Optimized Resource Management using Linear Programming in Integrated Heterogeneous Networks

Umar Toseef *†, Yasir Zaki‡, Andreas Timm-Giel*, and Carmelita Görg†
*Institute of Communication Networks, Hamburg University of Technology, Hamburg, Germany
Email: {umar.toseef, timm-giel}@tuhh.de
†TZI ComNets, University of Bremen, Bremen, Germany, Email: {umr, cg}@comnets.uni-bremen.de
‡Computer Science Department, New York University, Abu Dhabi, UAE, Email: yz48@nyu.edu

***Abstract*—There have been tremendous advances over the past decades when it comes to wireless access technologies. Nowadays, several wireless access technologies are available everywhere. Even mobile devices have evolved to support multiple access technologies (e.g., 3G, 4G or WiFi) in providing the best possible access to the Internet. However, all of these devices can communicate using only one access technology at a time. It is foreseen that an integration of these access technologies to offer users a network access through multiple simultaneous connections would be beneficial for both end users and the mobile network operators. This paper investigates how to tackle the simultaneous usage of multiple wireless access technologies in the downlink. For this purpose, a practical example of heterogeneous network is considered where 3GPP LTE and non-3GPP WLAN access technologies are integrated together. Furthermore, a novel decision mechanism is proposed, that focuses on optimizing the network resource management based on a mathematical formulation of the system. The mathematical model is implemented using the Linear Programming techniques. The paper demonstrates the gains that are achieved from using such innovative decision mechanism as well as the benefits that arise from the simultaneous usage of multiple wireless heterogeneous accesses.**

Keywords*— *LTE and WLAN interworking, User QoE optimization, Linear programming, Access link modeling, Heuristic methods

I. Introduction

The Long Term Evolution (LTE) of the Universal Mobile Telecommunication System (UMTS) is one of the latest milestones achieved in an 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. On the other hand, the technology of handheld mobile devices has also made significant advancements in 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 the 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 better 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.

Fig. 1 shows how 3GPP and non-3GPP access technologies can be integrated to build heterogeneous networks according to the 3GPP standards [2]. This architecture called as, System Architecture Evolution (SAE), also allows mobile users to roam between the two access technology types. For this purpose, a seamless mobility is achieved by employing Proxy Mobile IPv6 (network based mobility) or Dual Stack Mobile IPv6 (host based mobility) [2]. The 3GPP SAE architecture, however, has certain limitations when it comes to supporting multi-homed users. This implies that a user can be associated with only one of the available access networks but cannot connect to more than one network simultaneously.

This work is an extension of [1] which investigates how the multihoming support can be realized in 3GPP SAE architecture. In addition, it discusses how the network operators can make an optimum use of the aggregated bandwidth resources and network diversity in a multihoming scenario through traffic flow management. The rest of the paper is organized

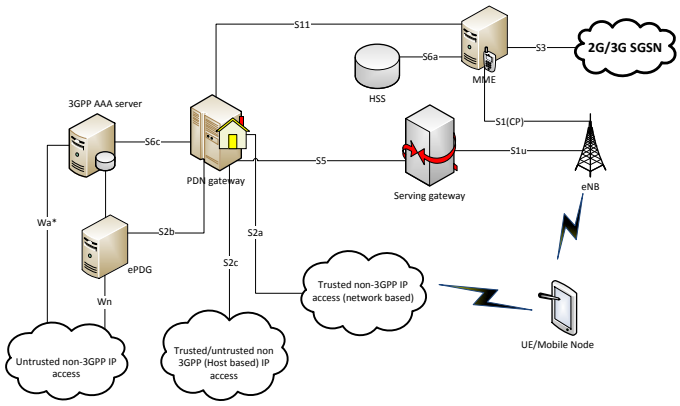


Fig. 1. 3GPP proposed SAE architecture for the integration of non-3GPP and 3GPP access technologies

as follows: related work has been discussed in Section II, Section III describes how the current 3GPP SAE architecture can be extended to provide user multihoming support. Section IV describes the importance of flow management function in a heterogeneous network, and Section V explains the Linear Programming technique to do optimized flow management operation. Section VI provides proof of concepts through the discussion of simulation results of the investigated realistic scenarios. Finally, in Section VII a heuristic based resource management algorithm is devised that offers near-optimum performance without high computational complexity.

II. Related Work

A number of research studies can be found in literature 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 [3] describe an intelligent scheme of vertical handover decisions in selecting the best handover target from several candidate heterogeneous networks. Several other proposals have been made to improve the performance of cellular and 802.11 networks. Song et. al. [4] has discussed admission control schemes to improve the performance of integrated networks. Fei and Vikram [5] propose a service differentiated admission control scheme based on semi-Markov chain that although very accurate but has high computational complexity. Similarly, Zhai et. al. [6] 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 a non-saturated state. Other studies have focused on developing solutions for load balancing in the integrated heterogeneous network environment. Such a proposal can be found in [8], where policy based load balancing framework has been presented to effectively utilize the aggregated resources of loosely coupled cellular/WLAN network. In this work, we explore the practical limits of the achievable performance in a heterogeneous network scenario by going down to the MAC layer functionalities of the involved access technologies. The goal is to maximize the spectral efficiency of the network bandwidth resources and fulfill the application QoS requirements at the same time. In contrast to other studies, we provide an analytical solution to the problem that adapts to the time varying channel conditions of the user and dynamically decides the best network paths for user traffic flows in order to achieve system wide optimized performance and improved user QoE. The focus of this work is, however, restricted to the downlink transmission of access technologies.

III. Network Simulation Model

This work follows the proposal of the 3GPP specifications in the integration of 3GPP access technology (namely, LTE) and trusted non-3GPP access technology (namely, legacy WLAN 802.11a) where host based mobility solutions, i.e., Dual Stack Mobile IPv6 is considered. For this purpose, a simulation network model has been implemented using the OPNET [9] network simulator. This includes the detailed implementation of LTE network entities following the 3GPP specifications. In this simulation model, home agent (HA) function is located at the Packet Data Network (PDN) gateway. The remote server acts as a correspondent node (CN) from where mobile users access application services like VoIP, video, HTTP and FTP (see Fig. 4).

OPNET model library implements 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 multi-homing, this basic support has been extended according to the IETF RFC for multiple care-of address (MCoA) registration [10]. This enables the user to register the care-of addresses from all of its active network interfaces with its home agent. This work assumes 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.

Though MCoA extension enables a user to register up to two care-of addresses with its home agent, 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 [11] that permits 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 [12]. Traffic selector helps 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. The 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. A comprehensive description of this heterogeneous network simulator can be found in [13].

It should be noted that our focus is only on the downlink access for LTE and WLAN. This implies that no uplink transmissions are performed for WLAN during the whole simulation time. Instead uplink traffic (e.g., TCP ACK packet etc.) is transmitted by the user through the LTE uplink access.

Fig. 2 presents the resource management architecture to be used in conjunction with the heterogeneous network simulator. This is an open and flexible architectural framework, where the resource management task is performed in 3 steps: (1) information collection; (2) decision making; (3) decision enforcement [14]. These steps are handled by three functional entities:

- Information Management Entity (IE) is in charge of gathering the information required by the decision making, e.g., link quality, power limitation, load and congestion of the networks. IE also pre-processes and filters the gathered information before it is delivered to the other entities.
- Decision Making Entity (DE) is the most intelligent part of this system architecture. It makes use of the 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. A decision making entity residing in the network is denoted with DE_n and that in the user terminal by DE_u.
- Execution and Enforcement Entity (EE) finally executes or performs the decision made by DE. In this work, DE_n entity takes all resource 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 that enforce them using the local EE entity.

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 at home agent as well as at user terminal.

IV. Flow Management

In the developed simulation environment, a user can communicate simultaneously through 3GPP access technology (i.e., LTE) as well as through non-3GPP access networks (i.e., WLAN). The question still remains how a network operator or a user can make an efficient use of the two network paths from two access technologies. For this purpose a flow management function is introduced at the home agent. In other words, it specifies how DE_n entity should function. The flow management function makes use of the MIPv6 extensions and allows controlling the user data rate on each network path. In general, there are two options of managing traffic flows for a multi-homed user. The first option is to carry one complete application traffic flow over one path of choice, this is known as "traffic flow switching". The second option is to divide the traffic flow into several smaller sub-flows where each sub-flow is carried over one network path. This will be called "traffic flow splitting".

If flow management is performed properly a considerable improvement in network capacity and user satisfaction can be achieved. That is why the decision engine (i.e., DE_n) of the flow management function, which controls the user data rate over the network paths is of core importance. In order to attain the goals of optimized network performance, the decision engine needs to know the precise information of the available network resources and the user demands. Once this information is available, DE_n with the help of the proposed mathematical techniques, can optimally assign network resources to the users fulfilling their demands while making use of all available network paths.

Each wireless access network usually has a fixed amount of network resources (e.g., spectrum bandwidth), and the network performance itself depends on the fact with what (spectral) efficiency these resources are utilized. In order to achieve higher data rates, a network resource management function should select those users more often who can attain high spectral efficiency. This work adopts the term "network path cost" to denote the required network resources per unit data rate. For a user, its network path cost can be accessed through cross layer information from the MAC layer of the corresponding access technologies. In the following subsections, it is shown how the network path cost can be computed for users in WLAN and LTE networks.

A. Network path cost for WLAN

Most modern wireless LAN access networks follow IEEE 802.11 standards, marketed under the name of Wi-Fi. In the infrastructure mode of 802.11 typically a number of stations are associated to an access point (AP) (which is normally a router) that serves as a bridge to a wired infrastructure network. 802.11 MAC uses one of the following three techniques to provide channel access control mechanisms.

1) Point coordination function (PCF): resides on the access point to coordinate the channel access for all associated stations through polling. A polled station can communicate with the access point in a contention free manner within a time slot. PCF is not part of the Wi-Fi Alliance interoperability and therefore is rarely found implemented on a portable device.
2) Distributed coordination function (DCF): is a random access scheme based on the Carrier Sense Multiple Access with Collision Avoidance protocol (CSMA/CA) with a binary exponential back-off algorithm. The DCF has two operating modes: the basic channel access mode and the RTS/CTS (Request-to-Send/Clear-To-Send) mode. DCF does not provide a contention free medium access and

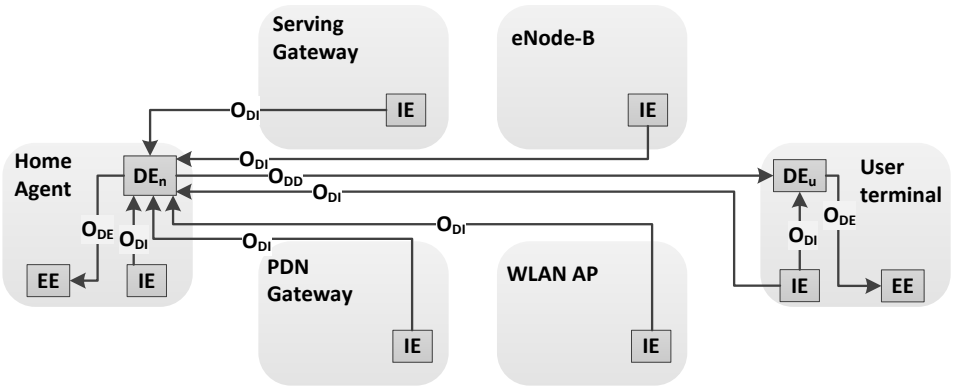


Fig. 2. Resource management architecture for the heterogenous network simulator.

therefore collisions can occur during the transmission if other stations also start transmitting at the same time. DCF is the de-facto default setting for Wi-Fi hardware.

3) Hybrid coordination function (HCF): has been designed to provide a differentiated medium access. Though HFC does not provide service guarantees, it establishes a probabilistic priority mechanism to allocate bandwidth based on traffic categories. HCF was introduced for the 802.11e standard, but it is hard to find complaint hardware. In recent time the 802.11n standard has incorporated HCF, which though becoming increasingly popular, is still available for a limited number of portable devices.

From the above description it can be deduced that a dominant percentage of today's portable Wi-Fi capable devices operate in the DCF mode of 802.11a/b/g. Three flavors of 802.11, i.e., a,b&g, follow very similar procedures in medium access mechanism therefore in this work we focus only on one of the flavors, i.e. 802.11a. The readers are encouraged to refer to [15]- [17] for further details on 802.11 specifications and performance.

Now, in order to explain the network path cost computations for WLAN, consider a scenario where a WLAN access network consists of a station associated with an access point. Assume that the station is just receiving a downlink traffic flow from the access point and does not transmit anything in the uplink. In this way, there is no contention for medium access. The transmission of one data frame with RTS/CTS enabled takes T_S seconds including the exchange of control frames such as RTS/CTS, SIFS (Short Interframe Space), DIFS (DCF Interframe Space), and ACK frames, where

$$T_S = T_{backoff} + T_{DIFS} + T_{RTS} + T_{CTS} + T_{data} + 3 \cdot T_{SIFS},$$

$$T_{backoff} = \frac{W_{min} - 1}{2} \cdot T_{slotTime},$$

$$T_{DIFS} = T_{SIFS} + 2 \cdot T_{slotTime}$$

All components of T_S except T_{data} can be found in the 802.11 standards (see Table II and I). The value of T_{data} can be computed based on the PHY data rate of transmission, i.e., $T_{data} = \frac{\sigma}{\varphi}$, where σ is the data frame size in kbit, and φ is the PHY transmission data rate in [kbit/sec]. Accordingly, the maximum downlink capacity η can be estimated as follows: $\eta = \frac{\sigma}{T_S}$ [kbit/sec].

TABLE I
MAC/PHYSICAL LAYER PARAMETERS OF 802.11A.

SIFS	SlotTime	RTS	CTS	ACK	W_{min}	W_{max}
16 μs	9 μs	160 bit	116 bit	116 bit	16	1024

TABLE II
DURATION OF CONTROL FRAMES IN 802.11A FOR DIFFERENT PHYSICAL LAYER DATA.

Data Rate (Mbps)	Modulation	Bits per Sym.	RTS		CTS/ACK	
			Sym.	μs	Sym.	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

It is clear that 802.11 MAC follows the Time Division Multiple Access (TDMA) like scheme, where 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 γ fraction of that interval to achieve a unitary data rate of 1 kbit/sec. This way, the γ that is expressed in units of $[\frac{sec}{kbit/sec}]$ represents the network path cost. Its value directly depends on T_S, which is the delay experienced in transmitting one data packet of average size σ [bit] operating at PHY data rate φ [kbit/sec]. That is, $\gamma = \frac{T_S}{\sigma}$.

B. Network path cost for LTE

In contrast to 802.11, LTE performs a managed scheduling of available bandwidth resources. The smallest unit of bandwidth resource is referred as a physical resource block (PRB) in the LTE specification. Based on the allocated frequency spectrum size, LTE has a certain number of PRBs. The LTE MAC scheduler residing at the eNodeB schedules these PRBs using a 1ms transmission time interval (TTI). The LTE MAC

scheduler has a very complex way of assigning resources to the associated users. Without digging into the details of the MAC scheduler operation, we focus on the last stage of the resource assignment procedure in a certain TTI. Upon reaching this stage, the MAC scheduler already builded up a list of users which will be transmitting/receiving data in that TTI. For each user entry in the list, there is a corresponding value of the allocated number of PRBs, as well as the channel dependent Modulation and Coding Scheme (MCS) index. The MCS index is then used to lookup the Transport Block Size Index (TBS index). With the help of TBS index and number of allocated PRBs, the Transport Block Size (TBS) is obtained from a table defined in the 3GPP specifications [21]. This is a two dimensional table where each row lists TBS sizes corresponding to the number of PRBs for a particular TBS index. The obtained TBS value defines the size of the MAC frame transmitted to the user in that TTI. In this way, the user received throughput at the MAC layer in a certain TTI can be estimated if the TBS value for that user is known.

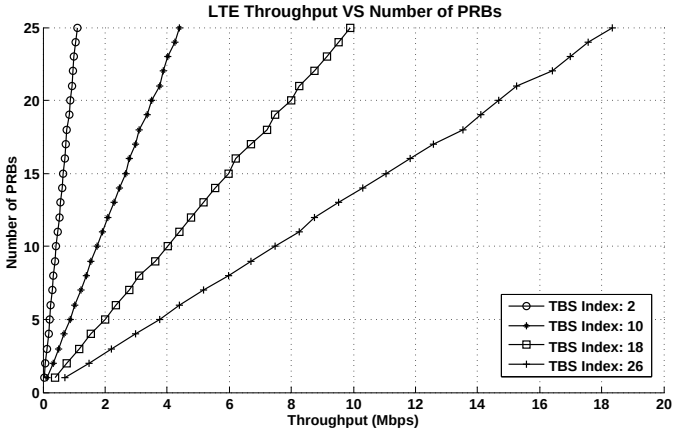


Fig. 3. Relationship of LTE air interface throughput and number of PRBs for different TBS index values [21]. Each curve represents one TBS index.

Fig. 3 shows that for a particular TBS index, the LTE throughput value has almost a linear relationship with the used number of PRBs. If described mathematically, this relationship can be used to determine the required number p of PRBs/TTI to achieve a certain data rate X [kbit/sec] for a user having TBS index i. That is

$$p = \alpha_i \cdot X + \beta_i$$

α_i is the slope of a straight line (as shown in 3) described in units of PRBs/kbps. β_i is the intercept of straight line at the y-axis and is expressed in units of number of PRBs. Both α and β together determine the network path cost of a user's LTE access link.

V. Optimized network resource utilization

When the network path costs for a multi-homed user are known, the problem of optimal resource utilization can be solved using mathematical techniques. In this work, we prefer Integer Linear Programming (ILP) to solve this problem. This choice has been made due to several reasons. For example, ILP guarantees an optimum solution for a correctly formulated problem, the problem formulated in ILP can be extended or restricted by introducing appropriate constraints as well as it saves additional implementation work by making use of already available Linear Programming solvers.

TABLE III
Mathematical model for the optimized resource utilization in algebraic form

Given

U	a set of users
α_j	Data rate dependent part of the LTE link cost in PRBs per kbps for user j, for each $j \in U$
β_j	Data rate independent part of the LTE link cost in PRBs for user j, for each $j \in U$
γ_j	Cost of WLAN link in seconds per kbps for user j, for each $j \in U$
δ_j	Minimum data rate (kbps) demand of a traffic flow destined to user j, for each $j \in U$
Δ_j	Maximum data rate (kbps) allocation for a traffic flow destined to user j, for each $j \in U$
Ω	Number of available PRBs for the LTE access network

Defined variables

X_j	Size of sub-flow in kbps sent over the LTE access link to user j, for each $j \in U$
Y_j	Size of sub-flow in kbps sent over WLAN access link to user j, for each $j \in U$
Z_j	Auxiliary binary variable; its value for a user j is either 1 if $X_j > 0$ or 0 otherwise, for each $j \in U$

Maximize

$\sum_{j \in U} X_j + Y_j$

Subject to

1. $\sum_{j \in U} \alpha_j \cdot X_j + \beta_j \cdot Z_j \leq \Omega$
2. $\sum_{j \in U} \gamma_j \cdot Y_j \leq 1$
3. $\delta_j \leq X_j + Y_j \leq \Delta_j$ for each $j \in U$
4. $Z_j \leq X_j \cdot 10^{20}$ for each $j \in U$
5. $Z_j \geq X_j / \Delta_j$ for each $j \in U$
6. $0 \leq X_j \leq \Delta_j$ for each $j \in U$
7. $0 \leq Y_j \leq \Delta_j$ for each $j \in U$
8. $Z_j \in \{0, 1\}$ for each $j \in U$

Table. III shows the formulation of the problem in algebraic form. The model defines U as the set of multi-homed users. Each element of this set has a number of input parameters, e.g., network path costs for LTE (α,β) and WLAN network (γ) according to the user channel conditions in the corresponding network. The maximum and minimum range of user data rate demands (δ, Δ) based on the individual user application. The amount of available network resources in LTE (Ω) and WLAN (which is 1 second) are also considered as input parameters. The output parameters for each user in set U include the assigned data rate over the LTE network and the WLAN network paths (X,Y). It is obvious that the goal of this model is to achieve the highest possible spectral efficiency from the two network access technologies. The higher the spectral efficiency, the higher the network throughput. Hence,

the objective is to maximize the user data rate over the two network paths, i.e., X and Y for every multi-homed user.

The model imposes eight constraints, which are listed at the bottom of Table III. The first two constraints ensure that the available network resources should not be exceeded when allocating the data rates for users. The third constraint dictates that the user data rate allocation should lie in the specified range. The 4th and 5th constraint determine the value of variable Z based on the X value. If there is a need, a user is allowed to receive its whole demanded data rate over a single network path as shown in constraint number 6 and 7. Constraint 8 is set in order to emphasis that Z is a binary variable, which has value either 0 or 1.

It is assumed here that each user is running only one application. For a constant bit rate application, e.g., VoIP or video the minimum data rate is set equal to the maximum data rate in the model input parameters. For TCP based flows, these two values can be set according to the network operator's policy. It should be noted that the problem has been formulated in a way that it guarantees the minimum data rate for all users and then assigns an additional data rate up to the maximum data rate while optimizing the spectral efficiency of the access networks.

In the investigated scenario, the LTE coverage is available in the whole area of user movement while WLAN coverage is limited in a circular area of 100 meter radius around a hotspot. This implies the users always have LTE access available and WLAN coverage is only found in the vicinity of the hotspot (see Fig. 4). During the resource assignment process, the flow management function classifies users into the following three categories (i) users with LTE access only and running VoIP or video applications (ii) users with LTE and WLAN access running any type of application (iii) users running FTP or HTTP applications with LTE access only. Users in the first category must be assigned the required minimum data rate through LTE as there is no other access available for them. Users in the second category are multi-homed users whose data rate will be decided by the aforementioned mathematical model. For users belonging to the third category, they must get their traffic through the LTE path; however, it is not clear how much data rate should be allocated to them in order to achieve the optimized resource allocation objective. This issue is resolved by using the following work around: the users are assigned a WLAN network path cost greater than unity and they are put into the second category. The WLAN network cost greater than unity will refrain the LP solver to assign any data rate for these users over the WLAN path while the data rate for the LTE path will be decided based on the global objective of the optimized resource allocation.

The resource assignment process by the flow management function is carried out periodically every 100ms[1] in order to adapt to any changes in the user channel conditions. For this purpose user channel condition parameters are obtained through cross layer information from the base stations of the two access technologies. With the help of these parameters, costs for each user network path is computed and fed to the above described mathematical model as the input arguments accompanied with the user data rate demands. As described earlier, the mathematical model is formulated using Linear Programming and solved using the C application programming interface (API) of ILOG CPLEX from IBM [22], which has been integrated inside the OPNET simulator by the authors. The output of this process consists of user data rates on each network path. These decided data rates are then implemented for each user through a traffic shaping function residing at the home agent.

[1]Section VI-A analyzes the selection of this time period

VI. Simulation Results

In this section, the performance of the proposed scheme for optimized resource allocation is evaluated with the help of a simulation scenario. Fig. 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 IV 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 [26], i.e., "3GPP-HO" and "Channel Aware". In the "3GPP-HO" approach user 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 Quality of Experience (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 model shown in Table III. 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 the EE entities.

Fig. 5 shows the performance of the FTP downlink application in terms 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 increasing 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 Fig. 6. It

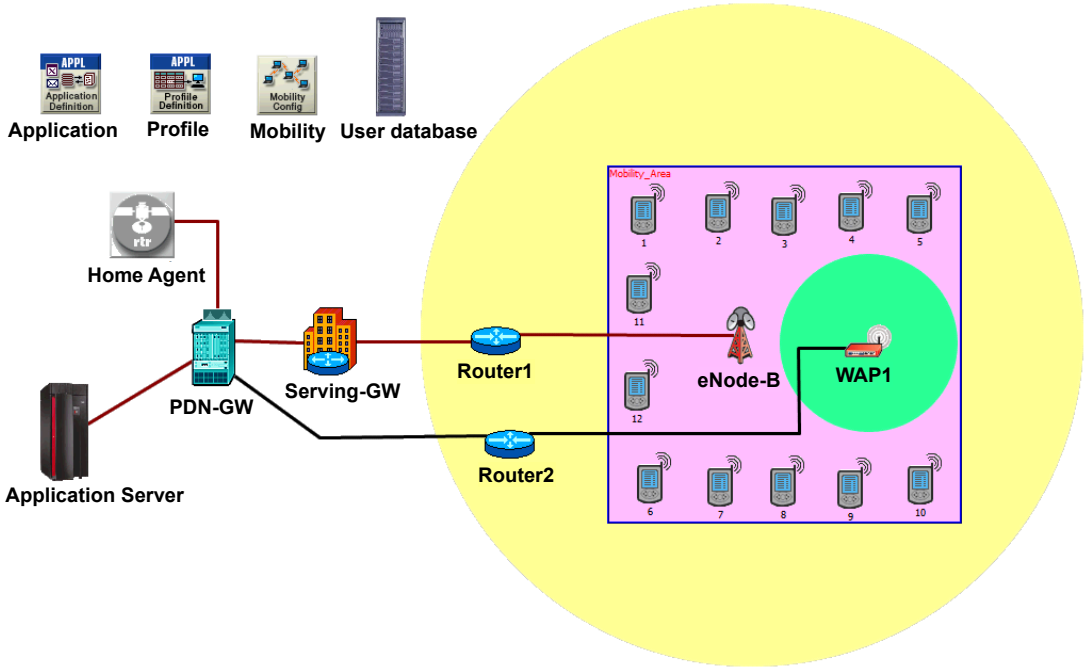


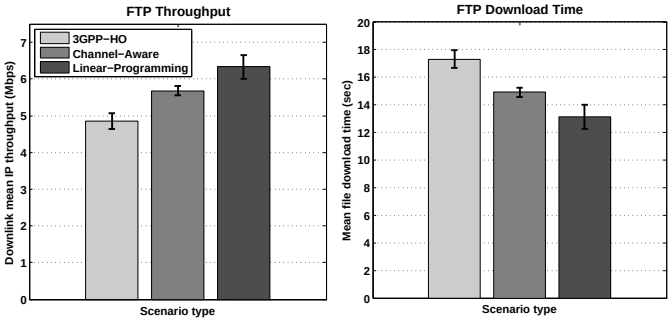
Fig. 4. Simulation scenario setup in the OPNET simulator. The large circular area shows the coverage of LTE and the smaller circular area represents the WLAN network coverage. The user movement is restricted to the rectangular area.

TABLE IV
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 [23], Correlated Slow Fading.
LTE MAC scheduler	Time domain: Optimized Service Aware [24], Frequency domain: Iterative RR approach [25]
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, 23.05 kbps data rate and 50frame/s
Skype video model	MPEG-4 codec, 512 kbps, 30frame/s, 640x480 resolution, play-out delay: 250ms
Streaming video model	MPEG-4 codec, 1 Mbps, 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 $[\delta, \Delta]$	[200 kbps, 25 Mbps] for FTP and HTTP users
Simulation run time	1000 seconds, 10 random seeds, 95% Confidence interval

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 pages. The download of these objects finishes before the TCP connection could achieve the maximum possible throughput. 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 the 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.

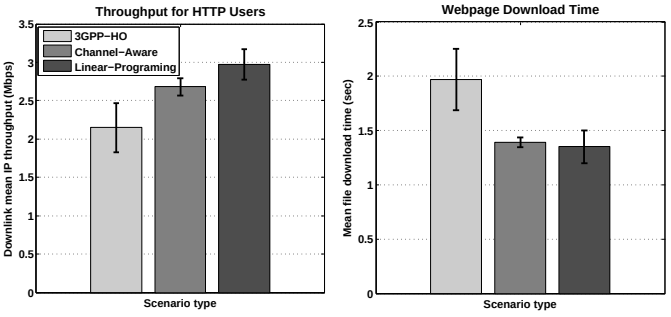
The performance gain of "Channel Aware" and "Linear Programming" over "3GPP-HO" approach can be attributed to manifold factors. For example, both of them are capable of aggregating bandwidth resources from multiple access links, they can accurately estimate the capacity of an access link and utilize it accordingly, they can periodically reevaluate their



(a) (b)

Fig. 5. FTP downlink performance comparison for "3GPP-HO", "Channel Aware", and "Linear Programming" approaches.

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 that 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 that 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 exact sufficient amount of data to the air interface schedulers that 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.



(a) (b)

Fig. 6. HTTP downlink performance comparison for "3GPP-HO", "Channel Aware", and "Linear Programming" approaches.

The user QoE for VoIP application has been depicted in Fig. 7. The Box plot shows the sample values of Mean Opinion Score (MOS)[2] [28] computed for a user employed wideband codec. A Box plot graphically depicts the groups of numerical data through their five 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.

It can be seen in Fig. 7 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.

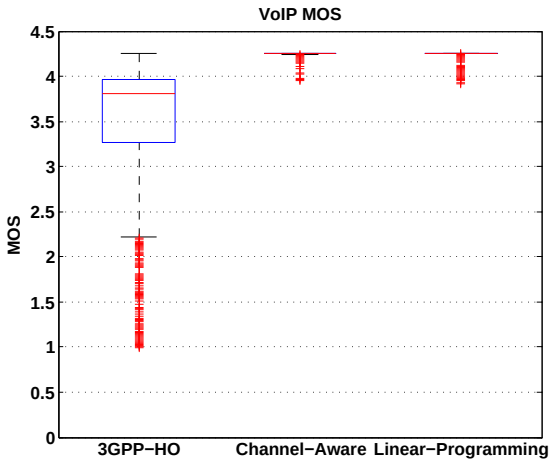
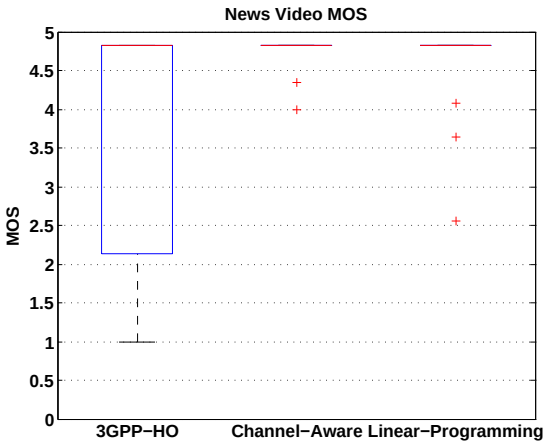


Fig. 7. Downlink performance comparison of VoIP application for "3GPP-HO", "Channel Aware", and "Linear Programming" approaches.

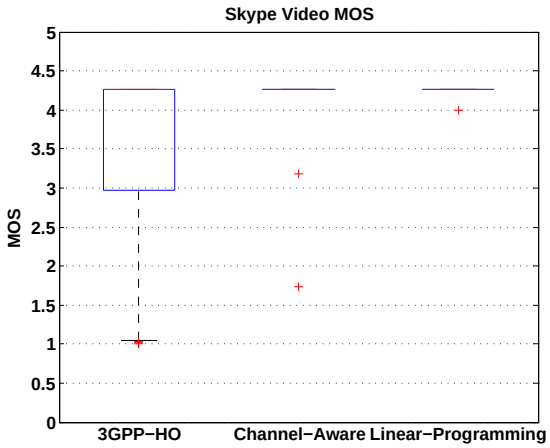
The "Linear Programming" approach also offers excellent user QoE for video applications. This can be confirmed by referring to Fig. 8, which shows the Box plot of user experienced MOS score for their video applications, i.e., video conference and video streaming. Almost all video quality evaluations 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 Fig. 9. Considering the

[2]MOS gives a numerical indication of the perceived user QoE of realtime applications. MOS is expressed on a scale from 1 to 5, 1 being the worst and 5 the best.

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 Fig. 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 that 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 higher that they have to encounter a congested network.



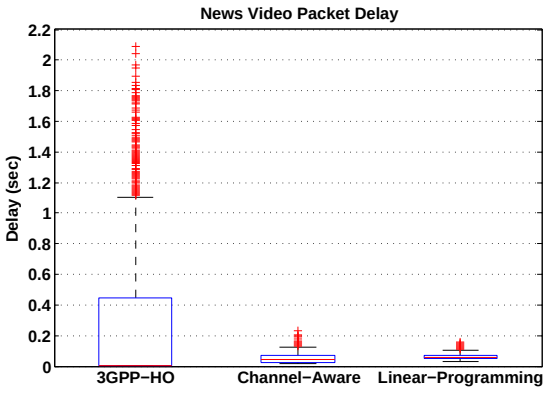
(a)



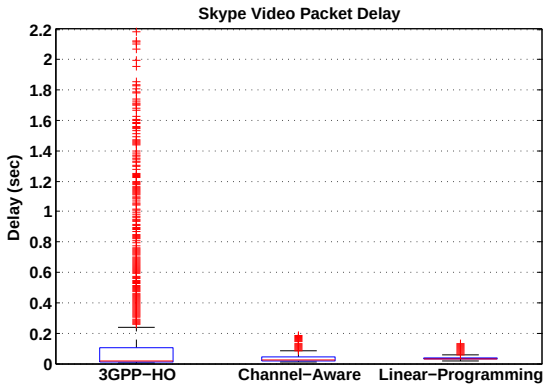
(b)

Fig. 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



(a)



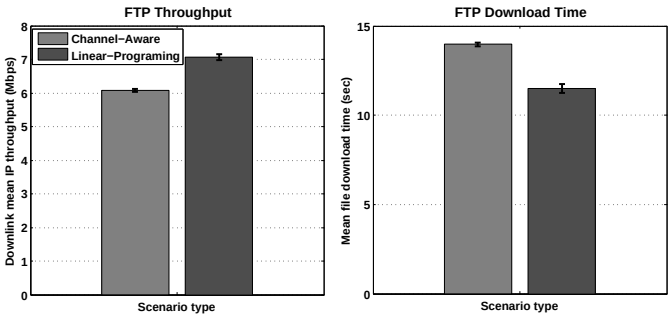
(b)

Fig. 9. Packet delay comparison of video applications for "3GPP-HO", "Channel Aware", and "Linear Programming" approaches.

have stringent QoS requirements, which have to be fulfilled at all costs in order to keep users satisfied. Therefore, 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 another scenario 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. Fig. 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 the 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 the '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)

Fig. 10. FTP downlink performance comparison between the "Channel Aware" and the "Linear Programming" approaches.

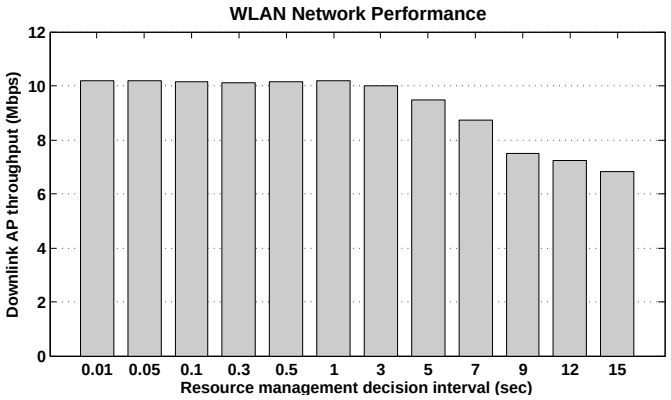


Fig. 11. Downlink throughput variations of WLAN access point for different values of the DE_n decision interval. The "Linear Programming" approach has been used for resource management decisions.

A. Sensitivity Analysis of DE_n Decision Intervals

It has been explained that the decision making entity (DE_n) of the flow management overlay architecture that 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 the 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 the aforementioned situation. As soon as the 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 the 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?

The above questions can be answered with the help of the simulation results shown in Figs. 11, 12, and 13. For this purpose the same simulation scenario, which has been discussed earlier in this section and whose configurations has been listed in Table IV is employed. In this simulation study, the DE_n decision interval is varied from 10ms to 15s and the "Linear Programming" approach is used to make the 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 Fig. 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 would 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 24 Mbps to 36 Mbps 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 36 Mbps to 24 Mbps during the decision interval. This will cause that user to experience large packet delays due the fact that some of the data are 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 Fig. 11 that increasing the DE_n decision interval to 15s causes approximately 30% degradation in the network capacity.

Fig. 12 shows a similar behavior for the LTE network, which may undergo cell throughput degradations due to the 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

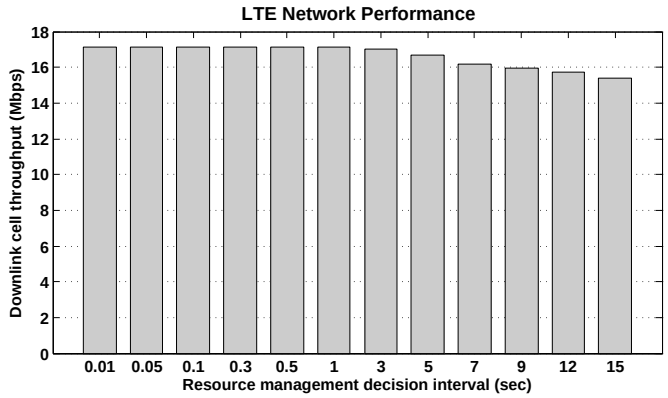


Fig. 12. LTE downlink cell throughput variations for different values of the DE_n decision interval. The "Linear Programming" approach has been used for resource management decisions.

has a smaller coverage area and the user 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 the resource management decisions.

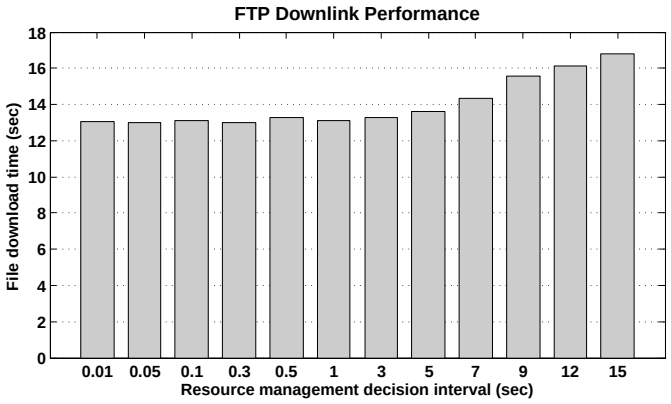


Fig. 13. Mean file download time experienced by FTP users. The figure shows how the FTP performance is affected by different values of DE_n decision interval. The "Linear Programming" approach has been used for resource management decisions.

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 Fig. 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.

VII. Heuristic Algorithms

The solution of the resource allocation problem obtained through mathematical modeling using Linear Programming provides an upper limit on the 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 high complexity of Linear Programming problem. The high complexity requires substantial computing power and time to solve these problems. This makes 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 it 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. Fig. 14 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 [3]. 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.

TABLE V
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

Before the development of the heuristic algorithm, an understanding of the resource allocation problem is developed

[3]Microsoft Windows Server 2008 R2 Enterprise 64bit, Intel©Xeon CPU @ 2.67GHz, 48GB RAM

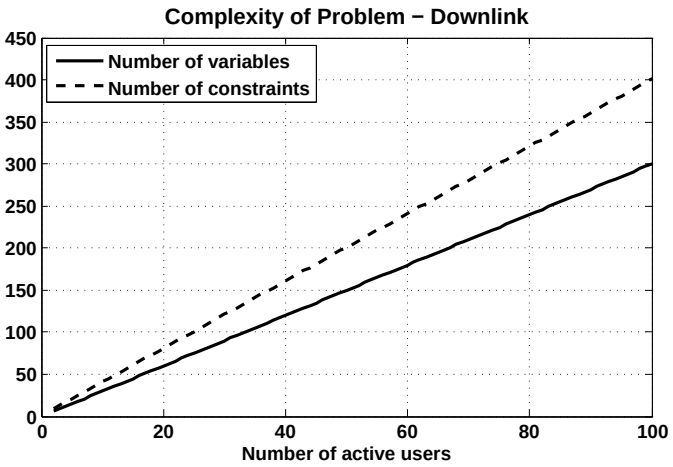


Fig. 14. The complexity of Linear Programming problem for downlink communication described in Table III.

with the help of a simple example presented in Table V. In this example, there are two multihomed users who require a fixed data rate of 1 Mbps 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 to 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.

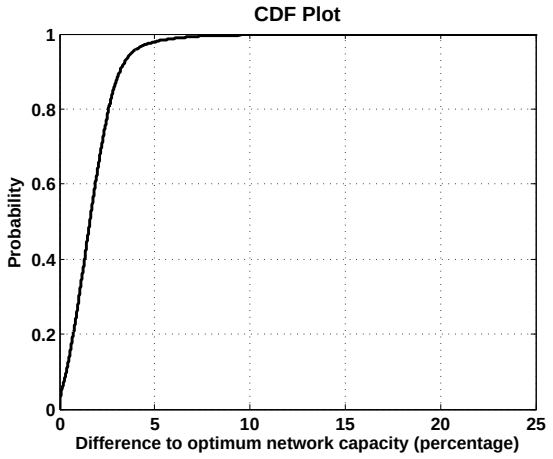
The most suitable strategy to allocate resources in such a situation is through the greedy approach. This implies that users should be served over that particular access link that costs less network resources. It can be noticed from the table that both users have less normalized cost for LTE access links 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 Fig. 16. 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.

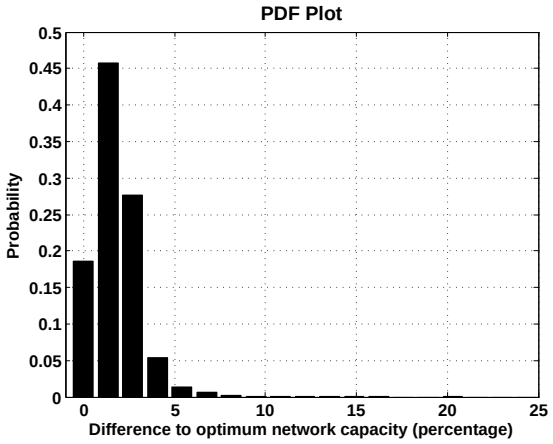
After fulfilling the minimum data rate demands of all users, the left over network resources should be assigned to the users whose maximum data rate demand is greater than their minimum data rate demand. Typically, they are the FTP/HTTP users. Though the same greedy approach can also be employed here 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 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 Fig. 16. A flow chart of subprocess (A) has been shown in Fig. 17. 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 Figs. 16 and 17, 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 that 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). Fig. 15 summarizes the outcome of this evaluation process. The figure shows a CDF and PDF curves of values representing how large the 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 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 VI. It is a slightly modified version of the example presented in Table V where the maximum data rate demands of users have been raised to 23.75 Mbps. 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., 1 Mbps) 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



(a)



(b)

Fig. 15. 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.

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 22.75 Mbps. This way UE1 is assigned with a total data rate of 23.75 Mbps considering also 1 Mbps 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 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 11.1 Mbps. This way, UE2 gets a total data rate allocation of 12.2 Mbps considering also 1 Mbps data rate allocation in the first step. Hence, the total network capacity amounts to 22.75+12.2=34.95 Mbps, in this case.

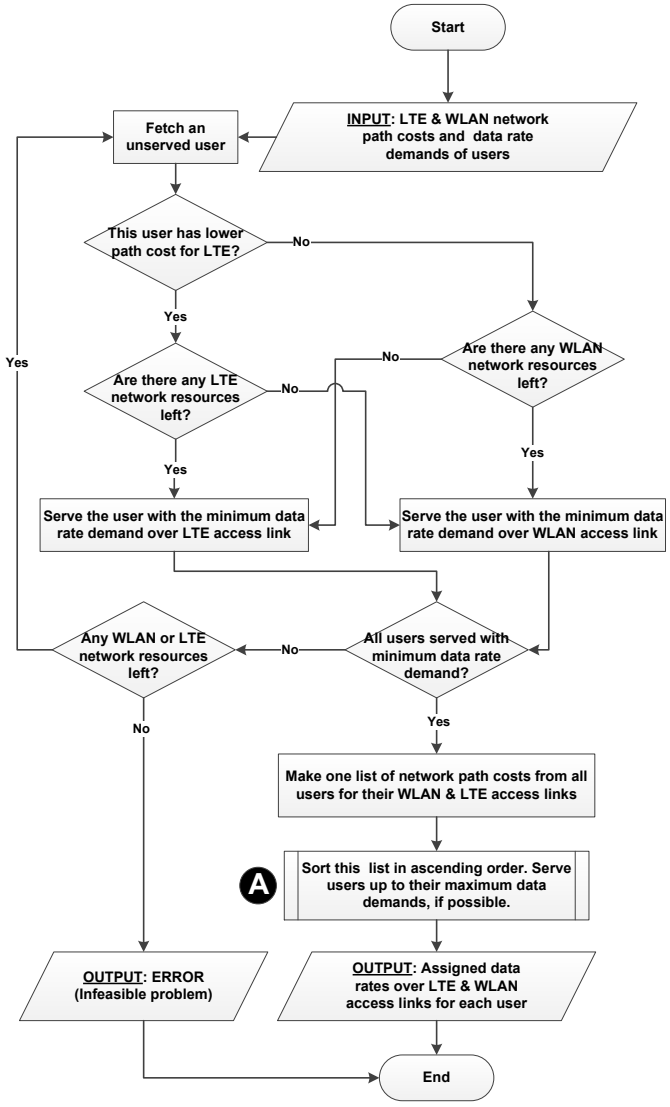


Fig. 16. Flow chart of the heuristic algorithm to solve the resource allocation problem in downlink communication.

TABLE VI
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

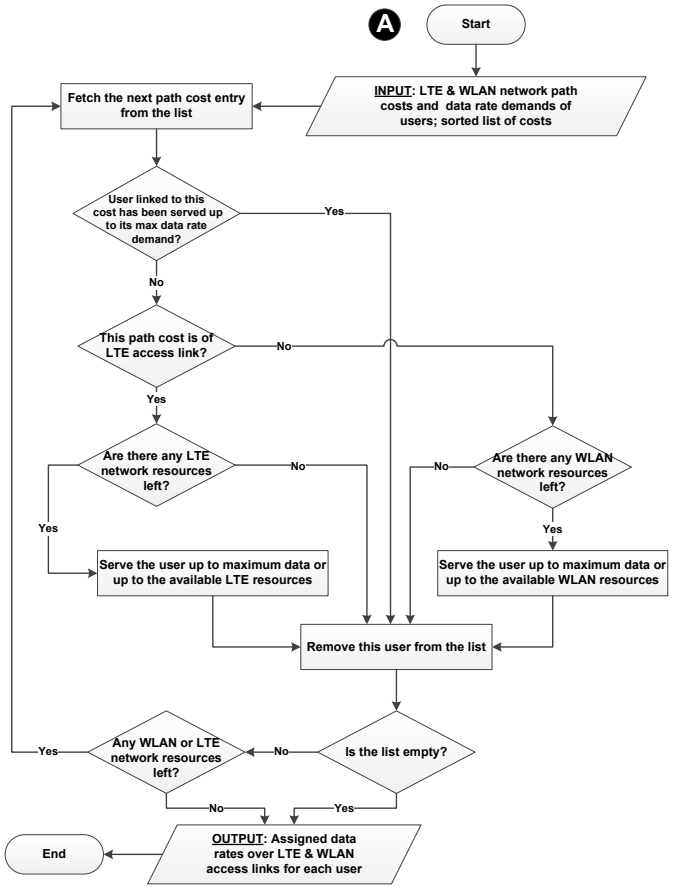
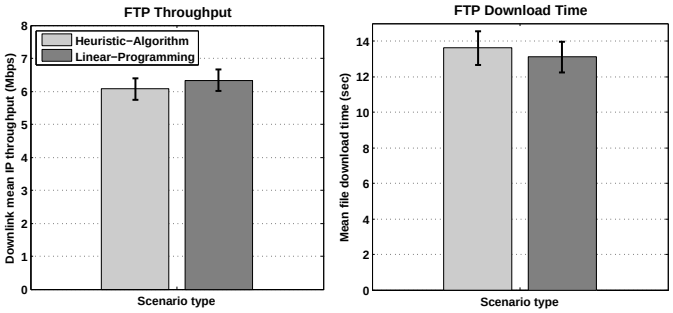


Fig. 17. Flow chart of the subprocess (A) in Fig. 16.

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 that has the highest path cost. This would be a bad move that 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 16.67 Mbps and UE1 over the LTE access network with data rate of 20 Mbps. This way, total network capacity amounts to 36.67 Mbps which is 4.9% higher than that attained by using the heuristic approach.

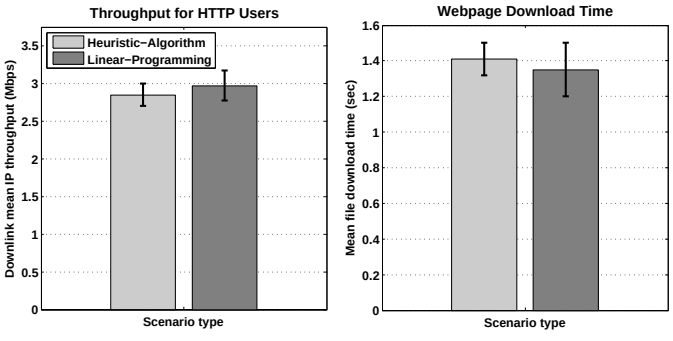
A sophisticated heuristic algorithm that mimics the "Linear Programming" approach in conceiving the effects of resource allocation of a user on the achievable spectral efficiency of the other users will be overly complex. This is because as the user count increases, each resource allocation will have 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.

The performance of the developed heuristic approach is evaluated using the simulation scenario discussed in Section VI. This way, the results of the heuristic approach can be compared with that of the "Linear Programming" approach. Fig. 18 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 performance 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)

Fig. 18. FTP downlink performance comparison for "Heuristic Algorithm" and "Linear Programming" approaches.



(a) (b)

Fig. 19. HTTP 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 always 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 the 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.

VIII. Conclusion

This work highlights the importance of multihoming support in the integrated heterogeneous wireless networks of 3GPP and non-3GPP access technologies. The existing 3GPP specifications for integration of two types of the access technologies are extended to realize multihoming support for the users. Following the proposed extensions, a network simulation model is developed, where LTE and WLAN co-exist. This work also focuses on the problem of optimum resource utilization in such heterogeneous networks, where the users and network operators can take advantage of multihoming support. The problem of optimum network resource allocation is mathematically modeled using the Linear Programming technique. The proof of concept is provided through the simulation results. With the help of simulation results it is shown that the proposed scheme of resource allocation brings twofold benefits when compared to the 3GPP proposal. On the one hand, it significantly improves the network capacity and on the other hand it fulfills the user application QoS demands, which otherwise cannot be satisfied from QoS unaware non-3GPP access technologies. In addition to the Linear Programming based solution, this work also proposes a heuristic based method for network resource management. This method not only exhibits less computational complexity but also accomplishes a performance gain close to that attained by mathematical optimization techniques. This makes it feasible for use in real world network equipment.

References

[1] U. Toseef, Y. Zaki, A. Timm-Giel, and C. Görg, Optimized Flow Management using Linear Programming in Integrated Heterogeneous Networks, The Seventh International Conference on Systems and Networks Communications, Lisbon, Portugal, November 2012.

[2] 3GPP Technical Report TS 23.402, Architecture enhancements for non-3GPP accesses, 3rd Generation Partnership Project, v10.6.0, December 2011.

[3] Q. Song and A. Jamalipour, Network Selection in an Integrated Wireless LAN and UMTS Environment using Mathematical Modeling and Computing Techniques, IEEE Wireless Commun., June 2005.

[4] W. Song, H. Jiang, and W. Zhuang, Performance analysis of the WLAN-first scheme in cellular/WLAN interworking, IEEE Trans. Wireless Commun., vol. 6, May 2007.

[5] F. Yu and V. Krishnamurthy, Optimal Joint Session Admission Control in Integrated WLAN and CDMA Cellular Networks with Vertical Handoff, IEEE Transaction on Mobile Computing, vol. 6, January 2007.

[6] H. Zhai, X. Chen, and Y. Fang, How Well Can the IEEE 802.11 Wireless LAN Support Quality of Service?, IEEE Trans. Wireless Commun., vol. 4, 2005.

[7] W. Song, H. Jiang, and W. Zhuang, "Call admission control for integrated voice/data services in cellular/WLAN interworking", IEEE ICC06, vol. 12, June 2006.

[8] S. Lincke-Salecket, Load shared integrated networks, Personal Mobile Communications Conference, 2003.

[9] OPNET website, http://www.opnet.com, as accessed in June 2013.

[10] R. Wakikawa, V. Devarapalli, G. Tsirtsis, T. Ernst, and K. Nagami, Multiple care-of addresses registration (RFC 5648), 2009.

[11] G. Tsirtsis, H. Soliman, G. Giaretta, and K. Kuladinithi, Flow bindings in mobile IPv6 and NEMO basic support (RFC 6089), 2010.

[12] G. Tsirtsis, G. Giaretta, H. Soliman, and N. Montavont, Traffic selectors for flow bindings (RFC 6088), 2011.

[13] U. Toseef, Y. Zaki, A. Timm-Giel, C. Görg., Development of Simulation Environment for Multi-homed Devices in Integrated 3GPP and non-3GPP Networks, The 10th MobiWAC conference, Paphos, 2012.

[14] SAIL consortium, D.C.1: Architectural Concepts of Connectivity Services, July 2011.

[15] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications ISO/IEC 8802-11:1999(E); ANSI/IEEE Std 802.11.

[16] G. Bianchi, Performance Analysis of the IEEE 802.11 Distributed Coordination Function, IEEE Journal on Selected Areas in Communications, Vol. 18, No. 3, pp. 535-547, March 2000.

[17] R. Litjens, F. Roijers, J. L. van den Berg, R. J. Boucherie, and M. Fleuren, Performance Analysis of wireless LANs: an Integrated Packet/Flow Level Approach, ITC Conference, Berlin, Germany, August 2003.

[18] J. Klaue, B. Rathke, and A. Wolisz, EvalVid - A Framework for Video Transmission and Quality Evaluation, 13th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation, pp. 255-272, Illinois, USA, September 2003.

[19] Recommendation ITU-T G.722.2, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)", Approved in July 2003.

[20] U. Toseef, M. Li, A. Balazs, X. Li, A. Timm-Giel, C. Görg, Investigating the Impacts of IP Transport Impairments on VoIP service in LTE Networks, 16th VDE/ITG Fachtagung Mobilkommunikation, 2011

[21] 3GPP Technical Report TS 36.213, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, v10.2.0, June 2011.

[22] IBM CPLEX Optimizer, http://www.ibm.com, as accessed in June 2013

[23] 3GPP Technical Report TS 25.814, Physical layer aspects for E-UTRA, 3rd Generation Partnership Project, v.7.1.0, September 2006.

[24] N. Zahariev, Y. Zaki, T. Weerawardane, C. Görg, and A. Timm-Giel. Optimized service aware lte mac scheduler with comparison against other well known schedulers. In 10th International Conference on Wired/Wireless Internet Communications, WWIC 2012, June 2012.

[25] 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, 8th International Wireless Communications and Mobile Computing Conference, Limassol, Cyprus, August 2012.

[26] U. Toseef, Y. Zaki, L. Zhao, A. Timm-Giel, and C. Görg, QoS Aware Multi-homing in Integrated 3GPP and non-3GPP Future Networks, The 7th International Conference on Systems and Networks Communications, Lisbon, Portugal, November 2012.

[27] Y. Zaki, T. Weerawardane, C. Görg, and A. Timm-Giel, Multi-QoS-Aware Fair Scheduling for LTE, VTC Spring, 2011.

[28] Recommendation P.800, Methods for subjective determination of transmission quality, Approved in August 1996.