “3GPP-HO” approach fails to achieve a matched performance. Though the median value lies close to the MOS score 4.0, the other values show quite lower score due to long queuing delays at WLAN access point. Even some of the outliers fall below MOS score 2.2 which could be very annoying for the users. The reason for the “3GPP-HO” approach to sometimes achieve a very high MOS score (i.e., above 4.0) lies in the fact that when the VoIP users are being served over the LTE access network, they are provided with the guaranteed QoS service. The problem arises only when these users are handed over to the WLAN access network.



(a) (b)

Figure 6.6: HTTP downlink performance comparison for “3GPP-HO”, “Channel Aware”, and “Linear Programming” approaches.

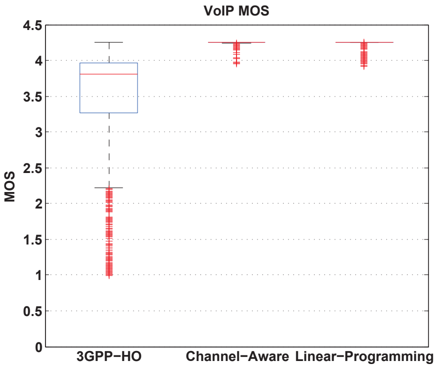
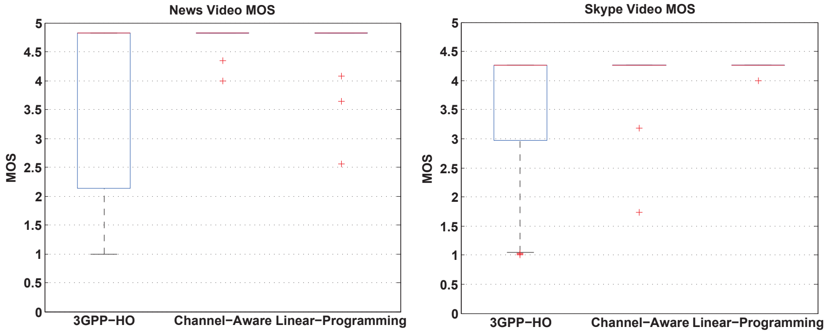


Figure 6.7: Downlink performance comparison of VoIP application for “3GPP-HO”, “Channel Aware”, and “Linear Programming” approaches.

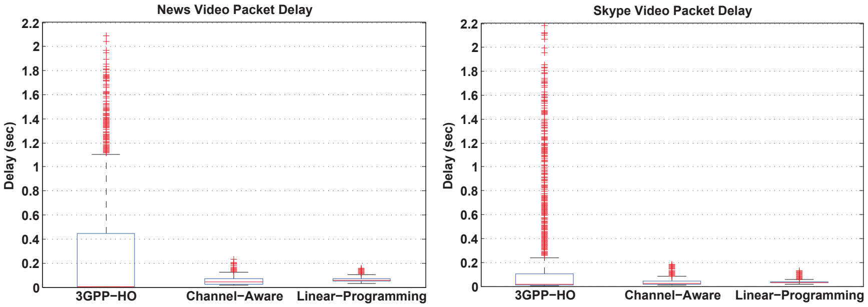
The “Linear Programming” approach also offers excellent user QoE for video applications. This can be confirmed by referring to Figure 6.8 which shows the Box plot of user experienced MOS score for their video applications, i.e., video conferencing and video streaming. Almost all video quality evaluations result in the best MOS score for video applications for both “Channel Aware” and “Linear Programming” approaches. However, “3GPP-HO” fails again to offer an acceptable performance for video application users. Its performance pattern is similar to that of the VoIP application, i.e., the median value stays at the best MOS score while the 3^{rd} & 4^{th} quartiles show the suboptimal performance. As already explained for VoIP case, this phenomenon can be understood with the help of end-to-end packet delay plots shown in Figure 6.9. Considering the play-out delay limit of 250ms, any packet arriving later than this limit is assumed as lost by the video quality evaluation mechanism. Such packet losses in turn lead to performance degradation. It is evident from Figure 6.9 that a large number of packets experience more than 250ms delay for the case where "3GPP-HO" approach is employed. It is mainly the large MAC queue occupancy at the WLAN access point which is the main reason behind these delays. During the times when video users are served over LTE packet end-to-end delays remain under control due to QoS aware scheduling employed in the LTE MAC scheduler. In these situations, the users are satisfied with the service as indicated by the best MOS score. However, during the time when users receive their video application traffic over the WLAN access link, the chances are high that they have to encounter a congested network.



(a) (b)

Figure 6.8: Downlink performance comparison of video applications for "3GPP-HO", "Channel Aware", and "Linear Programming" approaches.

Now that the performance of all applications has been observed, it can be inferred that both the "Linear Programming" and "Channel Aware" approaches provide similar performance for realtime applications. However, the "Linear Programming" approach excels when it comes to non-realtime applications like FTP, HTTP etc. This is because realtime applications have stringent QoS requirements which have to be fulfilled at all costs in order to keep users satisfied. Therefore,



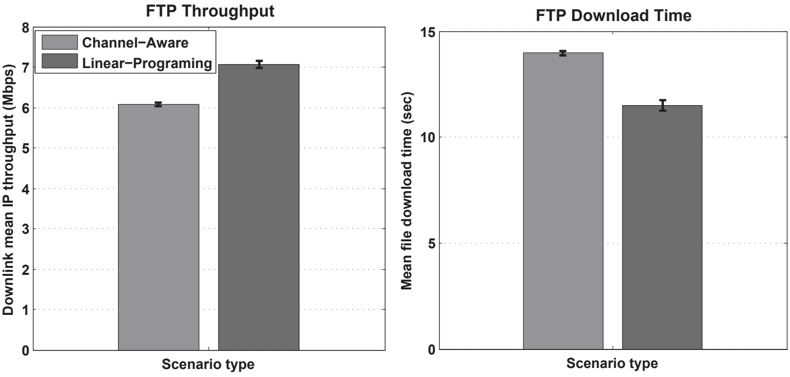
(a)

(b)

Figure 6.9: Packet delay comparison of video applications for “3GPP-HO”, “Channel Aware”, and “Linear Programming” approaches.

both approaches preferably deliver the data rate demands of the realtime services. However, the “Linear Programming” approach, with the help of optimized resource allocation techniques, manages to offer these data rates by consuming lower network resources. This way, larger network resources are made available to non-realtime application users in order to enhance their QoE as well as to increase network capacity.

The discussion on downlink communication is concluded by comparing the performance of “Channel Aware”, and “Linear Programming” approaches in scenarios where only FTP users exist within an area of complete LTE and WLAN coverage overlap. Each of these seven FTP users download 10Mbyte files continuously, i.e., as soon as one file download ends, a new file download is started. Figure 6.10 shows the FTP application throughput and file download time as experienced by the users. In this particular scenario, the “Linear Programming” approach manages to achieve 16% higher throughput compared to the “Channel Aware” approach. This is slightly higher than 13% which was observed in case of the previous scenario with mixed traffic. The reason behind this improved performance is the lower ‘minimum data rate’ requirement of FTP users compared to video users. Owing to the fact that ‘minimum data rate’ requirements must be fulfilled, the users with bad channel conditions consume lots of resources in achieving that data rate. On the other hand, if this requirement is less, fewer network resources will be consumed even when a user is suffering from bad channel conditions.



(a) (b)

Figure 6.10: FTP downlink performance comparison between the “Channel Aware” and the “Linear Programming” approaches.

6.3.1.2 Sensitivity Analysis of DE_n Decision Intervals

It has been explained that the decision making entity (DE_n) of the flow management overlay architecture which resides in the network is responsible for the resource management decisions. These decisions are based on the network information (e.g., user channel conditions, application QoS demands, traffic load, congestion, etc.) supplied by the information management entities (IE) installed at different monitoring points across the network and at UE. Owing to the dynamic load conditions of the networks and variable channel conditions of mobile users, the information provided by IEs has a short validity period after which it must be refreshed. In this way, the resource management decisions made by DE_n at a certain time instant remain no longer the optimal decisions as soon as the IE supplied information on which these decisions were based becomes obsolete. Therefore, the IEs must send the fresh information to the DE_n periodically to prevent aforementioned situation. As soon as DE_n receives the updated information, it revises its resource management decisions and enforces them to achieve an optimal network performance over time. There can be two reasons that the DE_n receives a delayed information from IE entities set up across the network. First, there exists congestion in the network due to which it takes longer for the information data to reach DE_n. Second, an operator wants to cut down the signalling traffic load generated by that information element by reducing the frequency of updates. In such situations the question is how long is the validity period of the IEs provided information and what could be the consequences if resource management decisions are not updated in due time?

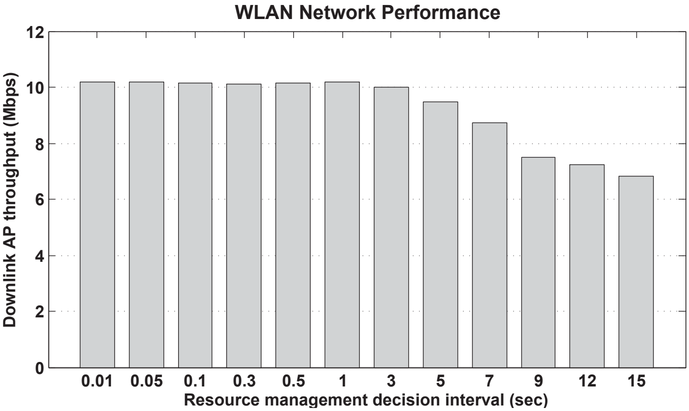


Figure 6.11: Downlink throughput variations of WLAN access point for different values of the DE_n decision interval. The “Linear Programming” approach has been used for resource management decisions.

The above questions can be answered with the help of simulation results shown in Figure 6.11, 6.12, and 6.13. For this purpose the same simulation scenario is employed which has been discussed earlier in this section and whose configurations has been listed in Table 6.5. In this simulation study, the DE_n decision interval is varied from 10ms to 15s and the “Linear Programming” approach is used to make resource management decisions. It is clear that the optimal resource management decisions should provide an optimum network capacity for both WLAN and LTE networks. It can be seen in the Figure 6.11 that the WLAN access point throughput which represents that network’s capacity, remains at the optimum point until the DE_n updates the resource management decisions at least every 1 second. Any further delays cause the system throughput to reduce. This is due to the user movements (at 6 km/h speed) because of which their channel conditions vary and hence their PHY data rates change. When these variations are not tracked by the DE_n due to lack of fresh information the optimal resource management decisions cannot be carried out. For example, if a user’s PHY data rate has increased from 24Mbps to 36Mbps during the elongated decision interval, his throughput will not be upgraded by the DE_n until the information about this change reaches DE_n and it revises the resource management decisions. Similarly a high traffic volume will be continuously sent to a user whose PHY data rate has decreased from 36Mbps to

24Mbps during the decision interval. This will cause that user to experience large packet delays due the fact that some of the data is being buffered at the access point due to PHY data rate downgrade. Due to such events the WLAN network performance degrades. It can be noticed from the Figure 6.11 that increasing the DE_n decision interval to 15s causes approximately 30% degradation in the network capacity.

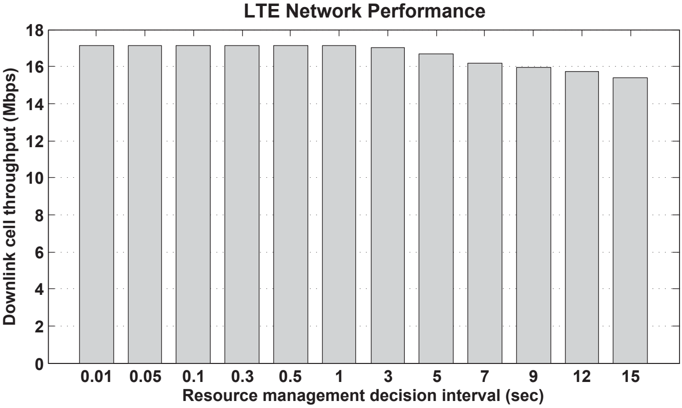


Figure 6.12: LTE downlink cell throughput variations for different values of the DE_n decision interval. The “Linear Programming” approach has been used for resource management decisions.

Figure 6.12 shows a similar behavior for the LTE network which may undergo cell throughput degradations due to elongated DE_n decision intervals. The cell throughput can reduce up to 9% compared to its optimal value if a decision interval of 15s is considered. However, it can be observed that the performance of the LTE network is less sensitive to DE_n decision intervals compared to WLAN network. For example, a noticeable capacity degradation is seen for the LTE network for a 5s decision interval while such a behavior was observed for the WLAN network at a 3s decision interval. The reason for this phenomenon lies in the fact that WLAN has a smaller coverage area and the user the PHY data rates decrease sharply when commuting away from the access point. Therefore, the information about user PHY data rate becomes stale relatively faster and, in turn, affects the optimality of resource management decisions.

It has been seen that when the resource management decisions are not optimal, capacities of access networks are reduced. A natural consequence of this will be deteriorations in the perceived user QoE. An example of this is illustrated in

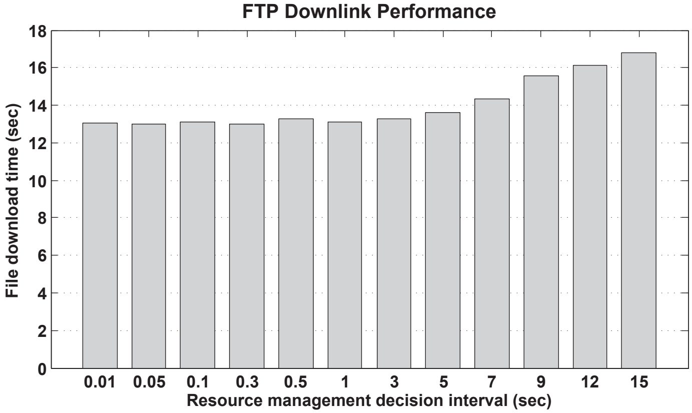


Figure 6.13: Mean file download time experienced by FTP users. The figure shows how the FTP performance is affected by different values of DE_n decision interval. The "Linear Programming" approach has been used for resource management decisions.

Figure 6.13 which shows the mean file download time for FTP users. It can be observed that the users have to wait longer for file download completion if DE_n fails to make optimal resource management decisions. Actually, this is because of the reduced network capacities that the users can no longer achieve high throughput and hence suffer from QoE degradations. The simulation results show that file download time can increase up to 24% compared to the optimal value, if DE_n makes resource management decisions every 15s.

The above discussion implies that a decision interval of at most 1 second should be employed in order to achieve an optimal network performance. However, this value is specific to the simulation scenario being discussed and may not hold for other scenarios. For example, in the current scenario the users are moving with a speed of 6km/h following the random direction mobility model. If this configuration is changed or some additional dynamic background traffic load is added to the network, a rerun of this sensitivity analysis will be needed.

6.3.2 Uplink Communication

In 'pure' uplink communication, the users attached to the WLAN access network may compete for the medium access which results in packet transmission collisions and elongated back-off periods. In this case, the network path cost must be determined from the analytical relation presented in Equation 6.22. However, it is clearly a nonlinear relation which must be linearized using some work-around for use with linear programming. For this purpose, the Equation (6.20) is split into two parts. The first part (i.e., $f_x(n)$) depends only on n, a variable which represents the total number of active stations in the WLAN network. The other part incorporates both a function of n (i.e., $f_y(n)$) and the variable $\widehat{T_s}$. So that,

$$\widehat{E[slot]} = f_x(n) + f_y(n) \cdot \widehat{T_s}$$
$$\text{where} \quad f_x(n) = (1 - P_{tr})\sigma + P_{tr}(1 - P_s) \cdot T_c,$$
$$f_y(n) = P_{tr} \cdot P_s$$

In the above equation, $\widehat{T_c} = T_c^{rts} = T_c$ which holds due to the assumption that WLAN network operates with RTS/CTS enabled. Moreover, Equation (6.15) shows that $E[X]$ is a function of only one variable, i.e., n. Owing to this fact, it is possible to rewrite Equation (6.21) in the following form:

$$\widehat{E[D]} = E[X] \cdot \widehat{E[slot]} = f_1(n) + f_2(n) \cdot \widehat{T_s} \tag{6.23}$$
$$\text{where} \quad f_1(n) = E[X] \cdot f_x(n)$$
$$\text{and} \quad f_2(n) = E[X] \cdot f_y(n)$$

In order to simplify the relation in equation (6.23), $f_1(n)$ and $f_2(n)$ are approximated using 3rd order polynomial curve fitting as shown below:

$$f_1(n) \approx A_{11} \cdot n^3 + A_{12} \cdot n^2 + A_{13} \cdot n + A_{14},$$
$$f_2(n) \approx A_{21} \cdot n^3 + A_{22} \cdot n^2 + A_{23} \cdot n + A_{24}$$

where all occurrences of A_{xx} represent constant value numbers. Fig. 6.14 shows that the curve fitting process generates an accurate approximation for $f_1(n)$ and $f_2(n)$ with norm of residuals as 3.9×10^{-5} and 0.14, respectively.

Substituting the approximated functions for $f_1(n)$ and $f_2(n)$ in Equation (6.23) produces

$$\widehat{E[D]} = A_{11} \cdot n^3 + A_{12} \cdot n^2 + A_{13} \cdot n + A_{14} + \left(A_{21} \cdot n^3 + A_{22} \cdot n^2 + A_{23} \cdot n + A_{24}\right) \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} T_{S_i} \tag{6.24}$$

Substituting the value of $\widehat{E[D]}$ in Equation (6.22) and assuming $E[P] = G$, the following relation can be derived after a few manipulation steps of algebra.

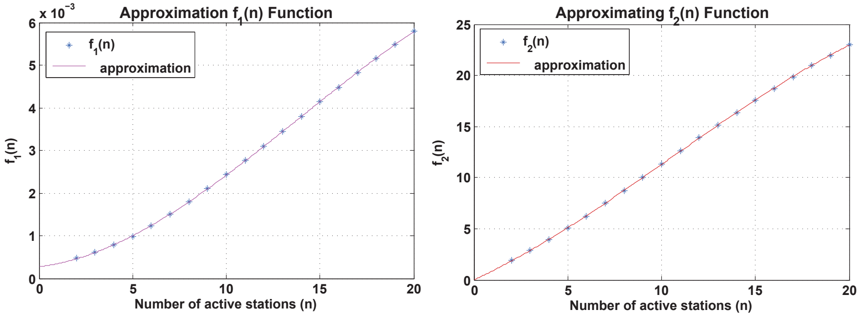


Figure 6.14: Approximation of $f_1(n)$ and $f_2(n)$ using polynomial curve fitting.

$$n \cdot G = Y \cdot \{A_{11} \cdot n^4 + A_{12} \cdot n^3 + A_{13} \cdot n^2 + A_{14} \cdot n + (A_{21} \cdot n^3 + A_{22} \cdot n^2 + A_{23} \cdot n + A_{24}) \cdot \sum_{i=1}^{n} T_{S_i}\} \quad (6.25)$$

The variable n, in Equation (6.25), can be replaced with a summation of binary variables F_i which represents whether a station i is transmitting over the WLAN access or not. If there are a total of Z number of users in the WLAN network out of which only n users are active then

$$n = \sum_{i=1}^{Z} F_i \quad (6.26)$$

A second order term of n can be linearized as follows

$$n^2 = \left(\sum_{i=1}^{Z} F_i \right)^2 = \sum_{i,j=1}^{Z} F_i \cdot F_j = \sum_{i,j=1}^{Z} \chi_{i,j}^{\mathrm{F2}} \quad (6.27)$$

where $\chi_{i,j}^{\mathrm{F2}}$ is itself a binary variable and represents the product of two binary variables, i.e., F_i and F_j. Its value is determined by following three constraints:

$$\chi_{i,j}^{\mathrm{F2}} \le F_i, \qquad \chi_{i,j}^{\mathrm{F2}} \le F_j, \qquad \chi_{i,j}^{\mathrm{F2}} \ge F_i + F_j - 1$$

Now Equation 6.26 and 6.27 can be used to get a linear relation for the cubic term of variable n, i.e.,

$$n^3 = n^2 \cdot n = \left(\sum_{i,j=1}^{Z} \chi_{i,j}^{\mathrm{F2}} \right) \cdot \left(\sum_{k=1}^{Z} F_k \right) = \sum_{i,j,k=1}^{Z} \chi_{i,j}^{\mathrm{F2}} \cdot F_k = \sum_{i,j,k=1}^{Z} \chi_{i,j,k}^{\mathrm{F3}} \tag{6.28}$$

The value of the binary variable $\chi_{i,j,k}^{\mathrm{F3}}$ is decided by the following three constraints.

$$\chi_{i,j,k}^{\mathrm{F3}} \leq \chi_{i,j}^{\mathrm{F2}}, \qquad \chi_{i,j,k}^{\mathrm{F3}} \leq F_k, \qquad \chi_{i,j,k}^{\mathrm{F3}} \geq \chi_{i,j}^{\mathrm{F2}} + F_k - 1$$

Adopting this strategy any higher order term of variable n can be made linear. Another non-linear term which happens to appear in Equation 6.25 is the product of the continuous variable Y and n. Such a product can be linearized as follows

$$Y \cdot n = Y \cdot \left(\sum_{i=1}^{Z} F_i \right) = \sum_{i=1}^{Z} Y \cdot F_i = \sum_{i=1}^{Z} \chi_i^{\mathrm{YF}} \tag{6.29}$$

here χ_i^{YF} is a continuous variable which substitutes the product of binary variable F_i and continuous variable Y. Taking $\check{Y}$ as the maximum possible value of Y, the following three constraints help settle the value of χ_i^{YF}, i.e.,

$$\chi_i^{\mathrm{YF}} \leq \check{Y} \cdot F_i, \qquad \chi_i^{\mathrm{YF}} \leq Y, \qquad \chi_i^{\mathrm{YF}} \geq Y - \check{Y} \cdot (1 - F_i)$$

In the same way, the product of the continuous variable Y and any higher order term of n can be made linear. For example, using the value of n^2 from Equation (6.27), its product with Y can be written as

$$Y \cdot n^2 = Y \cdot \left(\sum_{i,j=1}^{Z} \chi_{i,j}^{\mathrm{F2}} \right) = \sum_{i,j=1}^{Z} Y \cdot \chi_{i,j}^{\mathrm{F2}} = \sum_{i,j=1}^{Z} \chi_{i,j}^{\mathrm{YF2}} \tag{6.30}$$

where the value of the continuous variable $\chi_{i,j}^{\mathrm{YF2}}$ is determined by three constraints as follows

$$\chi_{i,j}^{\mathrm{YF2}} \leq \check{Y} \cdot \chi_{i,j}^{\mathrm{F2}}, \qquad \chi_{i,j}^{\mathrm{YF2}} \leq Y, \qquad \chi_{i,j}^{\mathrm{YF2}} \geq Y - \check{Y} \cdot (1 - \chi_{i,j}^{\mathrm{F2}})$$

The summation term $\sum T_{s_i}$ in equation (6.25) represents the addition of the T_s parameters of active WLAN users. This can be rewritten as following

$$\sum_{i=1}^{n} T_{S_i} = \sum_{i=1}^{Z} F_i \cdot T_{S_i}$$

Linearizing the product terms of Equation (6.25) in the above described manner, a complete linear relation is obtained, i.e.,

$$\sum_{i=1}^{n} F_i \cdot G = \sum_{j,k,l,m=1}^{Z} (A_{11} + A_{21} \cdot T_{S_j}) \cdot \chi_{j,k,l,m}^{\text{YF4}} + \sum_{j,k,l=1}^{Z} (A_{12} + A_{22} \cdot T_{S_j}) \cdot \chi_{j,k,l}^{\text{YF3}} + \sum_{j,k=1}^{Z} (A_{13} + A_{23} \cdot T_{S_j}) \cdot \chi_{j,k}^{\text{YF2}} + \sum_{j=1}^{Z} (A_{14} + A_{24} \cdot T_{S_j}) \cdot \chi_{j}^{\text{YF}} \tag{6.31}$$

It should be noted that equation (6.31) is valid for $n > 1$. If there is only one active user in the system then no medium contention would take place. In that particular case

$$\widehat{E[D]} = T_s + T_{\text{back-off}} = T_s + \frac{W_{min} - 1}{2} \cdot \sigma = \widetilde{T_s} \tag{6.32}$$

Equation (6.31) and equation (6.32) can be combined by introducing another binary variable H whose value is 1 if there is only one active user (i.e., $n = 1$) and 0 otherwise. The value of the H is determined by following constraints

$$2 - H \cdot \xi \leq \sum_{j=1}^{Z} F_j \quad \text{and} \quad 1 + (1 - H) \cdot \xi \geq \sum_{j=1}^{Z} F_j$$

where ξ represents a large constant real number value, e.g., 10 or higher. The ξ helps establish the logic for linear programming to determine the intended value of the H. Finally the linearized version of equation (6.25) which is valid for $n \geq 1$ is given as below:

$$\sum_{i=1}^{n} F_i \cdot G = \sum_{j,k,l,m=1}^{Z} \left(A_{11} + A_{21} \cdot T_{S_j}\right) \cdot \chi_{j,k,l,m}^{\text{YF4}} + \sum_{j,k,l=1}^{Z} \left(A_{12} + A_{22} \cdot T_{S_j}\right) \cdot \chi_{j,k,l}^{\text{YF3}} + \sum_{j,k=1}^{Z} \left(A_{13} + A_{23} \cdot T_{S_j}\right) \cdot \chi_{j,k}^{\text{YF2}} + \sum_{j=1}^{Z} \left(A_{14} + A_{24} \cdot T_{S_j}\right) \cdot \chi_{j}^{\text{YF}} - \sum_{j=1}^{Z} \chi_{j}^{\text{YFH}} \cdot \left(A_{11} + A_{12} + A_{13} + A_{14} + (A_{21} + A_{22} + A_{23} + A_{24}) \cdot T_{S_j} - \widetilde{T_{S_j}}\right) \tag{6.33}$$

Now after linearizing the WLAN path cost function, the resource allocation problem for uplink communication can be mathematically modeled as shown in Table 6.6. Owning to the fact that this is an extension of the model described in Table 6.4, it inherits all input parameters such as $\alpha, \beta, \phi, \lambda, \Lambda, \Omega$, etc. from the parent model. In addition, it also declares a new parameter G which is the mean size of IP packets belonging to the traffic flows received by users in the WLAN network. The variables of the model include R^{lte} and R^{wlan}, which represent the user data rate carried over the LTE and WLAN access links, respectively. The user throughput in the WLAN network is defined by Y while $E\&F$ are used as auxiliary variables. The model targets to increase the network capacity by maximizing throughput of users over their LTE access links. Furthermore, a suitable set of users is selected which can achieve the maximum possible throughput in WLAN network, i.e., maximize Y.

The model has to work with a number of constraints as shown in the lower part of Table 6.6. Constraints 1–4 are exactly the same as described in Table 6.4 for downlink communication. In Constraint 5, it is elaborated that $R_j^{wlan} = Y$ is valid only for a user j who have been selected as an active WLAN user (i.e., $F_j = 1$) by the model. The 6th Constraint ensures that there is at least one active user in the WLAN access network. The last constraint has been written in its compact form for the sake of brevity; this constraint includes Equation (6.33) and all of its associated constraints which come into being during the linearization of this relation. The constraints related to Equation 6.29 and 6.30 are examples of such constraints.

6.3.2.1 Simulation Scenarios and Results

The performance of the "Linear Programming" approach for uplink is studied with the help of a simulation scenario. The scenario comprises 3 VoIP, 2 Skype video, and 7 FTP uplink users. The other parameter configurations are shared from the scenario of the downlink communication as listed in Table 6.5. Moreover, for the sake of comparison the default "3GPP-HO" as well as the "Time Round-Robin" approaches are considered as from Chapter 5. The resource allocation policies of the two aforementioned approaches have already been discussed. As far as the "Linear Programming" approach is concerned, the DE_n assigns data rates to the users by repeatedly solving the mathematical model presented in Table 6.6.

FTP uplink application performance has been indicated by Figure 6.15 in terms of uplink throughput and file upload time. The figure clearly attribute "Linear Programming" approach as the best among the other competing approaches. By employing the clever strategy for optimized resource allocation, the "Linear Programming" approach manages to enhance user throughput up to 40% compared to the default "3GPP HO" approach. This performance gain is substantially higher than that observed in downlink communication scenarios. The rationale behind this is the medium contention among the users in the WLAN network. These users when transmitting in uplink direction, encounter packet collisions and, therefore,

Table 6.6: Mathematical model for the resource allocation in uplink communication.

Given

U a set of multihomed users

α_j Data rate dependent part of the LTE link cost in [PRB/kbps] for user j, $\forall j \in U$

β_j Data rate independent part of the LTE link cost in [PRB] for user j, $\forall j \in U$

ϕ_j Path cost of WLAN access link in [sec/kbps] for user j, $\forall j \in U$

λ_j Minimum data rate [kbps] demand of a traffic flow destined to user j, $\forall j \in U$

Λ_j Maximum data rate [kbps] allocation for a traffic flow destined to user j, $\forall j \in U$

Ω Number of available PRBs for the LTE access network

G Mean IP packet size for traffic flows of active WLAN users in [bit]

Defined variables

R_j^{lte} Data rate in [kbps] carried over the LTE access link to user j, $R_j^{lte} \geq 0$, $\forall j \in U$

R_j^{wlan} Data rate in [kbps] carried over WLAN access link to user j, $R_j^{wlan} \geq 0$, $\forall j \in U$

Y Average user throughput in [kbps] experienced by each active user
in WLAN network, $Y > 0$

E_j Auxiliary binary variable which hints the use of LTE access link by user j;
its value for a user j is 1 if $R_j^{lte} > 0$ and 0 otherwise, $\forall j \in U$

F_j Auxiliary binary variable which hints the use of WLAN access link by user j;
its value is controlled by the model in optimization of Y, $\forall j \in U$

Maximize

$$\sum_{j \in U} R_j^{lte} + Y$$

Subject to

1. $\sum_{j \in U} \left(\alpha_j \cdot R_j^{lte} + \beta_j \cdot E_j\right) \leq \Omega$

2. $\lambda_j \leq \left(R_j^{lte} + R_j^{wlan}\right) \leq \Lambda_j, \quad \forall j \in U$

3. $\left(R_j^{lte}/\Lambda_j\right) \leq E_j \leq \left(R_j^{lte} \cdot 10^{20}\right), \quad \forall j \in U$

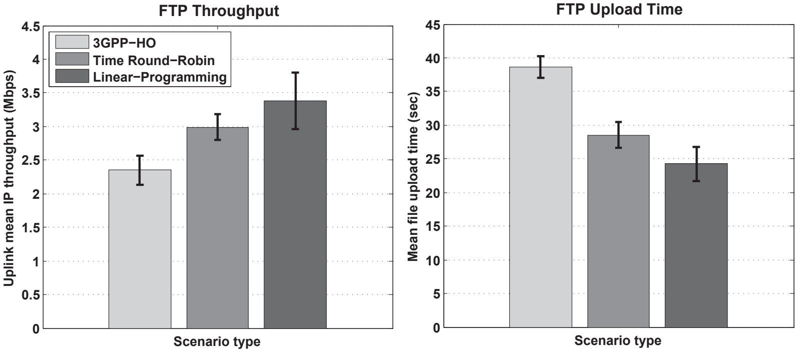
4. $0 \leq R_j^{lte} \leq \Lambda_j, \quad \forall j \in U$

5. $R_j^{wlan} = F_j \cdot Y, \qquad \forall j \in U$

6. $\sum_{j \in U} F_j \geq 1$

7. Equation (6.33) and associated constraints

experience less throughput. On the other hand, in downlink communication, it is mainly the access point which is transmitting and all users act as mere receivers.



(a) (b)

Figure 6.15: FTP uplink performance comparison for “3GPP-HO”, “Time Round-Robin”, and “Linear-Programming” approaches.

Therefore, no medium contention occurs in that case. That is why, a potential performance boost can be realized in the uplink communication scenarios by minimizing the medium contention. This is exactly one of the strategies used in the ‘Time Round-Robin” and the “Linear Programming” approaches. The “Linear Programming” approach even outperforms the “Time Round-Robin” approach by additionally exploiting access network diversity through the use of the optimized resource allocation scheme. Due to this reason, it achieves 15% higher uplink throughput compared to that of “Time Round-Robin”.

Figure 6.16 shows the performance of realtime applications, i.e., VoIP and video. The Boxplots of MOS scores indicate an excellent performance promised by the “Time Round-Robin” and the “Linear Programming” approaches for both application types. The results accomplished by “3GPP-HO” are good, as well. For example, in case of VoIP, all MOS score values stay close to 4.0 which represents an “all users satisfied” state of the service. However, for video applications, several outliers can be observed which indicate the user dissatisfaction about the quality of certain video calls. The cause of this phenomenon is explained by the end-to-end packet delay plot of Figure 6.17. Though most of the time packet delays are confined within 250ms threshold value, there exists a small probability that video packets arrive at the destination later than the aforementioned delay threshold. Such packets are discarded by the de-jitter buffer and hence cause a QoE degradation. These elongated packet delays are mainly caused by a congested WLAN access network due to its FTP users. Both “Time Round-Robin” and “Linear Pro-

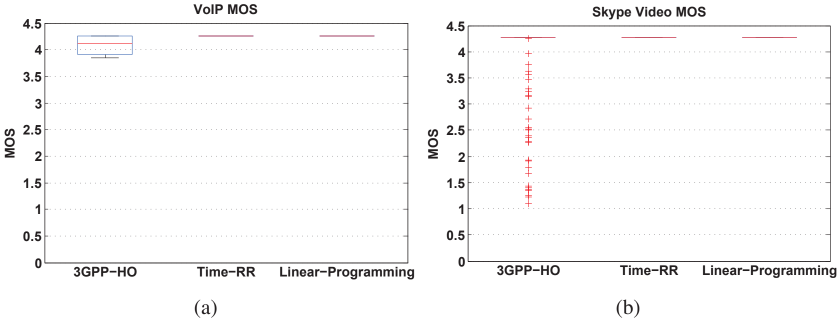


Figure 6.16: Uplink performance comparison of realtime applications for “3GPP-HO”, “Time Round-Robin”, and “Linear Programming” approaches.

gramming” approaches circumvent these stretched delays by precisely regulating the traffic load in the access networks. This makes packet delays stay well below 100ms excluding a few outliers.

When comparing the performance of realtime applications in uplink and downlink communication, it can be observed that “3GPP-HO” delivers much improved service in case of uplink communication. The reason of this behavior can be recalled from the discussion of results in Section 6.3.1.1. That is, in downlink, a single MAC queue at the WLAN access point is shared by both VoIP/video and FTP traffic. Owing to the fact that it is the only bottleneck point in the network, large buffer occupancy is created by the TCP protocol at this MAC queue. This leads to large queuing delays for the packets of realtime and non-realtime applications. In contrast to this, in uplink communication, although the users have their individual WLAN MAC queues in their devices, but in order to successfully transmit a packet they have to contend for the medium access. Therefore, if the number of active users are not excessively large, a packet can be transmitted within a dejitter threshold value of 250ms.

6.4 Heuristic Algorithms

The solution of the resource allocation problem obtained through mathematical modeling using linear programming provides an upper limit on achievable network capacity. A common practice in this regard is to consider that maximum achievable performance as a target and then devise some heuristic algorithms which try to attain a similar performance. The rationale behind this practice is the involved

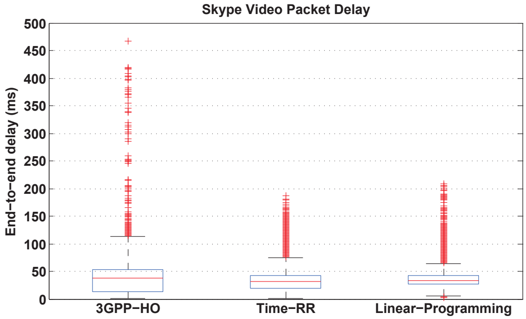


Figure 6.17: Packet delay comparison of the Skype video application for “3GPP-HO”, “Time Round-Robin”, and “Linear Programming” approaches.

high complexity of linear programming problem. The high complexity requires substantial computing power and time to solve these problems. This make the use of linear programming unsuitable for realtime optimization tasks in most of the cases. In this section, first of all the complexity of the proposed linear programming approaches is discussed and then two heuristic algorithms are developed for downlink and uplink communication scenarios. The effectiveness of the suggested algorithms is also evaluated by comparing with the corresponding linear programming approaches.

A customary way of analyzing the complexity of a linear programming problem is through the number of involved variables and constraints. Figure 6.18 depicts the complexity of the linear programming problem for downlink communication. The two curves indicate that the number of variables and constraints increase linearly with the number of active users in the network. Moreover, even for a large number of users (e.g., 100) the linear programming problem seems to have fairly small computational complexity. This is because only few hundreds of variable and constraints are involved at that user count. This fact is also verified by examining the wall-clock time required to solve these linear programming problems on a laboratory server computer [1]. The machine was able to solve any of such problems in less than 10ms of wall-clock time. The observation is based on an analysis involving 20,000 random problems with active number of users varying from 3 to 100.

Figure 6.19(a) shows the complexity of the linear programming problem used to optimally allocate resources in uplink communication. The figure indicates that

[1] Microsoft Windows Server 2008 R2 Enterprise 64bit, Intel©Xeon CPU @ 2.67GHz, 48GB RAM

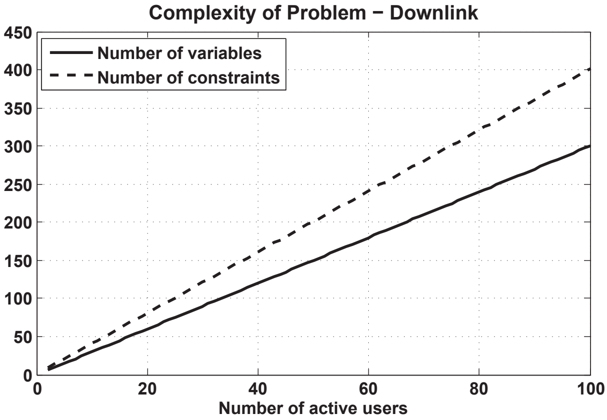


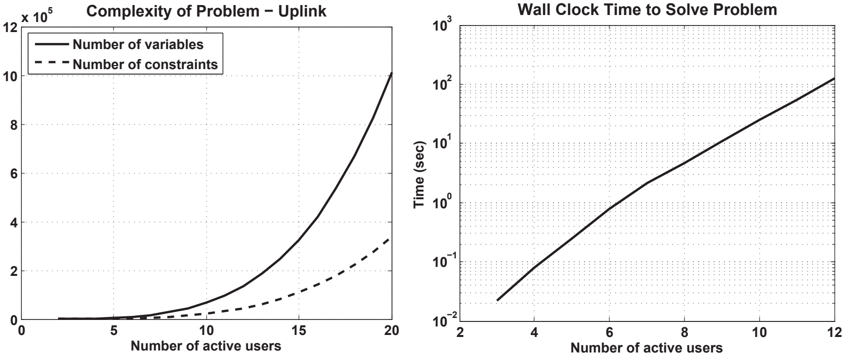
Figure 6.18: The complexity of linear programming problem for downlink communication described in Table 6.4.

the variables and constraints count grows exponentially with the number of multihomed users. This makes the problem extremely hard to solve even just for 20 users which involves approximately 1 million variables and 400,000 constraints. Accordingly, the wall-clock time to solve such problems also grows exponentially increasing number of users as indicated in Figure 6.19(b). Therefore, in this case, there is a particular need for an efficient heuristic algorithm to efficiently solve the optimum resource allocation problems in realtime. Over the next two subsections, the heuristic algorithms are developed for both downlink and uplink communication scenarios.

6.4.1 Downlink Communication

Table 6.7: An example problem of resource allocation in downlink communication.

User	Normalized network path cost per kbps		Data rate demand [kbps]	
	WLAN	LTE	Minimum	Maximum
UE1	6×10^{-5}	4×10^{-5}	10^3	10^3
UE2	9×10^{-5}	5×10^{-5}	10^3	10^3



(a) (b)

Figure 6.19: The complexity of the linear programming problem for uplink communication described in Table 6.6. The figure on right hand side is a semi-log graph with y-axis plotted on a logarithmic scale.

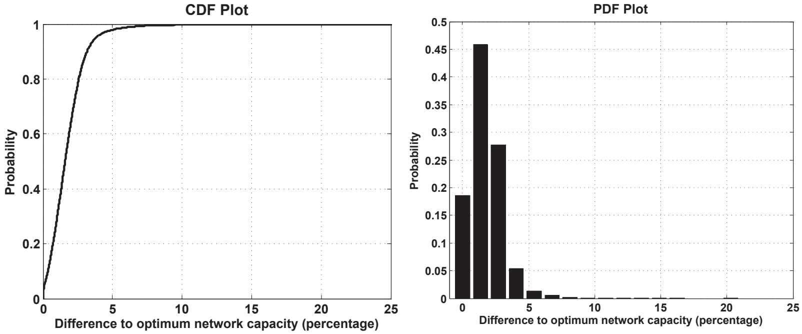
Before the development of the heuristic algorithm, an understanding of the resource allocation problem is developed with the help of a simple example presented in Table 6.7. In this example, there are two multihomed users who require a fixed data rate of 1Mbps to run a realtime service, e.g., video streaming. The normalized network path cost on each access link of the user is also mentioned in the table. The normalized cost represents the fraction of total access network resources to offer a user with 1 Kbps data rate over that access network. The normalized costs help directly compare the resource consumption of WLAN and LTE access networks for a given amount of data rate, e.g., it can be seen that UE1 has less path cost for the LTE access link compared to its WLAN access link.

The most suitable strategy to allocate resources in such a situation is through the greedy approach. This implies that users should be served over that particular access link which costs less network resources. It can be noticed from the table that both users have less normalized cost for LTE access links compared to that of WLAN access links. Therefore, according to the greedy strategy both users should be served over their LTE access link. Serving them with their minimum data rate over the LTE access network will consume 4×10^{-2} and 5×10^{-2} fraction of resources, respectively. In other words, it will require a total of 9% of the available LTE resources.

This strategy of the greedy approach is the main driver behind the heuristic algorithm developed for resource allocation in downlink communication as depicted in Figure 6.21. The algorithm takes the network path costs and user data rate demands as inputs. It traverses through the list of multihomed users and serve them with the minimum data demands over their less expensive access links. If it happens that the available network resources are already assigned, then the rest of the users are served over the other access network. In case, the network resources of both access networks are consumed without satisfying the minimum data rate demands of all users, the algorithm returns an error message. The error message indicates that the provided problem is infeasible and there is no solution to the problem.

After fulfilling the minimum data rate demands of all users, the left over network resources should be assigned to the users whose maximum data rate demand is greater than their minimum data rate demand. Typically, they are the FTP/HTTP users. Though the same greedy approach can also be employed here once again, it has to be slightly modified. This is because satisfying the maximum data rate demand of 'each' user is not compulsory. Therefore, only those users should be served who can achieve greater data rates with the available network resources. For this purpose, a list is prepared where the network path cost of each user for 'both' of its access links is added. The size of this list is twice the number of users. Sorting this list in the ascending order, users are served in the same order in which their access link costs appear in the list. This procedure of serving users up to their maximum data rate demand is performed in subprocess (A) in Figure 6.21. A flow chart of subprocess (A) has been shown in Figure 6.22. At the end, the heuristic algorithm returns the user data rate assignments over each of their access links.

In order to evaluate the performance of the heuristic algorithm described in Figure 6.21 and 6.22, its results are compared with that of the "Linear Programming" approach. The evaluation is made more comprehensive and thorough by solving 20,000 random problems of resource allocation using both the heuristic and "Linear Programming" approaches. The problems are generated automatically with the help of a script which considers a large range of active users from 3 to 100. A probability of 50% is used to determine if a user in a random problem should be using the realtime service (i.e., minimum and maximum data rate demands are same) or the non-realtime service (i.e., maximum data rate demand is greater than minimum data rate demand). Figure 6.20 summarizes the outcome of this evaluation process. It shows a CDF and PDF curves of values representing how large total network capacity is achieved using "Linear Programming" approach compared to that obtained by the heuristic algorithm in random resource allocation problems. The CDF curve depicts that in 90% of the problems, the heuristic approach achieved a network capacity which was at most 3% less than the optimum achievable capacity computed by the "Linear Programming" approach. Considering the simple complexity of the heuristic approach it is a great performance.



(a) (b)

Figure 6.20: The performance of the proposed heuristic algorithm for downlink communication. The CDF and PDF curves show the difference of the achieved network capacity using the heuristic algorithm compared to the optimum value obtained using “Linear Programming” approach.

A question can be raised at this point; why the simple greedy approach cannot achieve the same performance as shown by “Linear Programming” approach. This can be explained with the help of an example shown in Table 6.8. It is a slightly modified version of the example presented in Table 6.7 where the maximum data rate demands of users have been raised to 23.75Mbps. The resource allocation problem in this example is solved using the developed heuristic algorithm as follows. First of all, both users are served with their minimum data rate demands (i.e., 1Mbps) over LTE access network due to minimum involved resource consumption. This costs 9% of LTE resources. As there are still 91% of LTE and 100% of WLAN access network resources available, a sorted list of network path costs is prepared to utilize the remaining resources. The cost 4×10^{-5} of LTE access link from UE1 comes at the top, therefore 91% of LTE access network resources are allocated to UE1 which translates to a data rate of 22.75Mbps. This way UE1 is assigned with a total data rate of 23.75Mbps considering also 1Mbps data rate allocation in the first step. The next lowest access link cost is of UE2 for its LTE access link (i.e., 5×10^{-5}), however, there are no resources left on the LTE access network. Therefore, no action is taken for UE2 this time. The next lowest cost would be 6×10^{-5} of UE1 for its WLAN access link, but this user has already been served up to its maximum data rate demand. Hence no additional

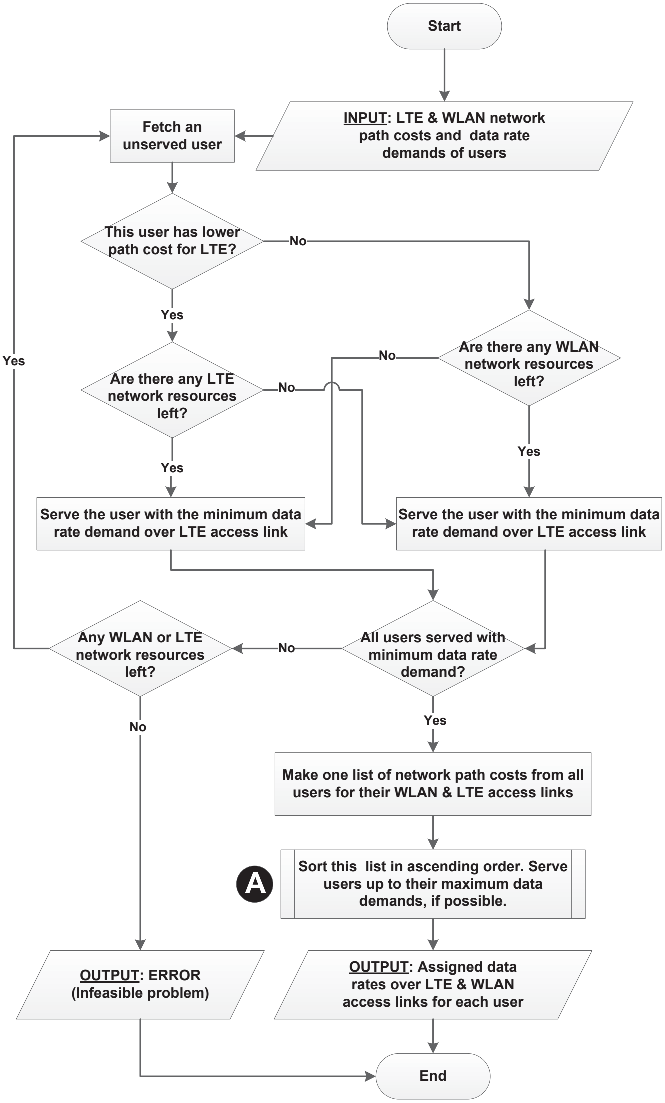

Start
INPUT: LTE & WLAN network path costs and data rate demands of users
Fetch an unserved user
This user has lower path cost for LTE?
No
Yes
Are there any WLAN network resources left?
No
Are there any LTE network resources left?
No
Yes
Yes
Yes
Serve the user with the minimum data rate demand over LTE access link
Serve the user with the minimum data rate demand over LTE access link
Any WLAN or LTE network resources left?
No
All users served with minimum data rate demand?
Yes
No
Make one list of network path costs from all users for their WLAN & LTE access links
A
Sort this list in ascending order. Serve users up to their maximum data demands, if possible.
OUTPUT: ERROR (Infeasible problem)
OUTPUT: Assigned data rates over LTE & WLAN access links for each user
End

Figure 6.21: Flow chart of the heuristic algorithm to solve the resource allocation problem in downlink communication.

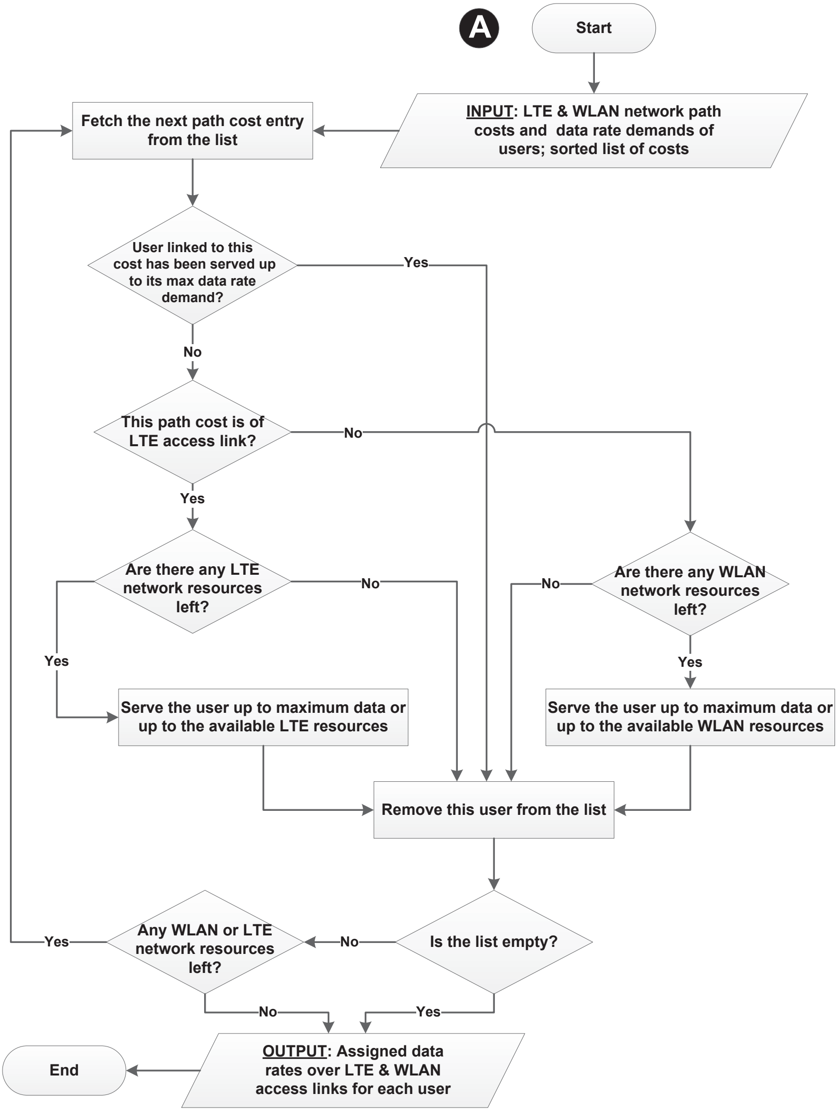

A
Start
INPUT: LTE & WLAN network path costs and data rate demands of users; sorted list of costs
Fetch the next path cost entry from the list
User linked to this cost has been served up to its max data rate demand?
Yes
No
This path cost is of LTE access link?
No
Yes
Are there any LTE network resources left?
No
Yes
Are there any WLAN network resources left?
No
Yes
Serve the user up to maximum data or up to the available LTE resources
Serve the user up to maximum data or up to the available WLAN resources
Remove this user from the list
Is the list empty?
No
Yes
Any WLAN or LTE network resources left?
Yes
No
OUTPUT: Assigned data rates over LTE & WLAN access links for each user
End

Figure 6.22: Flow chart of the subprocess (A) in Figure 6.21.

resource can be assigned to UE1. The last entry in the sorted list of cost would be 9×10^{-5} of UE2 over its WLAN access link. 100% of the WLAN access network resources are assigned to this user which amount to a data rate of 11.1Mbps. This way, UE2 gets a total data rate allocation of 12.2Mbps considering also 1Mbps data rate allocation in the first step. Hence, the total network capacity amounts to 22.75+12.2=34.95Mbps, in this case.

Table 6.8: Another example problem of resource allocation in downlink communication.

User	**Normalized network path cost per kbps**		**Data rate demand [kbps]**	
	WLAN	**LTE**	**Minimum**	**Maximum**
UE1	6×10^{-5}	4×10^{-5}	10^3	23.75×10^3
UE2	9×10^{-5}	5×10^{-5}	10^3	23.75×10^3

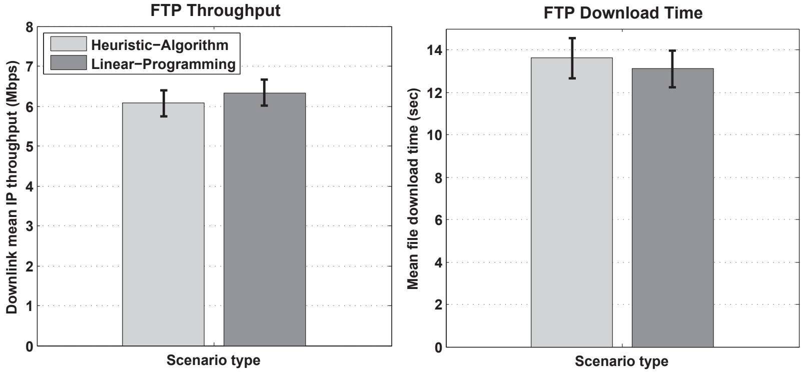
Solving the same problem using the "Linear Programming" approach serves UE1 completely over WLAN access network despite the fact that it has lower cost for LTE access link. This is because assigning all LTE resources to UE1 means that UE2 will have to be served over its WLAN access link which has the highest path cost. This would be a bad move which could decrease the over spectral efficiency of the network. Therefore, the "Linear Programming" approach takes an intelligent decision of serving UE1 over WLAN access network and keep LTE resources for UE2. Following this strategy, UE1 is served completely over WLAN access network with data rate of 16.67Mbps and UE1 over the LTE access network with data rate of 20Mbps. This way, total network capacity amounts to 36.67Mbps which is 4.9% higher than that attained by using the heuristic approach.

A sophisticated heuristic algorithm which mimics the "Linear Programming" approach in conceiving the effects of resource allocation of a user on the achievable spectral efficiency of the other users will be overly complex. This is because as the user count increases, each resource allocation will have to get feedback from many of the users in a recursive way. Based on this feedback, the algorithm would have to decide whether performing this resource allocation could degrade the achievable spectral efficiency of other users. Above all, devising such an advanced scheme would not offer a significant performance gain and would be against the idea of developing a simple alternative approach.

6.4.1.1 Simulation Scenarios and Results

This sections puts the developed heuristic approach to the test and evaluate its performance in a simulation scenario. For this purpose, the simulation scenario discussed in Section 6.3.1.1 is reused here so that the results of the heuristic approach can be compared with that of the “Linear Programming” approach.

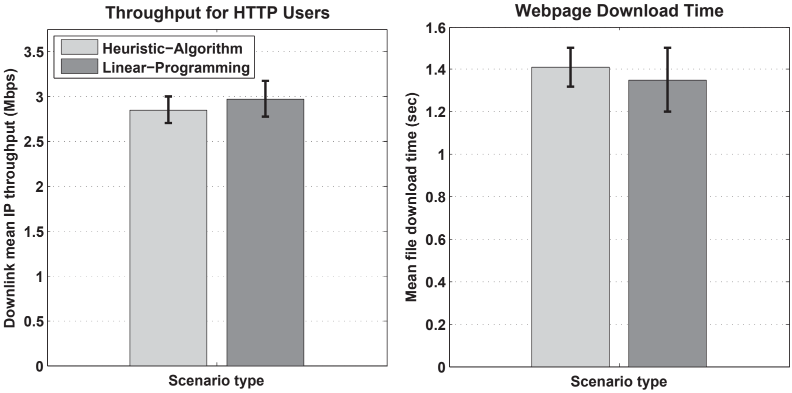
Figure 6.23 compares the performance of the FTP downlink application for two competing approaches. It is expected that the heuristic approach might not be able to deliver a performance matching to that of “Linear Programming”. The FTP downlink throughput as well as file download time verify this expected behavior. However, the performs degradation is not significant. A comparison of numerical values reveals that the loss of performance is as low as 4%. A very similar observation is also made for HTTP application performance. In this case users encountered just 3% degradation in their QoE for webpage downloads. Moreover, the absolute values of increase in download times are in the range of milliseconds. Such a slight increase in download time remains unnoticed for human users.



(a) (b)

Figure 6.23: FTP downlink performance comparison for “Heuristic Algorithm” and “Linear Programming” approaches.

Though non-realtime applications suffer slightly due to the use of the approach based on heuristic algorithm, the performance of realtime applications essentially remains unaltered. The reason behind this phenomenon has already been discussed. That is, the foremost target of the heuristic approach is to satisfy the minimum data rate demands of all users. Owing to the fact that realtime applications require a fixed amount data rate, their minimum data rate demands are al-



(a) (b)

Figure 6.24: HTTP downlink performance comparison for "'Heuristic Algorithm" and "Linear Programming" approaches.

ways fulfilled. Only after allocating the minimum required data rates to all users, the heuristic approach distributes the left-over resources among the non-realtime users. Therefore, if the resources are not utilized optimally, there will be fewer resources left to serve TCP users with the data rates surplus to their minimum data rate demands.

The simulation results of realtime applications (i.e., VoIP and video) has not been shown here in order to avoid unnecessary repetitions.

6.4.2 Uplink Communication

In uplink communication, users in the WLAN access network have to contend for the medium access when carrying out any transmission. This implies that the expected data rate of users over the WLAN access link would be degraded if other users also start transmitting at the same time. The extent, to which the data rate degrades, depends upon the PHY data rate of the involved users. In such situations the simple greedy scheme will not be of much use because the resources required to serve a user will change in response to the selection of other users to be served over the WLAN. A work-around to this problem has been found through the study of resource allocation solutions obtained using the "Linear Programming" approach. The analysis indicates that in the majority of the problems, at most three users are served over the WLAN access network and the rest are served over the LTE access network. This is because, serving large number of users over the WLAN degrades the network capacity due to the involved contention among them. However, the question still remains which particular users should be served over the WLAN access network. One option could be to select those users who have the better channel conditions compared to the other users. The findings of the analysis reveal that the aforementioned criterion is not applicable in all situations. The reason of which is very similar to the one explained during the discussion of the example presented by Table 6.8.

This issue is resolved by searching for the most appropriate group of users which, on being served over WLAN access network, can maximize the overall capacity of both networks. The search is executed by trying out all possible groups or combinations of up to three users. For each such combination of users, the total capacity of both networks is computed where these users are served over WLAN and the rest are served over LTE following the greedy approach. Now the user combination which promises the highest total capacity of networks is selected to be served over their WLAN access links. The number of user combinations from which the best choice has to be searched, depends on the number of multihomed user as described below

$$\begin{aligned}\text{Number of combinations} &= Z + \frac{Z!}{2!\cdot(Z-2)!} + \frac{Z!}{3!\cdot(Z-3)!}\\ &= Z + \frac{Z\cdot(Z-1)}{2} + \frac{Z\cdot(Z-1)\cdot(Z-2)}{6}\end{aligned} \tag{6.34}$$

where Z is the number of active users in a WLAN access network.

In real world scenarios, the number of users associated with a WLAN access point is typical assumed close to 20. Solving a resource allocation problem for the 20 users requires heuristic algorithm to search among 1,350 possible user combinations. Even extending this count up to 50 users would make the search process to look for one best user combination out of 20,875 possibilities. The machine used in the analysis presented in Figure 6.19(b), was able to execute such a search process in less than 10ms.

As described earlier, the analysis of solutions to the resource allocation problems reveals that mostly up to three users are served over the WLAN access network. The question can be asked; what happens in the rest of cases? This can be explained with the help of an example presented in Table 6.9. Here the first four users have very low path cost for their WLAN access links while the fifth user is economical if served over its LTE access link. Owing to the fact that the WLAN access network is capable of serving the first four users up to their maximum data rate demands, the “Linear Programming” approach will instruct them to transmit

Table 6.9: An example problem of resource allocation in uplink communication.

User	**Normalized network path cost per kbps**		**Data rate demand [kbps]**	
	WLAN	**LTE**	**Minimum**	**Maximum**
UE1	5×10^{-5}	40×10^{-5}	10^3	3×10^3
UE2	5×10^{-5}	40×10^{-5}	10^3	3×10^3
UE3	5×10^{-5}	40×10^{-5}	10^3	3×10^3
UE4	5×10^{-5}	40×10^{-5}	10^3	3×10^3
UE5	50×10^{-5}	4×10^{-5}	10^3	20×10^3

over their WLAN access links. The fifth user is then obviously served over LTE access network because serving this user over the WLAN access network will drag down the whole WLAN access network capacity.

In the heuristic algorithm, by limiting the user combination size to 3 would provide a suboptimal solution in the above mentioned kind of problems. This drawback can be circumvented using a ‘post-include-in’ strategy. In this strategy, after getting the best user combination, it is tried to serve more users over the WLAN access network without harming the existing users. For this purpose, an additional user is added to the best user combination and the total capacity of WLAN and LTE access networks is computed. If the resulted network capacity increases, that user is marked to be served over WLAN access network altogether with the best combination users. Otherwise, the same test is run on the other users, one by one.

The above mentioned strategies of the resource allocation process have been illustrated in the form of a flow chart in Figure 6.25. The process takes network path costs and data rate demands of users as inputs. The first step searches for the best user combination with the help of subprocess (B). Afterwards, the post-include-in strategy is applied to improve network capacity using subprocess (C). Finally, the user data rates are determined based on the outcome of subprocess (C).

The subprocess (B) has been outlined in Figure 6.26 which mainly assigns LTE access network resources to the users following the greedy approach. That is, users are first assigned the minimum data rate over the LTE access network. Then the users are sorted according to their LTE path cost and served up to their maximum data rate until all LTE access network resources are allocated. The subprocess (C) has been described in Figure 6.27 which tries to improve the network capacity by serving more users over the WLAN access network in addition to the users of the best combination.

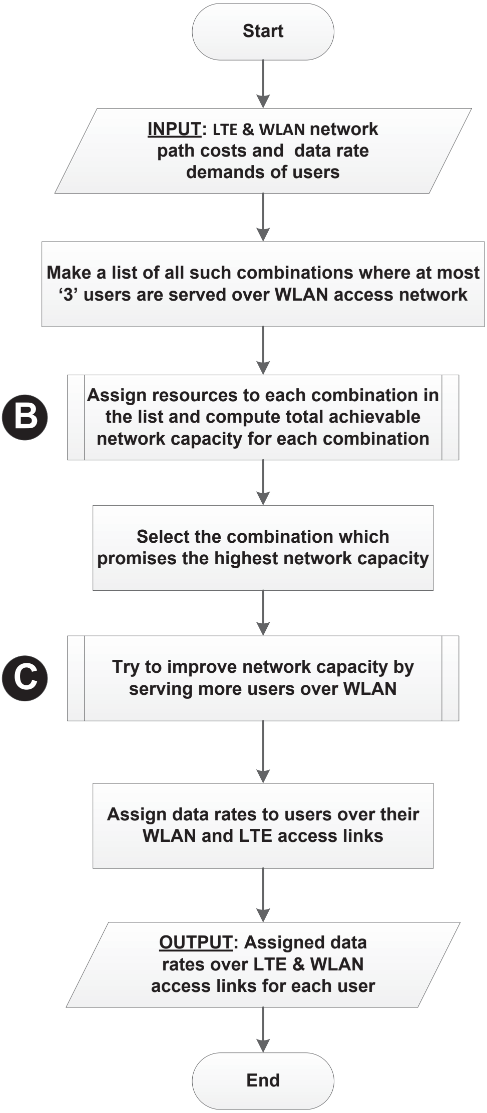

Start
INPUT: LTE & WLAN network path costs and data rate demands of users
Make a list of all such combinations where at most '3' users are served over WLAN access network
B
Assign resources to each combination in the list and compute total achievable network capacity for each combination
Select the combination which promises the highest network capacity
C
Try to improve network capacity by serving more users over WLAN
Assign data rates to users over their WLAN and LTE access links
OUTPUT: Assigned data rates over LTE & WLAN access links for each user
End

Figure 6.25: Flow chart of the heuristic algorithm to solve the resource allocation problem in uplink communication.

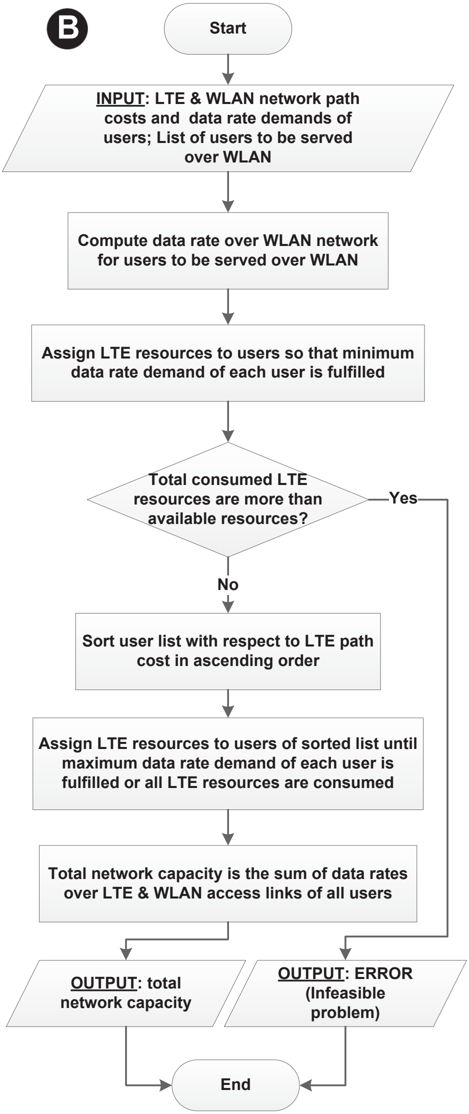

B
Start
INPUT: LTE & WLAN network path costs and data rate demands of users; List of users to be served over WLAN
Compute data rate over WLAN network for users to be served over WLAN
Assign LTE resources to users so that minimum data rate demand of each user is fulfilled
Total consumed LTE resources are more than available resources?
Yes
No
Sort user list with respect to LTE path cost in ascending order
Assign LTE resources to users of sorted list until maximum data rate demand of each user is fulfilled or all LTE resources are consumed
Total network capacity is the sum of data rates over LTE & WLAN access links of all users
OUTPUT: total network capacity
OUTPUT: ERROR (Infeasible problem)
End

Figure 6.26: Flow chart of the subprocess (B) in Figure 6.25.

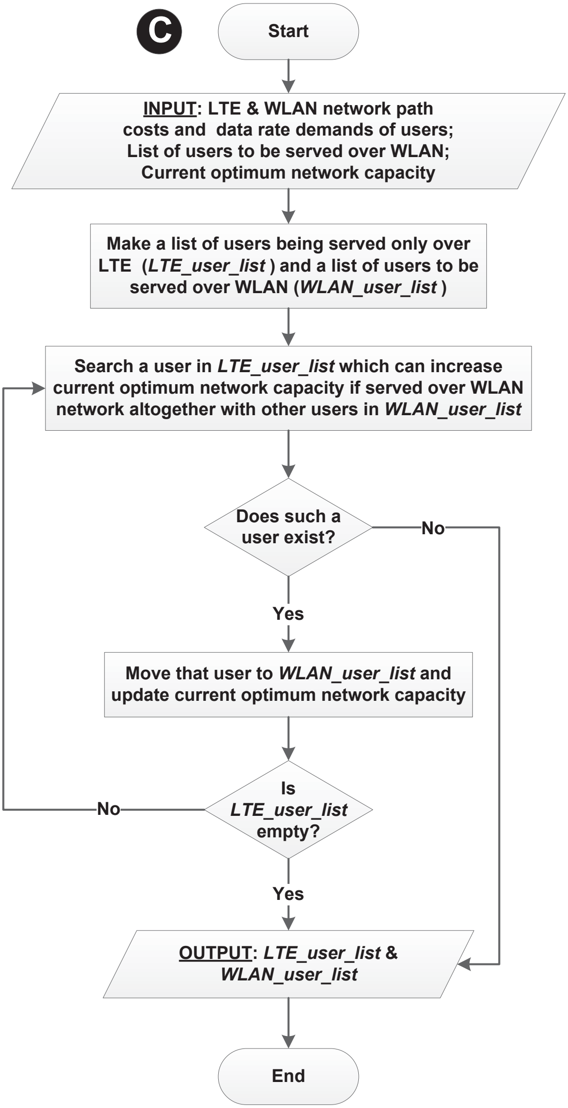

C
Start
INPUT: LTE & WLAN network path costs and data rate demands of users; List of users to be served over WLAN; Current optimum network capacity
Make a list of users being served only over LTE (LTE_user_list) and a list of users to be served over WLAN (WLAN_user_list)
Search a user in LTE_user_list which can increase current optimum network capacity if served over WLAN network altogether with other users in WLAN_user_list
Does such a user exist?
No
Yes
Move that user to WLAN_user_list and update current optimum network capacity
Is LTE_user_list empty?
No
Yes
OUTPUT: LTE_user_list & WLAN_user_list
End

Figure 6.27: Flow chart of the subprocess (C) in Figure 6.25.

The performance of the proposed heuristic algorithm is compared with that of the “Linear Programming” approach using a batch of 2,000 random tests. In these tests the number of users is selected using a uniform distribution in the range from 3 to 20 users. Furthermore, there is an equal chance for a user to select between realtime and non-realtime services. Out of all of these resource allocation problems, only 4 such cases were identified where the proposed heuristic algorithm could not match the network capacity computed by the “Linear Programming” approach. This way, the heuristic algorithm offered near optimal network capacity in 2,000 resource allocation problems with the probability of 99.8%. In order to further verify this claim the simulation scenario presented in Section 6.3.2.1 is rerun using the proposed heuristic algorithm. The obtained simulation results are then compared with those where the “Linear Programming” approach has been used for resource allocation. The two sets of results appear to be essentially the same and, therefore, have been omitted to avoid repetitions. This provides confidence to the claims that the developed heuristic algorithms exhibit far less complexity compared to “Linear Programming” approaches while delivering a matched performance in computation of optimum network capacity.

This chapter developed mathematical relations between network resources and achievable user data rate for LTE and WLAN access networks. These mathematical relations are then used to model the network resource allocation problem using linear programming both for uplink and downlink communication scenarios. The simulation results obtained by employing the “Linear Programming” approach are seen to excel over the other approaches discussed in Chapter 5. The complexity of the “Linear Programming” approach is also evaluated in terms of required computational power and time. As alternatives to the “Linear Programming” approaches, heuristic based algorithms are devised which have less computational complexity. Moreover, the performance of the proposed heuristic approaches in solving resource allocation problems is proved to be very close to that of the “Linear Programming” approaches.

7 Conclusions and Outlook

The main focus of this thesis work is to enhance user QoE as well as improve network capacity in existing and future wireless access networks. For this purpose, several optimizations are suggested in the 3GPP standards for Long Term Evolution (LTE), with special focus on the radio network. In addition, the work also provides a futuristic look at heterogeneous networks where non-3GPP networks (e.g., WLAN, WiMAX) are integrated into LTE networks. Such integration provides not only the means of traffic offloading but also paves the way to exploit a new dimension of multiuser diversity. As a result, the proposed heterogeneous networks are capable of living up to the demands of the mass-market by achieving increased spectral efficiency and improved services at a lower cost with better user QoE.

The implementation of a simulation model with the necessary details is an important and challenging task in the development and performance evaluation of communication networks. Thus already developed basic simulation models of LTE and WLAN were used in this work to build an integrated heterogeneous network simulation model. The integration of two network types according to 3GPP standards required a number of extensions and modifications in the E-UTRAN nodes (UEs, eNodeBs, PDN-GW, S-GW). Moreover, the realization of user multihoming also required an implementation of IETF specified extensions for the Mobile IPv6 protocol of OPNET. Another implementation task related to the simulator was the development of user QoE evaluation mechanisms for VoIP and video services. In addition, the developed heterogeneous network simulator also implemented the flow management system architecture which has been used to manage network bandwidth resources for multihomed users.

This work proposes valuable enhancements to the LTE air interface scheduler. For example, coordinated radio interface scheduling performs effective congestion avoidance for the LTE core network. This improves system stability and overall system performance in a number of ways, e.g., it saves UE battery power which would otherwise be consumed to retransmit the packets dropped in congested links. It also extends the network coverage and reduces the radio interference in the cell. Another novel LTE air interface scheduling algorithm proposed in this work dynamically adapts to network load conditions. During the time when congestion happens at the LTE radio interface, the algorithm enhances spectral efficiency at the expense of throughput fairness among the users. During the other times, it offers user throughput fairness and extended cell coverage. Finally, the performance evaluation study of different packet queue management schemes for the LTE air interface scheduler is also carried out in this work. The analysis of simulation results reveals that by employing the proposed schemes, not only the end user QoE is enhanced but also inter-site handover completion time is reduced.

The network bandwidth resource management of integrated heterogeneous networks is a cumbersome task for operators. This work introduces a comprehensive overlay architecture for resource management which complements 3GPP compliant heterogeneous network architecture in achieving enhanced user QoE and spectral efficiency. In addition, this work also presents several novel approaches for dynamic estimations of user access link capacity which is a prerequisite for an efficient network resource management. The effectiveness of the proposed architecture and the performance of the developed mechanisms for user link capacity estimation is studied using the results of various simulation scenarios. The results clearly indicate that the use of the proposed resource management schemes substantially enhances the user QoE for both realtime and non-realtime services in an environment of heterogeneous networks. This proves the superiority of heterogeneous networks with intelligent resource management mechanisms over the default 3GPP standardized networks.

In order to explore the limits of the achievable performance gain offered by the intelligent resource management in heterogeneous networks, mathematical optimization techniques are employed. For this purpose, the resource allocation problem is formulated using ‘Linear Programming’. The system performance of this approach excels over the other approaches as indicated by the simulation results obtained. The analytical study of the problem also provides an upper bound on achievable system performance which serves as a target for the designs of new resource management schemes. Inspired by the system performance achieved by mathematical optimization techniques, heuristic based algorithms are also devised in this work. These algorithms not only exhibit less computational complexity but also accomplish a performance gain close to that attained by mathematical optimization techniques. This make them feasible for use in real world network equipment.

The concepts, mechanisms, and system architecture for resource management presented in this work serve as a basis for further research in the area of user multihoming and heterogeneous networks. The current work has focused only on radio interfaces of LTE and WLAN. A natural extension would be to perform the resource management of the transport / core network and radio access networks simultaneously. This will provide the resource management scheme with an overall picture of the access network so that performance of both networks is optimized. For example, in such a scenario it will be possible to determine if the transport network can support the user data rates which are being allocated at the radio interfaces during the resource management process. Similarly, it will be possible to determine the suitable transport links for certain application types based on the dynamic QoS characteristics of these transport links.

Though the simulation results provided the proof of concept for integrated networks of LTE and WLAN, the presented concepts should be equally valid for other access technologies like HSPA, WiMAX etc. However, such a validation is another future work item. Moreover, this work focuses on network resource management controlled by operators, while a quantization of achievable benefits when users manage their own bandwidth resources, remains an open work item. Some work has already been initiated in this direction, e.g., [X. 12], [X. 13].

Appendix

A User Satisfaction Models

The rapid advancements in mobile communication devices like smart phones, tablets, PDAs etc. along with the evolutionary network technologies have opened doors for users to access a variety of multimedia services instead of relying only on voice communication. A major share of bandwidth resources from wireless access networks is being used to offer data services. That is why mobile data services are rapidly becoming an essential component of mobile operators' business strategies. It is expected that this trend will gain further pace in near future with the availability of new services and convergence of various access technologies. Owing to the fact that the requirements of new data services are continuously increasing, the growth of these services has posed big challenges in managing their performance with the constraint of scares wireless network resources.

In today's all IP networks, introducing some Quality of Service (QoS) improvement procedures may not necessarily translate to user satisfaction in the same order. For this purpose, another term, called Quality of Experience (QoE), is used which quantifies user satisfaction level from a service. This shifts the focus of service quality evaluation solely based on technical parameters to more subjective evaluation criteria. This places user QoE on a higher level than technical parameters when categorizing them with respect to their importance for network selection decisions in an environment of heterogeneous wireless networks. This, in turn, dictates that it is imperative for operators to estimate the user satisfaction or QoE for their services. This explains the need for a 'user satisfaction model' which can predict user QoE based on the expected QoS parameters leading to an efficient network resource management. This chapter presents the related work in user satisfaction modeling for various realtime and non-realtime applications. In addition, detailed discussions are made regarding the operation of the most accepted user satisfaction models in the research community.

A.1 Background

This section provides an overview on the relationship of QoS and QoE as well as state-of-the-art work in this area.

A.1.1 Quality of Service (QoS)

QoS describes the network's ability to provide guaranteed service in achieving predictable results. In order to provide end-to-end QoS all those functions and mechanisms in the network that ensure the provisioning of the negotiated service quality must play their role.

Network performance indicators within the scope of QoS include throughput, packet delay and jitter, packet loss etc. QoS guarantees can be provided either by enforcing performance measures, e.g., traffic prioritization, QoS aware scheduling etc. or by doing resource over-dimensioning.

A.1.2 Quality of Experience (QoE)

The Quality of Experience term is used to describe end user's perception of performance of a delivered service. There are several formal definitions of QoE found in literature. For example,

- *"The characteristics of the sensations, perceptions, and opinions of people as they interact with their environments. These characteristics can be pleasing and enjoyable, or displeasing and frustrating."* [SJB+04]
- *"Quality of Experience is the overall performance of a system from the point of view of the users. QoE is a measure of an end-to-end performance levels at the user perspective and an indicator of how well this system meets the user needs."* [Goo05]
- *"The user's perceived experience of what is being presented by the Application Layer, where the application layer acts as a user interface front-end that presents the overall result of the individual Quality of Services."* [SW03]

QoE is usually expressed on a scale of Mean Opinion Score (MOS) [IT98]. MOS value can be measured either by conducting subjective tests or using the mathematical models developed to predict user satisfaction based on number of parameters. In a broader sense, other than network QoS parameters, QoE is also affected by factors such as cost, reliability, efficiency, privacy, security, interface, user-friendliness and user confidence.

A.1.3 QoE versus QoS

The main difference between QoS and QoE is the reference perspective; QoS defines the network perspective of performance while QoE defines the user perspective of service performance. Though the perspective is different, QoE and QoS are so interdependent that no discussion on QoE can be concluded without referring to underlying QoS. In fact, the only way through which a network operator can offer the best QoE to the users in a cost-effective and efficient way is the proper management of QoS in all steps from network planning to implementation and optimization. In other words, the ultimate goal of achieving maximum user satisfaction (QoE) can effectively be accomplished by using the building blocks of QoS [SLC06].

From the above discussion one may get the impression that a better network QoS always results in an improved user QoE. However, this argument does not hold in all circumstances. For example, achieving high throughput and low packet loss using QoS enforcements in one part of the network might not help to satisfy an end user, if there is a severe bottleneck in another part of the network. This implies that QoS is essentially a bottom-up process which consists of a concatenation of point-to-point performance differentiation mechanisms with little focus on end user perception. In contrary to this, QoE is a top-down approach where end-user is the ultimate beneficiary of QoS. Therefore the implementation of QoS in a network can help attain better QoE if the perspective is end user and all service performance levels required for higher user satisfaction are assured.

The goal of delivering high QoE demands a comprehensive understanding of the factors which contribute to end user's perception of provided service. These factors encompass both technical and non-technical aspects of the service (see Figure. A.1). The technical factors are mainly covered by the end-to-end QoS and have been discussed intensively in research literature. However, non-technical factors which are often ignored in QoE estimations also bear an equal importance. Therefore, an accurate model of user QoE must take into consideration the effects of both factor types on the user satisfaction. That is why the user satisfaction model presented in this chapter conforms to this requirement by considering technical as well as non-technical factors.

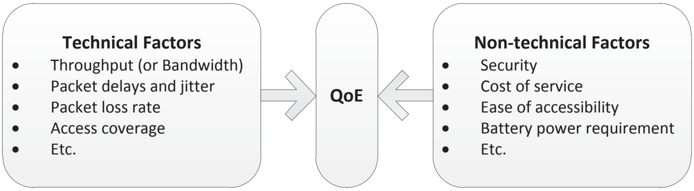


Figure A.1: Factors affecting end user QoE

A.1.4 Related Work

This section gives an overview of the research work found in literature with the focus on computation of user QoE for various types of applications using analytical models. The literature survey shows that speech quality assessment has been of particular interest to many researchers. This yielded a large number of signal based speech quality models and their modifications. Two most popular models which are referred to very often for voice call quality evaluation and proposed by ITU Telecommunication Standardization Sector (ITU-T), are E-model [IT09] and PESQ (Perceptual Evaluation of Speech Quality) [IT01]. Many discussions on the performance of these models have been carried out leading to a number of extensions to these two models. For example, in the context of network handovers and codec switch-over, a detailed study has been presented by Möller et. al. [MRK$^+$06]. Considering various roaming scenarios, they concluded that packet loss rate is the most dominant factor in deciding end user QoE while network handovers make nominal impact compared to packet loss rate and codec switch-over. Moreover, in situations when packet loss rate is very high even the most sophisticated codec cannot conceal these losses, hence, leading to poor voice quality.

Blazej et. al. [LWMV09] revealed that when narrowband and wideband codecs are used in the same call, E-model cannot predict the voice quality accurately. To help this situation they suggested a new impairment factor in E-model which reflects voice quality degradation due to codec switching. Similarly, another experimental study conducted by Mehmood et. al. [M.A10] explains how PESQ fails to provide accurate quality estimations due to several reasons like codec switching, internal time shifting of talk spurts due to instabilities of dynamic de-jitter buffer etc.

As far as video quality assessment is concerned, a range of objective models are available from simple models based on PSNR computations to advanced methods of comparing transmitted video contents with reference video using spatial and temporal correlations. For example PEVQ (Perceptual Evaluation of Video Quality) [IT08] model proposed by ITU-T is based on modeling the behavior of the human visual tract. PEVQ analyzes the picture pixel-by-pixel after a temporal alignment of corresponding frames of transmitted and reference video contents. Further discussions on video quality assessment can be found in [Net03], [BN] and [Y. 06].

Khirman et. al. [KH06] have investigated the correlation of objective measurements (QoS) and human perception (QoE) of HTTP service quality. The study focuses on the impact of content delivery latency on user satisfaction. In [H. 08], the authors consider various QoS parameters and suggest a sigmoid like function to show relationship between QoS and QoE. Another study regarding QoS and QoE relationship can be found in [F. 10], where authors present user rating as a function of response time for web applications. This study also encompasses QoS parameters like packet loss, delay & jitter as well as packet re-ordering. Further information on this subject can be accessed from [BH10] and [GR10].

A.2 Parameter Analysis for User Satisfaction Modeling

It has been mentioned earlier that a realistic representation of the user satisfaction requires the consideration of both technical and non-technical parameters. This discussion is further extended in this section, in order to identify and completely understand the influence of these parameters on perceived user satisfaction.

A.2.1 Technical Parameters

The most influential technical parameters in determining user satisfaction of both realtime and non-realtime applications are packet delay, packet loss, and bandwidth.

A.2.1.1 Impact of Delay on Different Application Types

For a VoIP application, network packet delays must be confined within a certain range in order to achieve user satisfaction. If the network latency grows beyond a certain limit, the listener hears the words and acknowledges the speaker later than a normal conversation which may cause an unnatural cadence of the conversation. Even further increasing the delays deems the conversation impractical. Figure A.2 graphically depicts the impact of end-to-end delay on the user satisfaction as found in the recommendations of the International Telecommunication Union (ITU) [IT03a]. According to this figure, three ranges of one-way delays can be established. 0–150ms delay provides transparent interactivity for the most of applications. 150–400ms delay is acceptable to allow flexible deployment of networks without making an excessive number of users annoyed. Above 400ms delay is unacceptable for general networking purposes. However, in some exceptional cases, this limit will be exceeded, e.g., double satellite hop for a hard to reach location. In practice, 200ms of delay is a reasonable goal and 250ms is the maximum acceptable latency allowable in a VoIP network.

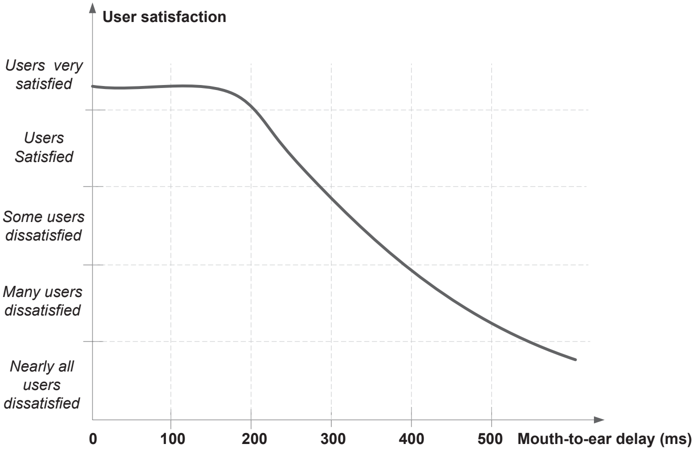


Figure A.2: Impact of VoIP packet delays on user satisfaction

The conversational or interactive video is also largely influenced by end-to-end packet delays. A delay of 150ms is an optimal value to achieve the best user satisfaction level. A value of 250ms is acceptable in most of the cases. The users get irritated over a threshold of 300ms and seriously annoyed at 500ms. Moreover, in order to achieve lip-synch (to match lip movements with spoken vocals) the audio and video streams should not be apart more than 50ms [SH04]. Similar recommendations can also be found in 3GPP standards where the delay budget for interactive video traffic has been defined as 150ms (see Table 2.2).

As far as the non-realtime video streaming is concerned, the long packet delays do not play any significant role in determining the user satisfaction. In this case, the effect of long packet delays can be eliminated by employing a play-out delay of 5 seconds or more based on network conditions and device capabilities [SH04]. However, for this application type, it is important to avoid the delay jitter grow excessively high which, otherwise, may lead to packet losses [S. 10]. The delay budget for non-interactive video streaming has been set as 300ms in LTE networks as shown in Table 2.2.

In TCP based non-realtime applications (e.g., FTP, HTTP etc.), the user satisfaction is usually determined by the provided throughput. Packet delay plays an important role in setting an upper bound the on the achievable TCP throughput. For example, in the absence of any packet loss, TCP throughput is determined by the following relation.

$$\text{TCP throughput} = \frac{W_{max}}{RTT} \tag{A.1}$$

where RTT is the TCP segment round trip time determined by the network latency. W_{max} is the maximum TCP window size in bytes. A typical value of W_{max} size is 64KBytes. This amount of W_{max} together with RTT as 300ms can provide a maximum of 1.7Mbps TCP throughput. However, when W_{max} is increased to 1MByte, it is sufficient enough to achieve 26.6Mbps throughput value. This implies that the effect of RTT can be nullified by using a proper value of TCP W_{max} if no packet losses are present.

A.2.1.2 Impact of Packet Loss on Different Application Types

There are several causes of packet losses in the network, e.g., high bit error rates of wireless access link, high levels of congestion that lead to buffer overflow in routers, link failure, high packet delays & jitter etc. In IP telephony packet losses must be controlled to make conversation possible. The extent to which packet losses can degrade user QoE depends on codec type, packet loss rate, burst length of packet losses, and packet concealment algorithm being employed. However, various investigations reveal that the quality of conversation will lag if packet loss rate exceeds 5%, provided the burst length is not very large [SU10], [Ins12], [DG03], [SXZS12]. Figure A.6 elaborates with examples how the user QoE deteriorates with packet loss rate.

Video applications are also very sensitive to packet losses and any amount of loss rate degrades the video quality. However, the extent of quality degradation depends on several factors like, burst rate of packet loss, bit rate, frame rate, and compression parameters of employed codec as well as the resolution of the video. For example, the most commonly used MPEG video codecs generate three different types of video frames (i.e., I, P, and B-frames). The composition of these frames varies for different video clips and the loss of each frame type has different effects on perceived video quality [J. 09]. Therefore, no single value of packet loss threshold can be specified as a rule of thumb to preserve user QoE for all videos. In practice, a packet loss rate value of 1% is used as a design parameter and in no case packet loss rate should exceed the upper limit of 5% [SH04] [MR02].

TCP based non-realtime applications are most susceptible to packet losses. Their performance degrades sharply even with a slight increase in packet loss rate. This is explained in greater details in Section A.3.3. As a design parameter, 3GPP standards propose packet loss rate in LTE networks to be 10^{-6} which is 1000 times higher than that of VoIP (see Table 2.2).

A.2.1.3 Impact of Bandwidth on Different Application Types

Interactive voice and video applications usually have a certain data rate requirement which is determined by the bit rate of the employed codec. If a network fails to offer this data rate, the application simply stops working. Admittedly, there are some advanced scalable audio and video coding schemes which can adapt to available bandwidth. However, they also need a minimum data rate to function and improve on delivered application quality when available bandwidth increases.

TCP based non-realtime applications are elastic in nature that they can operate at any available bandwidth. Though they don't have stringent requirements of data rate, the perceived QoE of their users is directly influenced by their achieved throughput as discussed earlier.

A.2.2 Non-Technical Parameters

User satisfaction is also effected by various non-technical parameters including user preference over the service cost, reputation of the operator / service provider, etc. Such parameters are of diverse scope and have a relatively more subjective nature compared to technical parameters. Moreover, the influence of non-technical parameters on user satisfaction is specific to the service types. They can be normalized on expectancies of the lower the better, the higher the better, or the nominal the better. The degree of impact of these parameters on user satisfaction is purely attribute dependent, i.e., the decision of using linear, exponential, logarithmic functions and control parameters depend on the attribute under consideration, e.g., impact of security parameters may be modeled using a sigmoid like function where users remain satisfied if a certain level of data encryption is achieved. Further beefing up security beyond that level does not bring more satisfaction to the users.

A.3 Measurement and Evaluation of User Satisfaction

This section discusses the various means of predicting user perception of packet-switched service quality from parameters that objectively describe the access network quality. These methods are capable of translating the effects of packet delay & jitter and dropped packets on user assessment of the respective service. As far as real time services like conversational voice and video conferencing are concerned, there are two distinct ways of their quality evaluation which are described below.

1. ***Subjective Measurements***

   Subjective tests are considered as the most reliable medium for obtaining a measure of user perceived quality of a service. The reason is that, in a subjective test, user assessments of quality are elicited and collected directly from typical user of that service. The user responses are then mapped to the widely used Mean Opinion Score (MOS) scale which range from 1 to 5 where 1 is the worst and 5 is the best perceived quality. In order to achieve credible and meaningful results from subjective measurements, the testing must be carefully structured, controlled, and standardized for a particular service, creating daunting test requirements like those presented in [IT98] for voice applications.

Though subjective tests can produce results which are intuitively credible, operationally meaningful and scientifically defensible, they are not always preferred. This is due the fact that such tests are very time consuming, expensive, and require a lot of resources. Moreover, monitoring of real-time performance is not always practical. Subjective tests are also of no choice for network planning purposes. These limitations give rise to alternative ways of indirect quantification of service quality termed as 'objective measurement' as discussed in the following.

2. ***Objective Measurements***

   Objective measurements of quantifying service quality and usability are based on measures of characteristics of the underlying network connection (e.g., throughput, latencies etc.). In other words, they rely on technical parameters which are used to describe the performance of communication networks. There are numerous advantages associated with objective measurements like, input data can be collected readily and automatically, output data can be interpreted without having to deal with the vagaries of human opinion, the tests can be replicated in different environment, scalability issues are non-existent etc. Objective measurements involve a modeling of the human auditory and visual system, low level neural processing and higher level cognitive processing. An objective measurement of a service quality will be considered more accurate if it has higher correlation with the subjective measurements. Objective measurement methods fall into following three categories.

   - *Full Reference*: This approach requires the access to original reference multimedia contents (i.e., audio or video file) that is assumed to have perfect quality. This reference signal (at sender side) is compared with possibly degraded signal (at the receiver side) to compute distortion levels produced during the transmission. For example, a received video transmission can be evaluated using the sent reference video and performing pixel by pixel comparison. The full reference quality evaluation methods provide the highest accuracy and repeatability but tend to be processing intensive.
   - *No Reference*: This approach assumes no access to the original reference signal and therefore relies only on the received signal to make the quality estimation. Instead of analyzing the received signal in greater depth, commonly found 'no-reference' methods are based only on an analysis of the digital bit stream at an IP packet level. As a results, performance of 'no-reference' methods is usually inferior to that of full reference methods.

- *Reduced Reference*: This approach, which is usually used for video quality evaluation, lies between above described full reference and no-reference approaches. In this approach certain features are extracted from reference signal at the sender side which are used along with the received signal to evaluate the service quality at the receiving end (See Figure.A.3). An efficient reduced reference method requires minimum feature data to predict a service quality which has high correlation with the results obtained from full reference methods.

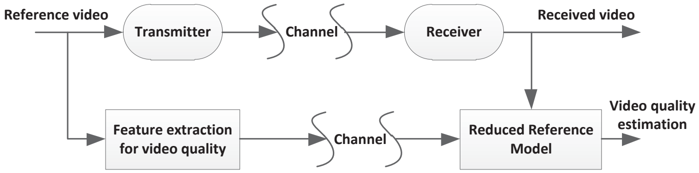


Figure A.3: Deployment of reduced-reference video quality assessment system

A.3.1 User Satisfaction Models for Conversational Voice

A.3.1.1 PESQ Model

PESQ (Perceptual Evaluation of Speech Quality) model provides a full reference objective voice quality measurement tool standardized in ITU-T recommendation P.862 [IT01]. PESQ has evolved from many new developments like PSQM, PAMS, MNB and, therefore, considered as the state-of-the-art model to assess end-to-end speech quality with confidence. PESQ takes into account following sources of voice signal degradation: packet loss, delay and jitter, coding distortions and errors as well as filtering in analog network components.

A overview of the structure of the PESQ algorithm can be seen in Figure A.4. In the first step, reference signal and received signal are aligned to a standard listening level. PESQ assumes 79dB as subjective listening level at the ear reference. In order to bring both signals at this level a gain function is applied. In the next step, PESQ compensates for any filtering that has taken place in the network with the help of input filters. Time alignment for the two signals is also required in order to counterbalance the variable delays of the network. Time alignment is performed in three stages: First, talk spurts are time aligned. Second, overlapping sections of the speech are aligned through the detection of delays which are variable over the length of a talk spurt. Finally, those sections of speech which undergo from very large distortion (called bad intervals) are realigned.

The next operation is the auditory transformation. This is essentially a psychoacoustic model that mimics certain key properties of human hearing. This gives a representation in

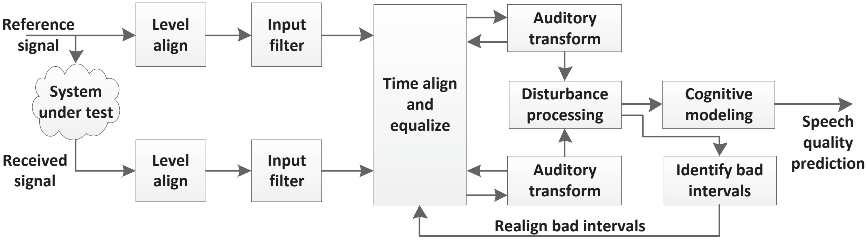


Figure A.4: Structure of perceptual evaluation of speech quality (PESQ) model [RBHH01]

time and frequency of the perceived loudness of the signal, termed as the sensation surface. The difference between sensation surfaces of reference and received signals represents audible differences introduced by the network. This analysis, at disturbance processing and cognitive modeling stage, yields two disturbance parameters, i.e.,

- *Absolute (symmetric) disturbance* - It is a measure of absolute audible error.
- *Additive (asymmetric) disturbance* - It is a measure of audible errors that are significantly louder than the reference signal.

In final step, these error parameters are linearly combined and converted to a quality score. The score which lies between 4.5 and 1 represents the measure of user's perception of quality. The highest score 4.5 indicates that the received signal bears no distortion. The score falls as the amount of distortion increases.

Rix et. al. [RBHH01] has shown in their experimental study that PESQ model provides significantly higher correlation with subjective measurements when compared to other popular models in the same class.

A.3.1.2 E-Model

E-model [IT09] is a no-reference computational model used as a transmission planning tool for assessing the effects of transmission parameters which decide conversational voice quality. The primary output of the E-model is the “rating factor" *R* or R-factor which can be mapped to MOS in order to estimate the user opinion on voice quality. The E-model assesses conversational voice quality by establishing a relationship between objectively measurable factors and subjective assessment of voice quality based on large scale of measurements. For a narrowband codec, the maximum value of R-factor computed by the E-model is 100, which corresponds to the best possible achievable voice quality. For a wideband codec, this value is ranged up to 129. The minimum value of *R* factor is 0, which represents the worst quality. *R* factor combines several transmission parameters which are considered relevant for an end-to-end transport connection: In addition it also reckons others elements which cover impairments due to low bit rate codecs, echo, background noise, and electronic equipment, etc.

The R factor is composed of five factors as stated below.

$$R = R_o - I_s - I_d - I_{\text{e-eff}} + A \tag{A.2}$$

The factor R_o represents the basic signal-to-noise ratio of a given environment of the talker. I_s is the sum of all impairments which may occur more or less simultaneously with the voice transmission. I_d represents impairments caused by mouth-to-ear path delay. The 'advantage factor' A provides compensation for impairments in return of other benefits enjoyed by the user, e.g. ease of access etc. $I_{\text{e-eff}}$ is packet loss dependent 'Effective Equipment Impairment' factor.

Considering the packet-loss probability P_{pl} and packet loss robustness factor B_{pl}, the value of $I_{\text{e-eff}}$ is calculated using the following equation.

$$I_{\text{e-eff}} = I_e + (95 - I_e)\frac{P_{pl}}{B_{pl} + \frac{P_{pl}}{BurstR}} \tag{A.3}$$

$BurstR$ is the so-called burst ratio; which is defined as:

$$BurstR = \frac{\text{Average length of observed bursts in an arrival sequence}}{\text{Average length of bursts expected under “random” loss}}$$

If packet losses are random (i.e., uncorrelated) $BurstR = 1$; and when packet losses occur in bursts (i.e., dependent or correlated) $BurstR > 1$. In this thesis work, pure random packet losses are considered, i.e. $BurstR = 1$. Planning values of the packet loss robustness factor B_{pl} are provided in ITU-T Recommendation G.113 [IT07] for several popular codec schemes.

The value of the I_e factor is obtained from subjective measurements of voice quality with various codecs and various operating conditions, e.g., packet loss, packet size, etc. The packet loss concealment algorithms of a given codec also influence I_e values.

Packet loss rate P_{pl} in equation A.3 has the following two components,

- *Packet loss rate*: All packet losses on the way from source node to destination node due to transport network link impairments, buffer overflows in transport network routers, etc.
- *Packet drop rate*: The packets which are dropped if they get delayed more than the length of the de-jitter buffer.

The factor I_s is a function of parameters which are independent of the underlying transport network. Therefore, a default planning value provided by ITU-T G.107 [IT09] can be used for simplification. Moreover, substituting a value of 100 for R_o shortens equation A.2 as follows:

$$R = 94.2 - I_d - I_{\text{e-eff}} + A \tag{A.4}$$

Cole et. al. [CR01] have analyzed I_d in great detail. They have provided a mathematical expression based on curve fitting functions to evaluate impairments caused by one way mouth-to-ear delay. This expression for I_d which is applicable for all codecs is defined as follows:

$$I_d = 0.024d - 0.11(d - 177.3)H(d - 177.3) \tag{A.5}$$

$H(\cdot)$ is heavyside function and d is one way delay. In pure IP networks d has the following four components.

- *Codec delay*: Encoding delay, look-ahead delay etc. This value is given in codec specifications.
- *Packetization delay*: It occurs when more than one voice frame are transported in one IP packet.
- *De-jitter buffer delay*: This is the delay associated with the packet waiting in de-jitter buffer. It shows up when packets are received out-of-sequence at destination or when there are packet losses on the way from sender to the destination. This delay value can go up to the maximum de-jitter buffer length value.
- *Compressing and de-compressing delay*: The processing delays associated with compression/decompression of data inside voice frames.

The R factor produced by the E-model is mapped to MOS scale which ranges from 1 to 5, 1 being the worst and 5 the best perceived quality. The expression used to map R factor of narrowband codecs onto MOS scale can be found in Appendix B of ITU-T G.107 [IT09]. A similar expression can also be derived for wideband codecs for whom R value ranges from 1 to 129. Table A.1 shows such a mapping of R factor on MOS scale.

Table A.1: User satisfaction level on MOS and R scale for wideband and narrowband codecs

R_{nb}–value (lower limit)	MOS–value (lower limit)	R_{wb}–value (lower limit)	User satisfaction
90	4.34	116.1	Very satisfied
80	4.03	103.2	Satisfied
70	3.6	90.3	Some users dissatisfied
60	3.1	77.4	Many users dissatisfied
50	2.58	64.5	All users dissatisfied

The Figure A.5 depicts the influence of packet loss rate and mouth-to-ear delay on achievable MOS score as predicted by E-model for wideband codec G.722.2 (23.05kbps) codecs.

The E-model defined by equations A.3-A.5 has been implemented by the author in OPNET simulator. This makes possible it to use E-model in evaluation of VoIP call quality, in simulation based studies presented in this work.

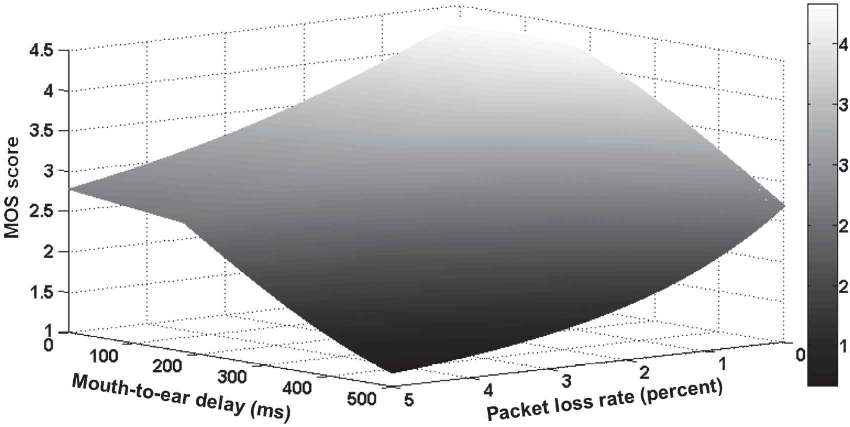


Figure A.5: 3-dimensional plot of VoIP wideband MOS score variations due to packet loss rate and mouth-to-ear-delay

A.3.1.3 $\mathrm{E^{IP}}$-Model

After having a detailed overview of PESQ and E-model, it would be interesting to compare their performances. Such a comparison study has been carried out by Uhl. et. al. [Uhl08]. Their study reveals that within the scope of IP networks PESQ model delivers improved results with high correlation to subjective measurements. However, E-model performance degrades substantially in the presence of a high packet loss rate in the IP networks.

In another research work [SU10] by the same authors, an enhancement has been proposed to the standard E-model in order to obtain more reliable results for lossy IP links. The enhancement comes from the investigations on the effects of average burst length of packet losses and speech sample length on voice quality. As a result, a new parameter BSLP (Burst Sample Length Product) is introduced. From the experimental study on PESQ behavior in diverse end-to-end network conditions and for various codecs, the values of B_{pl} and I_e are obtained in terms of BSLP parameter. For example, in the case of G.726(32 kbps) codec, these values are as shown below,

$$B_{pl} = 0.0634 \cdot BSLP + 20.815, \tag{A.6}$$

$$I_e = -0.01 \cdot BSLP + 17.76, \tag{A.7}$$

This modified E-model is named as $\mathrm{E^{IP}}$-Model by the developers. Figure A.6 shows that $\mathrm{E^{IP}}$-Model provided results are very close to that of the PESQ model for different packet loss rate and burst size values. On the other hand, the standard E-model provided curve deviates from the other two models.

The high accuracy of E^{IP}-Model makes it a preferable method for voice quality evaluation. Though this model has been implemented by the author in OPNET based network simulator, it has not been used in this thesis work due to two reasons. First, E^{IP}-Model has not yet been validated for any wideband codec. Owing to the fact that all simulation based studies presented in this thesis work employ wideband codecs, E^{IP}-Model is not a choice. Second, the investigations made in this thesis work do not deal with high loss IP links, therefore the original E-model still delivers the satisfactory performance in these scenarios.

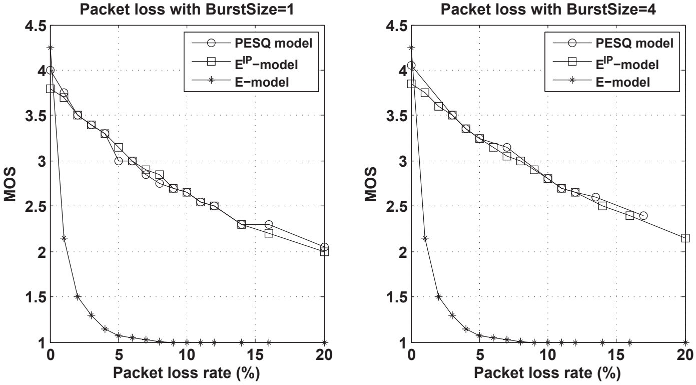


Figure A.6: MOS values as a function of nondeterministic distributed packet loss with BurstSize equal to 1 and 4 [SU10]

A.3.2 User Satisfaction Models for Conversational Video

Similar to VoIP, video quality at receiving end can also be determined using Subjective as well as Objective evaluation techniques. Most state-of-the-art objective evaluations of video quality metrics attempt to model the Human Visual System (HVS). The principle behind HVS based metrics is to process the visual data by simulating the visual pathway of the eye-brain system. Digital Video Quality (DVQ) metric [A. 01] and the Perceptual Distortion Model (PDM) [Win99] are the examples where HVS-based video quality metrics have been proposed. However, HVS-based quality metrics suffer from inaccurate modeling of the HVS. In particular, temporal mechanisms in the HVS is a likely source of performance loss as indicated in a study by Video Quality Experts Group [VQE00]. Hence, the performance of HVS based algorithms has a considerable room for improvement.

Among other popular video quality metrics are SNR (peak signal to noise ratio), Ssim (Structural similarity), and MDI (Media Delivery Index) which are computed by a large number of full reference models.

- **PSNR**: It is a derivative of the well-known signal to noise ratio (SNR) metric. PSNR term defines the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. When comparing two video files, signal is the original file and noise is the error which occurs due to compression or during transmission over the network. In the context of video quality evaluation, PSNR is taken as an approximation to human eye perception of image quality. It is measured in decibel units (dB).
- **Ssim Index**: Ssim exploits the well-defined structures found in natural image signals which carry important information about the visual scene. This information plays an important role in human visual system to perceive image quality. Ssim based models perform structural distortion measurements instead of just computing error signal power. Structural Similarity (Ssim) index gives a measure of the similarity between two images. Ssim index value ranges from -1 to 1. Higher the Ssim index value, higher the similarity between the two comparing images. For videos quality evaluation, Ssim index is computed image by image.
- **Media Delivery Index**: An interesting metric to evaluate IP based transport network performance for video streaming is MDI. Though it does not quantify user perception of a video quality, it provides a set of measures (e.g., packet delay and jitter in the transport network which are main causes for quality loss) to help monitor delivered video quality. MDI can be used in network planning phase as well as in network monitoring which allows network operators to take necessary corrective actions well in advance. A set of MDI values for different type of video streaming applications like SDTV, HDTV, Video-on-demand etc., have been recommended by IETF [WC06].

A.3.2.1 ITU-T G.1070 Model

The opinion model discussed in ITU-T G.1070 recommendation encompasses several input parameters related to video and speech quality which influences user satisfaction or QoE. This computational model consists of three functions, namely, video quality estimation, speech quality estimation, and multimedia quality integration functions. In first step, speech quality is estimated based on the E-model discussed in section A.3.1.2 and video quality V_q is calculated based on the relation given in equation A.8. In the second step, speech and video quality estimations are combined using an integration function to estimate overall multimedia quality.

$$V_q = 1 + I_{coding} \cdot exp\left(-\frac{P_{pl_V}}{D_{P_plV}}\right) \tag{A.8}$$

where I_{coding} represents the basic video quality determined by the codec distortion and is a function of the frame rate and bit rate. P_{pl_V} represents the packet loss rate and D_{P_plV} is the degree of video quality robustness against the packet losses. The basic video quality for a certain bit rate Br_V and frame rate Fr_V can be calculated using the following relation.

$$I_{coding} = I_{Ofr} \cdot exp\left(\frac{\left(ln(Fr_V) - ln(O_{fr})\right)^2}{2D_{FrV}^2}\right) \tag{A.9}$$

O_{fr} is an optimal frame for maximum achievable video quality and I_{Ofr} is the maximum video quality at bit rate Br_V so that,

$$O_{fr} = v_1 + v_2 Br_V, \quad 1 \leq O_{fr} \leq 30 \tag{A.10}$$

$$I_{Ofr} = v_3 \cdot \left(1 - \frac{1}{1 + (b/v_4)^{v_5}}\right), \quad 0 \leq I_{Ofr} \leq 4 \tag{A.11}$$

Moreover, D_{FrV} which represents the video quality robustness due to frame rate F_{r_V} and packet loss robustness factor D_{P_plV} are expressed as follows,

$$D_{FrV} = v_6 + v_7 Br_V, \quad D_{FrV} > 0 \tag{A.12}$$

$$D_{P_plV} = v_{10} + v_{11} \cdot exp\left(-\frac{F_{r_V}}{v_8}\right) + v_{12} \cdot exp\left(-\frac{Br_V}{v_9}\right), \quad D_{P_plV} > 0 \tag{A.13}$$

where coefficients v_1, v_2,...,v_{12} are dependent on codec type, video display size, key frame interval and display format. Provisional values for these coefficients have been provided in the ITU-T G.1070 document based on subjective tests for MPEG-4 codec in QVGA and QQVGA formats. Belmudez et. al. [BM10] has provided a new set of parameters for MPEG-2 codec. Another extension has been made by Yamagishi et. al. [YH08] who supplied coefficient values for H.264 codec in HD format.

The Figure A.7 shows the deterioration of video MOS score due to packet losses as predicted by G.1070 model.

A.3.2.2 PSNR Based Quality Evaluation

EvalVid is a framework and tool-set for evaluation of the video quality transmitted over a real or simulated communication network. It is capable of analyzing the transport network performance used for video streaming in terms of QoS parameters like, packet loss, delay & jitter. As an output, it provides both the frame by frame PSNR values and an overall MOS score as a prediction of end user's perception of received video quality. Evalvid toolkit is based on full reference service quality measurement methods which has been integrated into the heterogeneous network simulator developed in chapter 3. It has been used as a default video quality evaluation tool in all simulation based studies presented in this work.

The process of obtaining MOS values for a video streaming application in OPNET can be split into three phases i.e., pre-processing, online-processing, and post-processing. In the following a detailed description of each phase is provided.

- **Pre-processing phase**

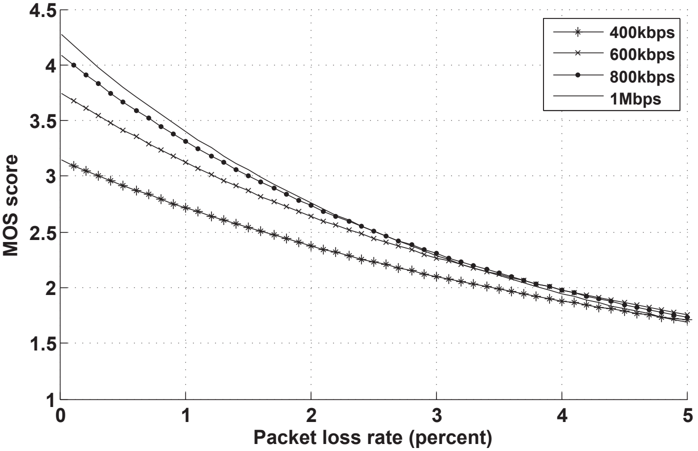


Figure A.7: Video MOS variations due to packet losses at different MPEG-4 codec bit rates

- In first step, a video clip of a certain time length, resolution, and frame size is selected which is used for video streaming application in OPNET. With the help of Evalvid took-kit the selected video clip is converted to raw YUV format making it ready for encoding in the desired format.
- Evalvid tool-kit supports MPEG-4, H.264, and H.263 codecs. Based on the scenario configuration the video clip is encoded to a preferred codec format with the desired bit rate. In this work, the default codec is selected as MPEG-4 due to its widespread use in Internet applications.
- EvalVid toolkit is then used to generate a trace of packet transmission for that particular encoded video clip. For this purpose, the video clip is streamed over a real IP network and all IP packets belonged to the transmission are captured using network sniffing tools, like, Tcpdump [Ana13] etc. A simple analysis of this trace file helps extract the information of packet sizes and their inter-arrival times.

- **Online-processing phase**
  - When running a simulation setup, the OPNET takes the above extracted information to generate UDP based video streaming traffic in the scenario. These UDP packets are marked with sequence numbers in order to detect packet loss and reordering at the receiving end.
  - At the receiving end in the simulation scenario, another trace file is generated which contains the information about the received video stream, like, end-to-end delay and losses for all packets.

- **Post-processing phase**
  - When the simulation ends, the two trace files generated in previous phases, along with the reference video clip, are fed to Evalvid took-kit. It compares two trace files to detect the packet losses during the transmission in simulation environment. With the help of this information and reference video clip, it constructs the received video file. In the construction of video file all those packets which are delayed greater than specified play-out or de-jitter buffer size are treated as lost packets.
  - In the final step, EvalVid took-kit takes the reference video file and received video file to compute peak signal-to-noise ratio (PSNR) values in frame by frame manner. Based on these PSNR values a MOS value is also computed by the tool following the mapping the Table A.2.

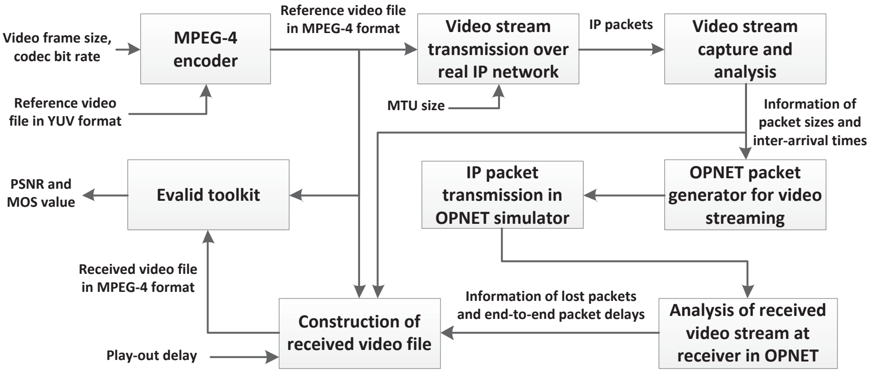

Video frame size, codec bit rate
MPEG-4 encoder
Reference video file in MPEG-4 format
Video stream transmission over real IP network
IP packets
Video stream capture and analysis
Reference video file in YUV format
MTU size
Information of packet sizes and inter-arrival times
PSNR and MOS value
Evalid toolkit
IP packet transmission in OPNET simulator
OPNET packet generator for video streaming
Received video file in MPEG-4 format
Information of lost packets and end-to-end packet delays
Construction of received video file
Analysis of received video stream at receiver in OPNET
Play-out delay

Figure A.8: Video quality evaluation in OPNET simulator using Evalvid

Table A.2: Mapping of PSNR values onto MOS scale [Ohm99]

PSNR (dB)	MOS
> 37	5 (Excellent)
31–37	4 (Good)
25–31	3 (Fair)
20–25	2 (Poor)
< 20	1 (Bad)

A.3.3 User Satisfaction Models for TCP Applications

Generally, realtime applications (i.e., VoIP call, video conference call etc.) are very sensitive to the path delays than the losses. This makes UDP transport protocol a natural choice for such applications. In contrast to this, non-realtime applications (i.e., FTP, HTTP etc.) require an error free delivery of the contents by compromising the transfer delays. Such requirements are fulfilled by the underlying TCP transport protocol. TCP retransmits the lost and corrupted packets, ensures in sequence deliveries, as well as, adapts to the available bandwidth of the link by controlling the data transfer rate. With the help of all these mechanism, TCP guarantees error free delivery of the data. Hence, for non-realtime applications, user satisfaction cannot be evaluated based on transmission errors in the delivered contents. Instead, the metric for user satisfaction is directly related to waiting time required for successful completion of data transfer, e.g., File download time, HTTP page response time etc. Owing to the fact that the content transfer time is determined by user throughput, the achieved user throughput also serves a QoE metric for TCP based non-realtime applications.

An error free delivery of contents is made possible in TCP through its error control mechanism. This ARQ (Automatic Repeat-reQuest) mechanism of TCP uses acknowledgements and timeouts to retransmit those packets which are lost or extensively delayed in the network. Such retransmissions, in turn, cause to reduce the achievable throughput as seen by the above application. Therefore, it is important to analyze the influence of the packet loss and delays on the achievable throughput of TCP based applications. In a simple scenario where negligible packet losses are introduced by the network, TCP throughput is inversely proportional to the end-to-end packet delay as indicated by equation A.1. However, in practice, TCP connection is subject to packet losses, for example, due to network congestion etc. Mathis et. al. [M. 97] studied the effect of such packet losses on the TCP performance with the help of TCP congestion avoidance algorithm model. They proposed a TCP throughput model with selective acknowledgements considering a wide range of Internet conditions. However, it is also applicable to other TCP implementations under restricted conditions. The model which predicts the achievable TCP throughput has been validated by the authors using simulations, as well as, the live Internet measurements. The model is described as follows:

$$\text{throughput} = \frac{MSS}{RTT} \cdot \frac{C}{\sqrt{p}} \tag{A.14}$$

where MSS is the maximum TCP segment size, p is the packet loss rate and C is the constant of proportionality which depends on TCP implementation. Nowadays, in most of the TCP implementations "Delayed Acknowledgement" algorithm [SW94] is used to reduce the number of tiny TCP acknowledgement packets in the network. For such TCP implementations C is normally less than 1. Thus, in many practical situations a simpler bound on TCP throughput can be used, i.e.,

$$\text{throughput} < \frac{MSS}{RTT} \cdot \frac{1}{\sqrt{p}} \tag{A.15}$$

As mentioned earlier, the above model is valid for packet loss rates up to a moderate level ($p < 2\%$). Padhye et. al. [P. 98] extended this investigation to develop an improved model which captures the effects of TCP retransmission mechanism and timeout mechanism on the achievable throughput. The predicted TCP throughput is given by the following relation:

$$\text{throughput} = \begin{cases} \dfrac{\frac{1-p}{p} + E[W] + \hat{Q}(E[W])\frac{1}{1-p}}{RTT(\frac{b}{2}E[W]+1) + \hat{Q}(E[W])T_0\frac{f(p)}{(1-p)}} & \text{if } E[W] < W_{max} \\ \dfrac{\frac{1-p}{p} + W_{max} + \hat{Q}(W_{max})\frac{1}{1-p}}{RTT(\frac{b}{8}W_{max} + \frac{1-p}{pW_{max}} + 2) + \hat{Q}(W_{max})T_0\frac{f(p)}{(1-p)}} & \text{otherwise} \end{cases} . \tag{A.16}$$

where W_{max} is the maximum TCP window size, b is the number of segments acknowledged by one TCP acknowledgement packet, p is the packet loss rate and T_0 is the initial retransmit timeout value. $\hat{Q}(w)$ is the probability that a loss in a window of size w is a TCP time-out.

$$E[W] = \frac{2+b}{3b} + \sqrt{\frac{8(1-p)}{3bp} + \left(\frac{2+b}{3b}\right)^2},$$

$$\hat{Q}(w) = \min\left(1, \frac{(1-(1-p)^3(1+(1-p)^3(1-(1-p)^{w-3}))}{1-(1-p)^w}\right) \approx \min\left(1, \frac{3}{w}\right),$$

$$f(p) = 1 + p + 2p^2 + 4p^3 + 8p^4 + 16p^5 + 32p^6$$

An approximation of the model in equation A.16 is also provided by the same authors as shown below,

$$\text{throughput} \approx \min\left(\frac{W_{max}}{RTT}, \frac{1}{RTT\sqrt{\frac{2bp}{3}} + \min\left(1, 3\sqrt{\frac{3bp}{8}}\right)p(1+32p^2)T_0}\right) \tag{A.17}$$

Figure A.9 provides a graphical illustration of how TCP throughput is influenced by the packet losses and path delays. The curves have been drawn using equation A.17 and considering a TCP windows size of 1MByte. It can be observed that both of the parameters (i.e., packet losses and delay) have a significant impact on user throughout and, hence, on the perceived QoE.

A.4 Generic User Satisfaction Model

The previous section has presented a number of objective models to evaluate user QoE for each type of realtime and non-realtime applications. Owing to the fact that the outcome of these models depends on a confined number of technical parameters, it is possible to define a generic user satisfaction model which encompasses all necessary parameters to evaluate user QoE for any predefined application type. Such a generic user satisfaction

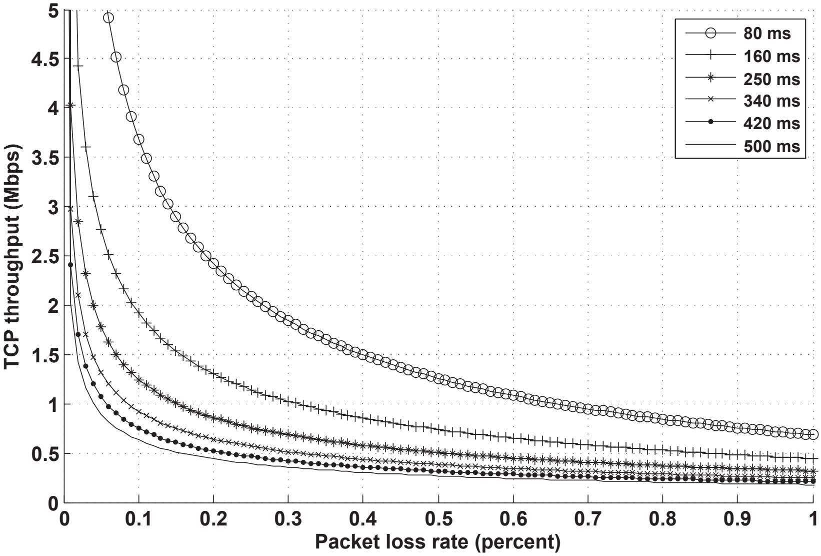


Figure A.9: TCP throughput degradation due to packet losses for different RTT values

model is composed of several linear and non-linear functions whose behavior for a certain application is determined by a set of predefined attributes for a set of predefined application types. In addition to technical parameters, such a model can also consider non-technical parameters as discussed in the beginning of this Chapter. The development of this generic user satisfaction model requires an extensive discussion and, therefore, is beyond the scope of this work. In another research work related to network selection, the author has realized such a model whose details can be found in [TKGTG11] and [KT11].

A generic user satisfaction model is of great importance in making crucial decisions at different levels of the telecommunication paradigm. At the user level, it can help users in network selection by predicting their satisfaction level achievable from a certain network. At the cell level, a base station's decisions about handover optimization and link adaptation can be derived from this model. At the network level, an operator can get help from this model in achieving optimal resource allocation and enhancing user QoE in an environment of heterogeneous networks. In addition, in an environment where no long term contracts exist and the users are free to choose a network operator on a per application usage basis, the estimation of the user satisfaction is of prime importance. In such an environment, the model can help an operator compete in the market by computing better service offers. The details of author's work on use of this model in aforementioned scenarios is available in [M. 10b], [M. 10a], and [TKGTG11].

B The Box Plot

The box plot or box-and-whisker plot was introduced by John Tukey in 1977. It is one of the standardized was of displaying the distribution of data based on the five number summary: (1) minimum, (2) maximum, (3) median (or second quartile), (4) the first quartile, and (5) the third quartile. An example of a standard box plot is shown in Figure B.1.

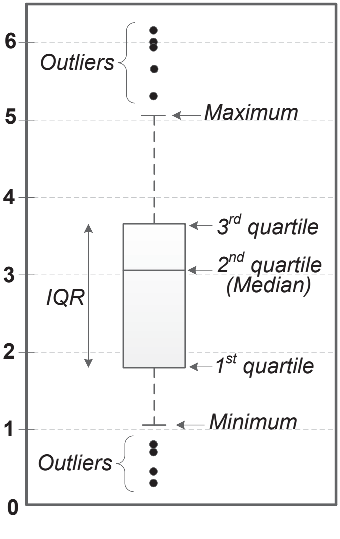


Figure B.1: An example of the box plot. IQR is the inter-quartile range.

It can be observed in the figure that:

- The box (the central rectangular portion of the plot) extends from the first quartile to the third quartile. The length of the box indicates the IQR (inter-quartile range) which is the middle half (the interquartile range) of the ordered data.
- A horizontal line segment inside the box shows the median. This helps illustrate the skewness pattern of the data. For example, if most of the data samples are concentrated on the low end of the scale, the distribution is skewed right. If the opposite is true then the distribution is skewed left. Moreover, if the median line evenly splits the box, it is an indication of symmetric distribution.

- Two vertical lines, called whiskers, extend from the top and bottom of the box.
- The lower whisker extends from minimum to the first quartile. The length of this whisker indicates the range of the lowest fourth of the ordered data.
- The upper whisker extends from the third quartile to maximum. The length of this whisker indicates the range of the highest fourth of the ordered data.
- The portion of the box between the first quartile and median indicates the range of the second fourth of the ordered data.
- The portion of the box between median and the third quartile indicates the range of the third fourth of the ordered data.
- The values that fall beyond the end of whiskers, have been plotted as dots. They are called outliers due to their extremeness relative to the bulk of the distribution.

C LTE Curve Fitting Data

The data presented in Table C.1 has been generated using “Curve Fitting Toolbox” of MATLAB software. The accuracy of the curve fitting is shown in the table using ‘norm of residuals’ and square of ‘correlation coefficient’. According to MATLAB help, the norm of residuals is computed using the following formula:

$$\text{norm of residuals} = \sqrt{\sum_{i=1}^{n} {d_i}^2}$$

where d represents the numerical difference of an original data point and its approximation. The n represents the number of data points in the sample.

Table C.1: Curve fitting data to represent a linear relationship between PRBs and LTE throughput at different TBS indices. The values of α and β are used in Equation 6.4 to get a relationship between 'number of PRBs' and 'achievable user throughput' for a certain TBS index.

TBS index	α	β	**Norm of residuals**	**Square of Correlation Coefficient**
0	3.581×10^{-2}	7.162×10^{-1}	0.9271	0.9993
1	2.749×10^{-2}	4.536×10^{-1}	1.2605	0.9988
2	2.250×10^{-2}	2.866×10^{-1}	0.9729	0.9993
3	1.716×10^{-2}	3.054×10^{-1}	0.8658	0.9994
4	1.387×10^{-2}	3.144×10^{-1}	0.8423	0.9995
5	1.124×10^{-2}	3.114×10^{-1}	0.6987	0.9996
6	9.637×10^{-3}	-2.859×10^{-2}	2.3870	0.9956
7	8.057×10^{-3}	2.486×10^{-1}	0.9057	0.9994
8	7.052×10^{-3}	2.270×10^{-1}	0.5489	0.9998
9	6.280×10^{-3}	1.599×10^{-1}	0.7211	0.9996
10	5.633×10^{-3}	1.719×10^{-1}	0.6484	0.9997
11	4.966×10^{-3}	5.924×10^{-2}	0.7752	0.9995
12	4.336×10^{-3}	1.306×10^{-1}	0.7100	0.9996
13	3.840×10^{-3}	1.404×10^{-1}	0.5839	0.9997
14	3.476×10^{-3}	5.404×10^{-2}	0.9189	0.9994
15	3.267×10^{-3}	1.358×10^{-2}	0.7915	0.9995
16	3.086×10^{-3}	-4.254×10^{-2}	0.8843	0.9994
17	2.758×10^{-3}	5.194×10^{-2}	0.8582	0.9994
18	2.515×10^{-3}	4.120×10^{-2}	0.6140	0.9997
19	2.323×10^{-3}	6.142×10^{-3}	0.5950	0.9997
20	2.159×10^{-3}	-2.557×10^{-3}	0.7571	0.9996
21	1.983×10^{-3}	7.328×10^{-2}	0.6418	0.9997
22	1.847×10^{-3}	8.933×10^{-2}	0.6624	0.9997
23	1.755×10^{-3}	-5.122×10^{-2}	0.8799	0.9994
24	1.628×10^{-3}	7.830×10^{-2}	0.5729	0.9997
25	1.563×10^{-3}	1.009×10^{-1}	0.6861	0.9996
26	1.359×10^{-3}	-4.971×10^{-2}	0.6814	0.9997

Bibliography

[A.07] A. Pokhariyal, K.I. Pedersen, G. Monghal, I.Z. Kovacs, C. Rosa, T.E. Kolding, and P.E. Mogensen. HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution. In *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, pages 2761–2765, April 2007.

[25.06] 3GPP Technical Report TS 25.814. Physical Layer Aspects For E-UTRA. Technical Report version 7.1.0, 3rd Generation Partnership Project, September 2006.

[36.11] 3GPP Technical Specification TS 36.300. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2. Technical Report Version 10.3.0, 3rd Generation Partnership Project, March 2011.

[36.12] 3GPP Technical Report TS 36.213. Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures. Technical Report version 10.7.0, 3rd Generation Partnership Project, September 2012.

[3GP08] 3GPP Technical Specification TS 23.401. General Packet Radio Services (GPRS) Enhancements For Evolved Universal Terrestrial Radio Access Network (EUTRAN) Specification. TS version 8.2.0, 3rd Generation Partnership Project (3GPP), December 2008.

[3GP10] 3GPP Technical Specification TS 36.322 . Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) Protocol Specification. Technical Report Version 10.0.0, 3rd Generation Partnership Project, December 2010.

[3GP11a] 3GPP Technical Specification TS 23.402. Architecture Enhancements For Non-3GPP Accesses. TS version 10.6, 3rd Generation Partnership Project (3GPP), December 2011.

[3GP11b] 3GPP Technical Specification TS 36.321 . Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification. Technical Report Version 10.2.0, 3rd Generation Partnership Project, June 2011.

[3GP11c] 3GPP Technical Specification TS 36.323. Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) Specification. Technical Report Version 10.1.0, 3rd Generation Partnership Project, March 2011.

[3GP12] 3GPP Technical Specification TS 23.203. Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture. TS version 10.6.0, 3rd Generation Partnership Project (3GPP), March 2012.

[72699] ETS 300 726. Digital Cellular Telecommunications System (Phase 2+) (GSM); Enhanced Full Rate (EFR) Speech Transcoding. Technical Report GSM 06.60 version 5.2.1, European Telecommunications Standards Institute, July 1999.

[A. 01] A. B. Watson, J. Hu, and J. F. Mcgowan. DVQ: A Digital Video Quality Metric Based on Human Vision. *Journal of Electronic Imaging*, 10:20–29, 2001.

[A. 03] A. Aditya, M. Bruce, S. Srinivasan, S. Anees, and S. Ramesh. A Measurement-based Analysis of Multihoming. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '03, pages 353–364, New York, NY, USA, 2003. ACM.

[A M13] A Mathematical Programming Language. `http://www.ampl.com/`, last accessed in May 2013.

[AB12] RJ. Atkinson and SN. Bhatti. ILNP Architectural Description. DRAFT 06, Internet Engineering Task Force, 2012.

[ABG03] J. Abley, B. Black, and V. Gill. Goals for IPv6 Site-Multihoming Architectures. RFC 3582, Internet Engineering Task Force, 2003.

[Agi11] Agilent. Introducing LTE-Advanced. *Application Note*, March 2011.

[AM07] T. Altiok and B. Melamed. *Simulation Modeling and Analysis with ARENA*. Academic Press, ISBN 978-0123705235, June 2007.

[Ana13] Tcpdump Packet Analyzer. `http://www.tcpdump.org/`, last accessed in May 2013.

[Ban98] J. Banks, editor. *Handbook of Simulation: Principles, Methodology, Advances, Applications and Practice*. John Wiley, New York, 1998.

[BGMA07] M. Bagnulo, A. Garcia-Martinez, and A. Azcorra. IPv6 Multihoming Support in the Mobile Internet. *IEEE Wireless Communications*, 14(5):92–98, October 2007.

[BH10] P. Brooks and B. Hestnes. Being Objective and Quantitative About User Measurements of QoE. *IEEE Network Communication Megazine*, March 2010.

[Bia00] G. Bianchi. Performance Analysis of The IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, March 2000.

[BM10] B. Belmudez and S. Möller. Extension Of The G.1070 Video Quality Function For The MPEG2 Video Codec. In *Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 7–10, June 2010.

[BN] White paper Brix Networks. *Video Quality Measurement Algorithms: Scaling IP Video Services for the Real World.*

[C. 96] C. Perkins. IP Mobility Support. RFC 2002, Internet Engineering Task Force, 1996.

[Cox12] C. Cox. *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*. John Wiley & Sons, 2nd edition, 2012.

[CR01] R. G. Cole and J. H. Rosenbluth. Voice over IP performance monitoring. *SIGCOMM Comput. Commun. Rev.*, 31(2):9–24, April 2001.

[CS92] F. Chance and L. W. Schruben. Estimating A Truncation Point In Simulation Output. *School of Operations Research and Industrial Engineering, Cornell University, New York*, 1992.

[Dah07] E. Dahlman. *3G Evolution: HSPA And LTE For Mobile Broadband*. Electronics & Electrical. Elsevier Academic Press, 2007.

[DG03] L. Ding and R.A. Goubran. Assessment of Effects of Packet Loss on Speech Quality in VoIP. In *The 2nd IEEE Internatioal Workshop on Haptic, Audio and Visual Environments and Their Applications*, pages 49–54, September 2003.

[DPS11] E. Dahlman, S. Parkvall, and J. Skold. *4G LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[EM02] B. Ethan and A. Mark. On Making TCP More Robust to Packet Reordering. *SIGCOMM Comput. Commun. Rev.*, 32(1):20–30, January 2002.

[Env13] OMNeT++ Discrete Event Simulation Environment. `http://www.omnetpp.org/`, last accessed in May 2013.

[F. 10] F. Markus, H. Tobias, and T. Phuoc. A Generic Quantitative Relationship Between Quality of Experience and Quality of Service. *Netwrk. Mag. of Global Internetwkg.*, 24:36–41, March 2010.

[FFML12] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis. Locator/ID Separation Protocol (LISP). DRAFT 23, Internet Engineering Task Force, 2012.

[Fis72] G. S. Fishman. Bias Considerations in Simulation Experiments. *Operations Research*, 20:785–790, 1972.

[FJ93] S. Floyd and V. Jacobson. Random Early Detection gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, August 1993.

[FMMP00] S. Floyd, J. Mahdavi, M. Mathis, and M. Podolsky. An Extension To The Selective Acknowledgement (SACK) Option For TCP. RFC 2883, Internet Engineering Task Force, 2000.

[For13] FortMP Official Web-Page. `http://www.optirisk-systems.com/products_fortmp.asp`, last accessed in May 2013.

[F.R04] F.R. Farrokhi, M. Olfat, M. Alasti, and K.J.R. Liu. Scheduling Algorithms for Quality of Service Aware OFDMA Wireless Systems. In *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, volume 4, pages 2689–2693, November 2004.

[FRHB12] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure. TCP Extensions for Multipath Operation with Multiple Addresses. DRAFT 09, Internet Engineering Task Force, 2012.

[FY02] W. Feng and Z. Yongguang. Improving TCP Performance Over Mobile Adhoc Networks with Out-of-order Detection and Response. In *Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing*, MobiHoc '02, pages 217–225, New York, NY, USA, 2002. ACM.

[G. 08] G. Mongha, K.I. Pedersen, I.Z. Kovacs, and P.E. Mogensen. QoS Oriented Time and Frequency Domain Packet Schedulers for The UTRAN Long Term Evolution. In *IEEE Vehicular Technology Conference*, pages 2532–2536, May 2008.

[G. 10] G. Tsirtsis, H. Soliman, N. Montavont, G. Giaretta, N. Montavont, and K. Kuladinithi. Flow Bindings In Mobile IPv6 And NEMO Basic Support. RFC 6089, Internet Engineering Task Force, 2010.

[GAM78] A. V. Gafarian, C. J. Ancker, and F. Morisaku. Evaluation of Commonly Used Rules for Detecting Steady-state in Computer Simulation. *Naval Research Logistics Quarterky*, 25:296–310, 1978.

[Gar07] V.K. Garg. *Wireless Communications and Networking*. Elsevier Morgan Kaufmann, 2007.

[Gas05] M. S. Gast. *802.11 Wireless Networks, The Definitive Guide*. O'Reilly Media Inc. CA 95472, 2nd edition, 2005.

[GLD$^+$08] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil. Proxy Mobile IPv6. RFC 5213, Internet Engineering Task Force, 2008.

[GLP13] GLPK (GNU Linear Programming Kit). `http://www.gnu.org/software/glpk`, last accessed in May 2013.

[Goo05] J. Goodchild. *IP Video Implementation And Planning Guide: Integrating Data Voice And Video - Part II*. United States Telecom Association, 2005.

[GR10] M.N. Garcia and A. Raake. Parametric Packet-Layer Video Quality Model For IPTV. *10th Internation Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, pages 349 – 352, May 2010.

[GUR13] GUROBI Optimizer version 5.0. `http://www.gurobi.com/`, last accessed in May 2013.

[H. 05] H. Zhai and X. Chen, and Y. Fang. How Well Can The IEEE 802.11 Wireless LAN Support Quality of Service? *IEEE Transactions on Wireless Communications*, 4(6):3084–3094, November. 2005.

[H. 08] H. J. Kim, D. H. Lee, J. M. Lee, K. H. Lee, W. Lyu, and S. G. Choi. The QoE Evaluation Method through the QoS-QoE Correlation Model, 2008.

[H. 09] H. Soliman. Mobile IPv6 Support for Dual Stack Hosts and Routers. RFC 5555, Internet Engineering Task Force, 2009.

[HT04] H. Holma and A. Toskala. *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. John Wiley & Sons, 3rd edition, 2004.

[HWG$^+$12] T. Hu, B. Wenning, C. Görg, U. Toseef, and Z. Guo. Statistical Analysis of Contact Patterns between Human-carried Mobile Devices. In *4th International Conference on Mobile Networks and Management*, Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Brussels, Belgium, September 2012. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[IBM13a] IBM ILOG. `http://www.ilog.com`, last accessed in May 2013.

[IBM13b] IBM ILOG CPLEX Optimization Studio. `http://www.ibm.com/software/websphere/products/optimization/`, last accessed in May 2013.

[Ins12] SANS Institute. Latency and QoS For Voice Over IP. *Whitepaper*, October 2012.

[IT88] ITU-T. Pulse Code Modulation (PCM) Of Voice Frequencies. Recommendation G.711, International Telecommunication Union, November 1988.

[IT98] ITU-T. Methods For Subjective Determination Of Transmission Quality. Recommendation P.800, International Telecommunication Union, June 1998.

[IT01] ITU-T. Perceptual Evaluation Of Speech Quality (PESQ): An Objective Method For End-To-End Speech Quality Assessment Of Narrow-Band Telephone Networks And Speech Codecs. Recommendation P.862, International Telecommunication Union, February 2001.

[IT03a] ITU-T. One-Way Transmission Time. Recommendation G.109, International Telecommunication Union, May 2003.

[IT03b] ITU-T. Wideband Coding Of Speech At Around 16 kbit/s Using Adaptive Multi-Rate Wideband (AMR-WB). Recommendation G.722.2, International Telecommunication Union, July 2003.

[IT07] ITU-T. Transmission Impairments Due To Speech Processing. Recommendation G.113, International Telecommunication Union, November 2007.

[IT08] ITU-T. Objective Perceptual Multimedia Video Quality Measurement In The Presence Of A Full Reference. Recommendation J.247, International Telecommunication Union, August 2008.

[IT09] ITU-T. The E-model: A Computational Model For Use In Transmission Planning. Recommendation G.107, International Telecommunication Union, 2009.

[IT12] ITU-T. Coding Of Speech At 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP). Recommendation G.729, International Telecommunication Union, June 2012.

[J. 05] J. Kang, Y. Zhang, and B. Nath. Accurate And Energy-efficient Congestion Level Measurement in Ad-hoc Networks. In *IEEE Wireless Communications and Networking Conference*, volume 4, pages 2258 – 2263 Vol. 4, Aarch 2005.

[J. 09] J. Greengrass, J. Evans, and A.C. Begen. Not All Packets Are Equal, Part 2: The Impact of Network Packet Loss on Video Quality. *IEEE Internet Computing Journal*, 13(2):74–82, March 2009.

[Kar84] N. Karmarkar. A New Polynomial-time Algorithm For Linear Programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, New York, NY, USA, 1984.

[KH06] S. Khirman and P. Henriksen. Relationship Between Quality-of-Service And Quality-of-Experience For Public Internet Service. *PAM*, 24, 2006.

[Kle75] L. Kleinrock. *Queueing Systems: Volume I - Theory*. Wiley, New York, ISBN 978-0471491101, January 1975.

[KNI13] KNITRO Optimization Software. `http://www.ziena.com/`, last accessed in May 2013.

[KS05] S. Kent and K. Seo. Security Architecture For The Internet Protocol. RFC 6040, Internet Engineering Task Force, 2005.

[KT11] M.A. Khan and U. Toseef. User Utility Function as Quality of Experience (QoE). In *The Tenth International Conference on Networks*, pages 99–104, January 2011.

[Law83] A. M. Law. Statistical Analysis of Simulation Output Data. *Operations Research*, 31(6):983–1029, 1983.

[LBD$^+$11] X. Li, W. Bigos, D. Dulas, Y. Chen, U. Toseef, C. Görg, A. Timm-Giel, and A. Klug. Dimensioning of the LTE Access Network for the Transport Network Delay QoS. In *IEEE 73rd Vehicular Technology Conference (VTC Spring)*, pages 1–7, May 2011.

[LG02] M. Laor and L. Gendel. The Effect of Packet Reordering in a Backbone Link on Application Throughput. *IEEE Network Journal*, 16(5):28–36, September 2002.

[LK91] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, Inc., 2nd edition, 1991.

[LK00] R. Ludwig and R. H. Katz. The Eifel Algorithm: Making TCP Robust Against Spurious Retransmissions. *ACM Computer Communication Review*, 2000.

[LLT$^+$12] X. Li, M. Li, U. Toseef, A. Timm-Giel, C. Goerg, D. Dulas, M. Nowacki, and R. Ruchala. Dimensioning Of The Shared Transport Network For Collocated Multiradio: LTE and HSDPA. In *IEEE 8th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 308–315, October 2012.

[lRG13] lp_solve version 5.5 lp_solve Reference Guide. `http://lpsolve.sourceforge.net/`, last accessed in May 2013.

[LS03] S. Lincke-Salecket. Load Shared Integrated Networks. In *5th European Personal Mobile Communications Conference*, pages 225–229, April 2003.

[LTB$^+$11] X. Li, U. Toseef, W. Bigos, D. Dulas, C. Görg, A. Timm-Giel, and A. Klug. Dimensioning of The LTE Access Network. *Telecommunication Systems*, pages 1–18, 2011.

[LTL$^+$11] M. Li, U. Toseef, X. Li, A. Balazs, A. Timm-Giel, and C. Görg. Investigating the Impacts of IP Transport Impairments on Data Services in LTE Networks. In *1st European Teletraffic Seminar (ETS)*, February 2011.

[LTW$^+$10a] X. Li, U. Toseef, T. Weerawardane, W. Bigos, D. Dulas, C. Görg, A. Timm-Giel, and A. Klug. Dimensioning of The LTE Access Transport Network For Elastic Internet Traffic. In *IEEE 6th International Conference Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 346–354, October 2010.

[LTW$^+$10b] X. Li, U. Toseef, T. Weerawardane, W. Bigos, D. Dulas, C. Görg, A. Timm-Giel, and A. Klug. Dimensioning of The LTE S1 Interface. In *Third Joint IFIP Wireless and Mobile Networking Conference (WMNC)*, pages 1–6, October 2010.

[LWMV09] B. Lewcio, M. Wältermann, S. Möller, and P. Vidales. E-Model Supported Switching Between Narrowband and Wideband Speech Quality. In *Proceedings of First International Workshop on Quality of Multimedia Experience*, San Diego, CA, United States, 2009.

[LY12] Z. Lu and H. Yang. *Unlocking the Power of OPNET Modeler*. Cambridge University Press, New York, ISBN 978-0521198745, 2012.

[M. 97] M. Matthew, S. Jeffrey, M. Jamshid, and O. Teunis. The Macroscopic Behavior Of The TCP Congestion Avoidance Algorithm. *SIGCOMM Comput. Commun. Rev.*, 27(3):67–82, July 1997.

[M. 03] M. Zhang, B. Karp, S. Floyd, and L. Peterson. RR-TCP: A Reordering-Robust TCP with DSACK. In *Eleventh IEEE International Conference on Networking Protocols*, pages 95–106, Atlanta, GA, USA, 2003.

[M. 10a] M. A. Khan, U. Toseef, S. Marx, and C. Görg. Auction Based Interface Selection with Media Independent Handover Services And Flow Management. In *European Wireless Conference (EW)*, pages 429–436, April 2010.

[M. 10b] M. A. Khan, U. Toseef, S. Marx, and C. Görg. Game-Theory Based User Centric Network Selection with Media Independent Handover Services and Flow Management. In *8th Communication Networks and Services Research Conference (CNSR)*, pages 248–255, May 2010.

[M.A10] M.A. Mehmood, B. Lewcio, P. Vidales, A. Feldmann, and S. Moeller. Understanding Signal-Based Speech Quality Prediction in Future Mobile Communications, 2010.

[Meh92] S. Mehrotra. On the Implementation of a Primal-Dual Interior Point Method. *SIAM Journal on Optimization*, 2(4):575–601, 1992.

[MIN13] MINOS Solver version 5.5. `http://www.sbsi-sol-optimize.com/asp/sol_products_minos_desc.htm`, last accessed in May 2013.

[MN06] R. Moskowitz and P. Nikander. Host Identity Protocol (HIP) Architecture. RFC 4423, Internet Engineering Task Force, 2006.

[MR02] S. Mohamed and G. Rubino. A Study of Real-time Packet Video Quality Using Random Neural Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1071–1083, December 2002.

[MRK$^+$06] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann. Impairment Factor Framework for Wideband Speech Codecs. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1969–1976, 2006.

[NB09] E. Nordmark and M. Bagnulo. Shim6: Level 3 Multihoming Shim Protocol For IPv6. RFC 5533, Internet Engineering Task Force, 2009.

[NCG13] White paper NMC Consulting Group. LTE X2 Handover
`http://www.nmcgroups.com/en/expertise/lte/x2handover.asp`,
last accessed in May 2013.

[Net03] Net Predic, White paper. *Performance Analysis For Video Stream Across Networks*, 2003.

[Net13] NetSim. `http://tetcos.com/`, last accessed in May 2013.

[NH06] T. Nguyen and Y. Han. A Proportional Fairness Algorithm with QoS Provision in Downlink OFDMA Systems. *Communications Letters, IEEE*, 10(11):760–762, November 2006.

[NS-13] NS-The Network Simulator. `http://www.nsnam.org/`, last accessed in May 2013.

[Ohm99] J. R. Ohm. Bildsignalverarbeitung Fuer Multimedia-Systeme. In *Proc. of Skript*, 1999.

[OPL13] OPL Development Studio. `https://www.ibm.com/software/integration/optimization/opl-dev-studio/`, last accessed in May 2013.

[OPN13] OPNET Modeler. `http://www.opnet.com/`, last accessed in May 2013.

[ORCTV] E. Oki, R. Rojas-Cessa, M. Tatipamula, and C. Vogt. *Internet Protocols, Services, and Applications*.

[P. 98] P. Jitendra, F. Victor, T. Don, and K. Jim. Modeling TCP Throughput: A simple Model and its Empirical Validation. *SIGCOMM Comput. Commun. Rev.*, 28(4):303–314, October 1998.

[P. 02] P. Chatzimisios, V. Vitsas, and A.C. Boucouvalas. Throughput And Delay Analysis of IEEE 802.11 Protocol. In *IEEE 5th International Workshop on Networked Appliances*, pages 168–174, October 2002.

[P. 08] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio. Dynamic Packet Scheduling Performance in UTRA Long Term Evolution Downlink. In *3rd International Symposium on Wireless Pervasive Computing*, pages 308–313, May 2008.

[PJA04] C. Perkins, D. Johnson, and J. Arkko. Mobility Support in IPv6. RFC 3775, Internet Engineering Task Force, 2004.

[Pla13] VLC Media Player. `http://www.videolan.org/`, last accessed in May 2013.

[R. 07] R. Stewart. Stream Control Transmission Protocol. RFC 4960, Internet Engineering Task Force, 2007.

[RBHH01] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual Evaluation Of Speech Quality (PESQ)-A New Method For Speech Quality Assessment Of Telephone Networks And Codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752, 2001.

[S. 03] S. Bohacek, J.P. Hespanha, J. Lee, C. Lim, and K. Obraczka. TCP-PR: TCP For Persistent Packet Reordering. In *23rd International Conference on Distributed Computing Systems*, pages 222–231, May 2003.

[S. 10] S. Ickin, K. De Vogeleer, M. Fiedler, and D. Erman. The Effects of Packet Delay Variation on the Perceptual Quality of Video. In *IEEE 35th Conference on Local Computer Networks (LCN)* , pages 663–668, October 2010.

[S. 12] S. N. K. Marwat, T. Weerawardane, Y. Zaki, C. Görg, and A. Timm-Giel. Design and Performance Analysis of Bandwidth and QoS Aware LTE Uplink Scheduler in Heterogeneous Traffic Environment. In *8th International Wireless Communications and Mobile Computing Conference*, August 2012.

[SA97] T. J. Schriber and R. W. Andrews. Interactive Analysis of Simulation Output by the Method of Batch Means. In *Proceedings of Winter Simulation Conference*, pages 513–525, Piscataway, N. J., 1997.

[Sch03] J. Schiller. *Mobile Communications*. Addison-Wesley Publishing Company, 2nd edition, 2003.

[Ser13] Architectural Concepts Of Connectivity Services. `http://www.sail-project.eu/wp-content/uploads/2011/08/SAIL_D.C.1_v1.0_Final_PUBLIC.pdf`, last accessed in May 2013.

[SH04] T. Szigeti and C. Hattingh. *End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs*. Cisco Press, 1st edition, 2004.

[SJ05] Q. Song and A. Jamalipour. Network Selection in an Integrated Wireless LAN and UMTS Environment Using Mathematical Modeling And Computing Techniques. *IEEE Wireless Communications*, 12(3):42–48, June 2005.

[SJB$^+$04] A. S.Patrick, S. Janice, C. Brian, N. Sylvie, E. Khatib, E. Bruno, Z. Todd, and M. Stephen. A QoE Sensitive Architecture for Advanced Collaborative Environments. In *Proceedings of the First International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks*, QSHINE '04, pages 319–322, Washington, DC, USA, 2004. IEEE Computer Society.

[SJZS06] W. Song, H. Jiang, W. Zhuang, and A. Saleh. Call Admission Control for Integrated Voice/Data Services in Cellular/WLAN Interworking. In *IEEE International Conference on Communications*, volume 12, pages 5480–5485, June 2006.

[SLC06] D. Soldani, M. Li, and R. Cuny. *QoS And QoE Management In UMTS Cellular Systems*. John Wiley and Sons, 2006.

[Sp13] Scalable and Adaptive Internet Solutions (SAIL) project. Official website `http://www.sail-project.eu/`, last accessed in May 2013.

[Sta13] StatSoft.
`http://www.statsoft.com/textbook/distribution-tables/`, last accessed in May 2013.

[STK$^+$06] S. Shalunov, B. Teitelbaum, A. Karp, J. Boote, and M. Zekauskas. A One-way Active Measurement Protocol (OWAMP). RFC 4656, Internet Engineering Task Force, 2006.

[Str13] Strategy Analytics. `http://www.strategyanalytics.com`, last accessed in May 2013.

[SU10] P. Stefan and T. Uhl. Adjustments for QoS of VoIP in the E-Model. September 2010.

[SW94] W.R. Stevens and G.R. Wright. *TCP/IP Illustrated: The Protocols*, volume 1 of *Addison-Wesley Professional Computing Series*. Addison-Wesley Publishing Company, 1994.

[SW03] M. Siller and J. Woods. Improving Quality of Experience for Multimedia Services by QoS arbitration on QoE Framework. In *Proceedings of the 13th Packed Video Workshop*, 2003.

[SXZS12] Choon Shim, Liehue Xie, Bryan Zhang, and C.J. Sloane. How Delay and Packet Loss Impact Voice Quality in VoIP. *Whitepaper*, October 2012.

[SZ05] W. Song and W. Zhuang. QoS Provisioning via Admission Control in Cellular/Wireless LAN Interworking. In *2nd International Conference on Broadband Networks*, volume 1, pages 543–550, October 2005.

[SZC07] W. Song, W. Zhuang, and Y. Cheng. Load Balancing for Cellular/WLAN Integrated Networks. In *IEEE Networks*, volume 21, January 2007.

[TFG$^+$11] U. Toseef, C. Fan, C. Görg, U. Iqbal, and A. Udugama. Method For Establishing Of A Point-To-Point Connection Between A Mobile Node And A Network Entity, A Corresponding Mobile Node And A Corresponding Network Entity. Patent Application Number EP20080803967, 2011.

[TGF$^+$09] U. Toseef, C. Görg, C. Fan, A. Udugama, and F. Pittmann. Methods, Apparatuses, System, And Related Computer Program Product For Session Initiation. Patent Application Number PCT/EP2008/050967, 2009.

[TGF$^+$10] U. Toseef, C. Görg, C. Fan, A. Timm-Giel, and A. Udugama. Multiple Interface Support In Proxy Mobile IPv6. Patent Application Number EP20090004708, 2010.

[TGP$^+$09] U. Toseef, C. Görg, F. Pittmann, A. Udugama, and V. Pangboonyanon. Method For Network Controlled Mobile IP-Based Flow Management. Patent Application Number EP20080001071, 2009.

[TGPU09a] U. Toseef, C. Görg, F. Pittmann, and A. Udugama. Method And Device For Flow Management And Communication System Comprising Such Device. Patent Application Number EP20070013407, 2009.

[TGPU09b] U. Toseef, C. Görg, F. Pittmann, and A. Udugama. Network Mobility For Multi-Level Networks. Patent Application Number PCT/EP/2008/056669, 2009.

[TGSM11] G. Tsirtsis, G. Giaretta, H. Soliman, and N. Montavont. Traffic Selectors For Flow Bindings. RFC 6088, Internet Engineering Task Force, 2011.

[TGTG12] U. Toseef, C. Görg, and A. Timm-Giel. Optimized Flow Management using Linear Programming in Future Wireless Networks. In *The International Conference on Mobile Services, Resources, and Users*, 2012.

[TGU$^+$10] U. Toseef, C. Görg, A. Udugama, C. Fan, and F. Pittmann. Method Of And Device For Defining A Data Flow Description In A Network. Patent Application Number EP20090004708, 2010.

[The13] The General Algebraic Modeling System. `http://www.gams.com/`, last accessed in May 2013.

[TK12] U. Toseef and M. A. Khan. OPNET Simulation Setup for QoE Based Network Selection. In *Simulation in Computer Network Design and Modeling: Use and Analysis, ed. Hussein Al-Bahadili*, pages 100–139, 2012.

[TKGTG11] U. Toseef, M.A. Khan, C. Görg, and A. Timm-Giel. User Satisfaction Based Resource Allocation in Future Heterogeneous Wireless Networks. In *Ninth Annual Communication Networks and Services Research Conference (CNSR)*, pages 217–223, May 2011.

[TLL$^+$11] U. Toseef, M. Li, X. Li, A. Balazs, A. Timm-Giel, and C. Görg. Investigating the Impacts of IP Transport Impairments on VoIP Services in LTE Networks. In *16. ITG Fachtagung Mobilkommunikation*, Osnabrück, Germany, May 2011.

[TNJ07] S. Thomson, T. Narten, and T. Jinmei. IPv6 Stateless Address Autoconfiguration. RFC 4862, Internet Engineering Task Force, 2007.

[TWG$^+$13] U. Toseef, T. Weerawardane, R. Golderer, S. Hauth, C. Görg, and A. Timm-Giel. Coordinated LTE Uplink Radio Interface Scheduling. In *6th Joint IFIP Wireless and Mobile Networking Conference (WMNC)*, pages 1–8, April 2013.

[TZGTG12a] U. Toseef, Y. Zaki, C. Görg, and A. Timm-Giel. Development of Simulation Environment For Multi-homed Devices In Integrated 3GPP And non-3GPP Networks. In *10th ACM international symposium on Mobility management and wireless access*, MobiWac ’12, pages 29–36, New York, NY, USA, 2012. ACM.

[TZGTG12b] U. Toseef, Y. Zaki, C. Görg, and A. Timm-Giel. Uplink QoS Aware Multihoming in Integrated 3GPP and non-3GPP Future Networks. In *4th International Conference on Mobile Networks and Management*, Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Brussels, Belgium, September 2012. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[TZTGG12] U. Toseef, Y. Zaki, A. Timm-Giel, and C. Görg. Optimized Flow Management using Linear Programming in Integrated Heterogeneous Networks. In *ICSNC 2012, 7th International Conference on Systems and Networks Communications*, Red Hook, NY, USA, November 2012. IARIA, Curran Associates.

[TZZ$^+$12] U. Toseef, Y. Zaki, L. Zhao, A. Timm-Giel, and C. Görg. QoS Aware Multihoming in Integrated 3GPP and non-3GPP Future Networks. In *ICSNC 2012, The 7th International Conference on Systems and Networks Communications*, Red Hook, NY, USA, November 2012. IARIA, Curran Associates.

[U. 07a] U. Toseef, A. Udugama, C. Görg, P. Varaporn, and F. Pittmann. LINE: Link Information Normalization Environment. In *Proceedings of the 1st inter-*

national conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications, MOBILWARE '08, 2007.

[U. 07b] U. Toseef, A. Udugama, C. Görg, P. Varaporn, and F. Pittmann. Realization of Multiple Access Interface Management and Flow Mobility in IPv6. In *Proceedings of the 1st international conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications*, MOBILWARE '08, 2007.

[U. 11a] U. Toseef, T. Weerawardane, A. Timm-Giel, and C. Görg. LTE System Performance Optimization by Discard Timer Based PDCP Buffer Management. In *High Capacity Optical Networks and Enabling Technologies (HONET)*, pages 116–121, December 2011.

[U. 11b] U. Toseef, T. Weerawardane, A. Timm-Giel, and C. Görg. Performance Comparison Of PDCP Buffer Management Schemes In LTE System. In *Wireless Days (WD)*, October 2011.

[U. 12a] U. Toseef, T. Weerawardane, A. Timm-Giel, and C. Görg. LTE System Performance Optimization by RED Based PDCP Buffer Management. In *17. ITG Fachtagung Mobilkommunikation*, Osnabrück, Germany, May 2012.

[U. 12b] U. Toseef, T. Weerawardane, A. Timm-Giel, C. Görg, and H. Kröner. Adaptive Fair Radio Interface Scheduling for LTE Networks. In *High Capacity Optical Networks and Enabling Technologies (HONET)*, December 2012.

[Uhl08] T. Uhl. E-Model and PESQ in the VolP Environment: A Comparison Study. In *Proceedings of Polish-German Teletraffic Symposium, Berlin*, pages 207–216, 2008.

[UIT$^+$09] A. Udugama, M.U. Iqbal, U. Toseef, C. Görg, C. Fan, and M. Schlaeger. Evaluation of a Network Based Mobility Management Protocol: PMIPv6. In *IEEE 69th Vehicular Technology Conference*, pages 1–5, April 2009.

[VQE00] VQEG. Final Report From The Video Quality Experts Group On The Validation Of Objective Quality Metrics For Video Quality Assessment. `http://www.vqeg.org/`, 2000.

[W. 07] W. Song, H. Jiang, and W. Zhuang. Performance Analysis of the WLAN-First Scheme in Cellular/WLAN Interworking. *IEEE Transactions on Wireless Communications*, 6(5):1932–1952, may 2007.

[WB11] M. Wasserman and F. Baker. IPv6-to-IPv6 Network Prefix Translation. RFC 6296, Internet Engineering Task Force, 2011.

[WC06] J. Welch and J. Clark. A Proposed Media Delivery Index (MDI). RFC 4445, Internet Engineering Task Force, 2006.

[WDT$^+$09] R. Wakikawa, V. Devarapalli, G. Tsirtsis, T. Ernst, and K. Nagami. Multiple Care-of Addresses Registration. RFC 5648, Internet Engineering Task Force, 2009.

[Wee11] T. Weerawardane. *Optimization and Performance Analysis of High Speed Mobile Access Network*. Springer publications, 2011.

[Wel81] P. D. Welch. On the Problem of the Initial Transient in Steady-state Simulations. *IBM Watson Research Center, New York*, 1981.

[Win99] S. Winkler. A Perceptual Distortion Metric for Digital Color Video. In *Proc. of SPIE*, pages 175–184, 1999.

[X. 12] X. Li, O. Mehani, R. Agüero, R. Boreli, Y. Zaki, and U. Toseef. Evaluating User-centric Multihomed Flow Management for Mobile Devices in Simulated Heterogeneous Networks. In *4th International Conference on Mobile Networks and Management Monami 2012*, September 2012.

[X. 13] X. Li, O. Mehani, R. Agüero, U. Toseef, Y. Zaki, and C. Görg. Evaluating User-Centric Multihomed Flow Management in Multi-User Scenarios. In *14th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, June 2013.

[Y. 06] Y. Wang. *Survey Of Objective Video Quality Measurements, Technical Report WPICS-TR-06-02, EBU Technical Review,*, 2006.

[Y. 11] Y. Zaki, N. Zahariev , T. Weerawardane, C. Görg, and A. Timm-Giel. Optimized Service Aware LTE MAC Scheduler: Design, Implementation and Performance Evaluation. In *OPNET workshop 2011*, Washington D.C., USA, August 29-September 1 2011.

[YH08] K. Yamagishi and T. Hayashi. Parametric Packet-Layer Model for Monitoring Video Quality of IPTV Services. In *Communications, 2008. ICC '08. IEEE International Conference on*, pages 110–114, May 2008.

[YK07] F. Yu and V. Krishnamurthy. Optimal Joint Session Admission Control in Integrated WLAN and CDMA Cellular Networks with Vertical Handoff. *IEEE Transactions on Mobile Computing*, 6(1):126–139, January 2007.

[Zak12] Y. Zaki. *Future Mobile Communications: LTE Optimization and Mobile Network Virtualization*. Springer publications, 2012.

[ZZU$^+$11] L. Zhao, Y. Zaki, A. Udugama, U. Toseef, C. Görg, and A. Timm-Giel. Open Connectivity Services For Future Networks. In *8th International Conference Expo on Emerging Technologies for a Smarter World (CEWIT)*, pages 1–4, November 2011.