SIP-based IMS Signaling Analysis for WiMax-3G Interworking Architectures

Arslan Munir and Ann Gordon-Ross

Abstract—The 3rd generation partnership project (3GPP) and 3GPP2 have standardized the IP multimedia subsystem (IMS) to provide ubiquitous and access network independent IP-based services for next generation networks via merging cellular networks and the Internet. The application layer Session Initiation Protocol (SIP), standardized by 3GPP and 3GPP2 for IMS, is responsible for IMS session establishment, management, and transformation. The IEEE 802.16 worldwide interoperability for microwave access (WiMax) promises to provide high data rate broadband wireless access services. In this paper, we propose two novel interworking architectures to integrate WiMax and 3rd generation (3G) networks. Moreover, we analyze the SIP-based IMS registration and session setup signaling delay for 3G and WiMax networks with specific reference to their interworking architectures. Finally, we explore the effects of different WiMax-3G interworking architectures on the IMS registration and session setup signaling delay.

Index Terms—IP multimedia subsystem (IMS), network architecture, Session Initiation Protocol (SIP).

1 Introduction and Motivation

The 3rd generation partnership project (3GPP) and 3GPP2 have standardized the IP multimedia subsystem (IMS) to provide IP-based rich multimedia services as well as content-based monetary charges for next generation networks. The Internet Engineering Task Force (IETF) has standardized the application layer Session Initiation Protocol (SIP) for the Internet and the 3GPP and 3GPP2 have standardized SIP for IMS session establishment (setup), management, and transformation. The IETF developed signaling compression (SigComp) for text-based protocol compression [1].

The evolving demand for mobile Internet and wireless multimedia applications has motivated the development of broadband wireless access technologies in recent years. The broadband wireless industry has recently focused on IEEE 802.16 worldwide interoperability for microwave access (WiMax) networks because WiMax addresses the issue of user equipment (UE) battery life, provide simultaneous support for high mobility and high data-rates, and provide a greater coverage area compared to wireless local area networks (WLANs). However, WiMax coverage area is limited when compared with 3G cellular networks, and 3G networks provide the added benefit of ubiquitous connectivity (although at lower data rates than WiMax networks). The complementary coverage area and data rate characteristics of WiMax and 3G networks motivate further exploration of their interworking with the intent of providing ubiquitous high speed wireless data access, and consequently, attracting a wider user base (WLAN-3G interworking architectures have been a large focus in previous work). The WiMax-3G interworking is interesting as WiMax is regarded as the next generation or fourth generation (4G) technology.

Although the 4G wireless networks are envisioned to provide better service than 3G wireless networks, the process of transitioning to 4G wireless networks is more than a simple technology upgrade, and requires significant changes to backhauls, radio sites, core networks, network management, service paradigms, and the mobile device distribution model [2]. In addition, it is important for 4G wireless technologies to reuse as much of the existing network and radio resources as possible, as well as provide an interworking with legacy systems. Like previous wireless technology deployment, 4G wireless network deployment would occur in distinct phases, thus filling the 4G network coverage gaps with legacy 2G/3G access technologies is necessary to provide a ubiquitous and seamless user experience. The integration of emerging 4G access technologies (e.g. mobile WiMax) with existing 2G/3G access technologies (e.g. code division multiple access (CDMA), Universal Mobile Telecommunications System (UMTS)) can be the first step towards the migration to mobile broadband networks that provide users with the best service experience at any place and anytime.

Even though near future WiMax enhancements may include global roaming support [3] to compete with the 3G cellular network market, it is important to consider that the existing 3G network customer base and infrastructure is much larger than that of WiMax, thus maximum revenue may only be achieved through the integration of these networks. Hence, in order to provide a uniform service experience and rich IP-based multimedia services to users, IMS is of particular importance with studying 3G and WiMax interworking. The 3G and WiMax interworking with IMS support would provide

- *Arslan Munir and Ann Gordon-Ross are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, 32611 USA. Ann Gordon-Ross is also with the NSF Center for High-Performance Reconfigurable Computing (CHREC) at the University of Florida, Gainesville, FL, 32611 USA e-mail: {amunir@ufl.edu, ann@chrec.org}*

users with access to heterogeneous wireless networks from any UE, and common billing and session management [4].

Two popular interworking system models for arbitrary wireless access networks (ANs) (such as 3G with WiMax in the case of this paper) are tight and loose coupling. In tightly coupled systems, the connecting AN integrates with the core 3G network similarly to any other 3G radio AN, using the same authentication, mobility, and billing infrastructures. To communicate with the 3G network, the connecting AN implements 3G radio access network protocols to route traffic through the core 3G elements. In loosely coupled systems, the connecting AN integrates with the core 3G network by routing communication traffic through the Internet, with no direct connection between the two networks. The two ANs use different authentication, billing, and mobility protocols, but however, may share the same subscriber databases for customer record management.

To the best of our knowledge, previous work does not specifically target the IMS infrastructure integration in WiMax-3G interworking architectures and only provides partial IMS signaling delay analysis. Specifically, "*reg event*" (informs users of their registration status within the IMS network) was not considered. The IMS session establishment signaling procedure is critical for quality of service (QoS) as session establishment negotiates the session between two UEs with agreed upon codecs. Additionally, previous work does not consider the Diameter Authentication procedures and provisional responses involved in the IMS session setup. Furthermore, the majority of previous work isolates the IMS signaling procedures from the interworking architectures. The architectural interaction effects are important because different interworking architectures contribute different delay and overhead to the IMS signaling procedures. Hence, IMS signaling delay analysis incorporating different interworking architectures evaluates the interworking architectures as well.

The main contributions of this paper are:

- We propose two novel WiMax-3G interworking architectures: the Loosely Coupled WiMax-Cellular (LCWC) and the Tightly Coupled WiMax-Cellular (TCWC) based on loosely and tightly coupling paradigms, respectively. The LCWC architecture enables independent WiMax and 3G network deployment and the TCWC architecture supports IMS sessions with QoS guarantees.
- We analyze the SIP-based IMS registration (including subscription to *reg event* state) and session setup signaling delay for various 3G and WiMax channel rates using a comprehensive model, which considers transmission, processing, and queueing delays at each network node. Our analysis considers provisional responses and DIAMETER authentication procedures involved in the IMS signaling as well as SigComp compression benefits.
- We investigate the effects of tightly and loosely coupled interworking architectures on the SIP-based IMS registration and session setup procedures. We also provide insights into the delay efficiency of WiMax-3G interworking architectures.

Our detailed analysis of IMS signaling procedures (registration and session setup) in 3G and WiMax networks will enable researchers in academia and industry to study SIP-based signaling performance before SIP protocol stack and IMS signaling capability implementation in WiMax-3G embedded devices. This analysis is important because IMS registration is a mandatory procedure before session establishment and IMS session setup negotiates the session between two UEs with agreed upon codecs.

The remainder of this paper is organized as follows. A review of related work is given in Section 2. Section 3 gives an overview of WiMax technology. Section 4 describes our proposed WiMax-3G interworking architectures. Section 5 describes signaling flows involved in the IMS registration and IMS session setup procedures. Section 6 outlines our proposed delay analysis model for studying SIP-based IMS signaling as well as our link layer analysis of SIP messages. Section 6 also explores the effects of WiMax-3G interworking architectures on IMS signaling. Numerical results are presented in Section 7. Section 8 gives future research work directions and Section 9 states conclusions.

2 Related Work

Interworking architectures have been studied in literature. Ruggeri et al. [5] presented interworking architectures based on loose and tight coupling paradigms. Other tightly and loosely coupled architectures were proposed and their costs analyzed in [6] and [7], respectively. Mahmood et al. [8] proposed an interworking architecture to integrate CDMA2000 (Code Division Multiple Access-based 3G system) with WLAN. Nguyen-Vuong et al. [9] presented WiMax to universal mobile telecommunications system (UMTS) handover signaling flows and proposed a UMTS-WiMax interworking architecture. Kim et al. [10] presented a loosely coupled UMTS-WiMax interworking architecture.

Much recent work exists in the area of interworking architectures. Lin et al. [11] presented a WiMax-WiFi integrated architecture that utilized a WiMax-WiFi access point device to combine the two technologies. [2] addressed the integration of mobile WiMAX with the 3GPP networks and proposed a handover mechanism that enabled seamless mobility between mobile access technologies with single radio mobile terminals. They concluded that the single radio handover (i.e. with terminals that do not need to simultaneously transmit on both access types) could mitigate the radio frequency (RF) coexistence issues that exist with dual-radio handover mechanisms with more intelligence in the network and mobile terminal. Munir et al. [12] proposed and analyzed

TABLE 1
WiMax MAC QoS classes

QoS class	Meaning	Description
UGS	unsolicited grant service	supports real-time constant bit rate (CBR) applications
rtPS	real-time polling service	supports real-time variable bit rate (VBR) applications
ertPS	extended real-time polling service	ensures a default CBR bandwidth and dynamically provides additional resources
nrtPS	non-real-time polling service	supports non-real-time VBR applications
BE	best effort	service with no bandwidth or delay guarantees

the cost of interworking architectures integrating 3G, WiMax, WLAN, and satellite ANs with IMS support.

An important aspect of interworking architectures is their IMS signaling efficiency, which is determined by the interworking architecture's ability to carry out the IMS signaling procedures (i.e., session establishment, registration, termination, and transformation) with minimum delay and overhead. Previous work provides limited IMS signaling delay analysis. Melnyk et al. [13] studied the IMS session establishment procedure when both the source node (SN) and the correspondent node (CN) are in CDMA2000. The IMS session establishment for a more general case where SN and CN are in different ANs was studied in [14]. Fathi et al. [15] studied SIP-based voice over IP (VoIP) IMS session establishment delay for 3G wireless networks using an adaptive lost packet retransmission timer. The authors studied different protocols such as transmission control protocol (TCP), user datagram protocol (UDP), and radio link protocol (RLP - an automatic repeat request (ARQ) medium access control (MAC) layer wireless interface protocol) for VoIP analysis. Xu et al. [16] presented an overview of 3GPP and WiMax networks based on IMS. They studied the 3GPP SIP extensions for QoS and authentication, authorization and accounting (AAA) provisioning.

Other interesting work exists in literature regarding IMS signaling analysis. Rajagopal et al. [17] analyzed IMS networks based on SIP signaling delay and formulated an IMS network utility function and calculated optimal network queue service rates considering delay constraints. Anzaloni et al. [18] conducted a performance study on the impact of authentication security levels for IMS in 3G networks. This performance study calculated the average authentication delay in integrated mobile IP and SIP mode where UE mobility required mobile IP registration and session mobility required SIP registration. The authors also suggested different IMS security levels for an authentication delay versus security level trade-off. Wu et al. [19] analyzed SIP-based vertical handoff (the delay incurred when a UE switches from one network to another) in WLAN and 3G networks. However, their work did not calculate IMS signaling delays, nor did the calculations consider processing delays. [14] studied SIP-based IMS session establishment for WiMax and 3G networks. However, [14] did not analyze the IMS registration procedure and the effects of different interworking architectures on the IMS signaling delay.

3 WiMax Overview

In this section, we give an overview of WiMax, WiMax QoS classes, and WiMax protocol structure. WiMax is the commercial name given to products that are compliant with the approved IEEE 802.16 standard [20] and associated enhancements, such as IEEE 802.16d, 802.16e, 802.16f, 802.16g, 802.16m [16]. The IEEE 802.16d standard supports low latency applications (i.e. audio, video) and provides broadband connectivity without requiring direct line of sight between UEs. IEEE 802.16e/f/g/m standards provide mobility support (referred to as mobile WiMax) and were adopted by the ITU (International Telecommunication Union) as one of the IMT-2000 (International Mobile Telecommunications-2000) technologies in November 2007 [21]. Since this adoption, mobile WiMAX has become a major global cellular wireless standard along with the 3GPP UMTS and 3GPP2 CDMA/EV-DO (Evolution Data Only).

WiMax can provide data rates up to 75 Mbps over long distances, with a theoretical coverage radius of approximately 50 km [22], and attempts to provide QoS guarantees for high-speed multimedia services (Table 1 depicts the WiMax MAC QoS classes). However, the IEEE 802.16 standard [20] leaves QoS support feature (e.g. traffic policing and shaping, connection admission control, and packet scheduling) implementation for WiMax vendors.

The IEEE 802.16 standard [20] specifies the MAC and physical (PHY) layers of the open system interconnection (OSI) model for WiMax. MAC and PHY functions can be classified into three categories: data plane, control plane, and management plane [23]. The data plane includes functions required for data processing such as header compression and MAC and PHY layer data packet processing. The control plane includes control functions necessary to support radio resource configuration, coordination, signaling, and management. The management plane includes functions required for external management and system configuration.

Fig. 1 depicts the WiMax protocol stack. The WiMax MAC layer consists of two sublayers: the convergence sublayer (CS) and the MAC common part sublayer (CPS). The WiMax protocol stack layers are integrated

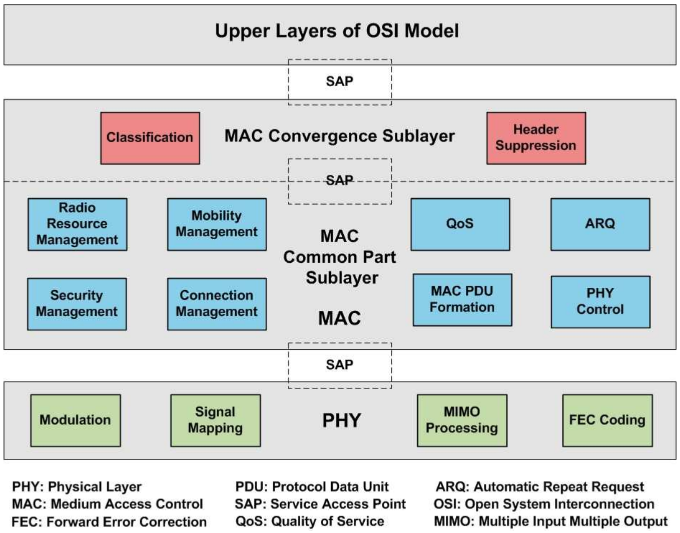


Fig. 1. WiMax Protocol Stack (adapted from [24]).

with service access points (SAPs) according to the IEEE 802.16 standard [25].

The CS enables the MAC layer to keep essential information, such as QoS parameters and destination addresses, for the upper layer service data units (SDUs). The *header suppression* block performs header suppression for the upper layer protocol packets. The *classification* block transforms/maps the IP address (from the upper layer or external network) into several service flow identifiers (SFIDs) and vice versa (from SFIDs to IP address). The CS records the mapping between a SFID and a transport connection ID (TCID) [26].

The MAC CPS maintains the MAC operations and generates management messages such as the ranging request/response (RNG-REQ/RNG-RSP), the downlink/uplink channel descriptor (DCD/UCD), and the downlink/uplink map (DL-MAP/UL-MAP) [26]. The CPS's main functions are mobility management, radio resource management, connection management, security management, MAC protocol data unit (PDU) formation, PHY control, QoS, and ARQ.

The *radio resource management* block adjusts radio network parameters according to user traffic load and includes functions for load balancing, admission control, and interference control. The *mobility management* block assists in handover operation. The *security management* block performs key management, data encryption/decryption, and authentication for secure communication. The *connection management* block allocates connection identifiers (CIDs) during initialization and/or handover, and interacts with the CS to classify MAC service data units (MSDUs) from the upper layers [22]. The *QoS* block performs rate control based on QoS input parameters from the connection management function for each connection. The *ARQ* block performs the MAC ARQ function. The *MAC PDU formation* block constructs MAC protocol data units (PDUs) for transmission of user traffic and/or management messages via the PHY. The *PHY control* block performs PHY signaling functions such as ranging and channel quality measurement.

Mobile WiMAX PHY uses orthogonal frequency division multiple access (OFDMA) and supports channel bandwidths of 3.5 MHz to 20 MHz, with up to 2048 subcarriers. The *Modulation* block supports QPSK (Quadrature Phase Shift Keying), 16-QAM (Quadrature Amplitude Modulation), and 64-QAM modulation schemes in the down link (DL) (from the base station to the UE) and the up link (UL) (from the UE to the base station). Mobile WiMAX supports link adaptation using adaptive modulation and coding (AMC) and power control. The *Signal Mapping* block handles bit mappings to the signal constellation. The *MIMO Processing* block provides support for multiple-input multiple-output (MIMO) antennas to provide good non-line-of-sight (NLOS) characteristics. The *FEC Coding* block performs forward error correction (FEC) coding for error correction purposes [27].

4 WIMAX-3G INTERWORKING ARCHITECTURES

We explain our proposed TCWC and LCWC interworking architectures along with the functionalities of various network nodes with reference to the 3GPP specification [28]. The “3GPP IP Access” refers to accessing the external IP networks such as IMS, 3G operators network, corporate Intranets or the Internet, through the 3GPP system. The TCWC architecture provides 3GPP IP Access. The “direct IP Access” refers to accessing a locally connected IP network from a WiMax network directly. The LCWC architecture provides direct IP Access.

4.1 TCWC: A Tightly Coupled WiMax-3G Interworking Architecture

Figure 2 depicts our proposed tightly coupled interworking architecture with IMS infrastructure support integrating WiMax and 3G wireless cellular networks. This tightly coupled paradigm directly interconnects an AN (such as WLAN or WiMax) with a 3G core network (as opposed to a loosely coupled paradigm in which an AN interconnects to a core 3G network via the Internet or Intranet). The dotted lines in Figure 2 represent 3G and WiMax base station coverage areas.

The WiMax AN consists of WiMax base stations that are controlled by the WiMax base station controller (WBSC). The WiMax network controller (WNC) controls several WBSCs and is connected to a wireless access gateway (WAG) to provide WiMax users with 3GPP packet-switched (PS) and IMS services. In the TCWC architecture, the WiMax-3G interworking function (WMIF) is responsible for abstracting WiMax network details and 3G protocol implementation for mobility management, authentication, etc. from the 3G core network. The WMIF connects to the serving GPRS (General Packet Radio Service) support node (SGSN) (in case of 3G UMTS) or the packet control function (PCF) (in case of 3G CDMA) of the 3G core network.

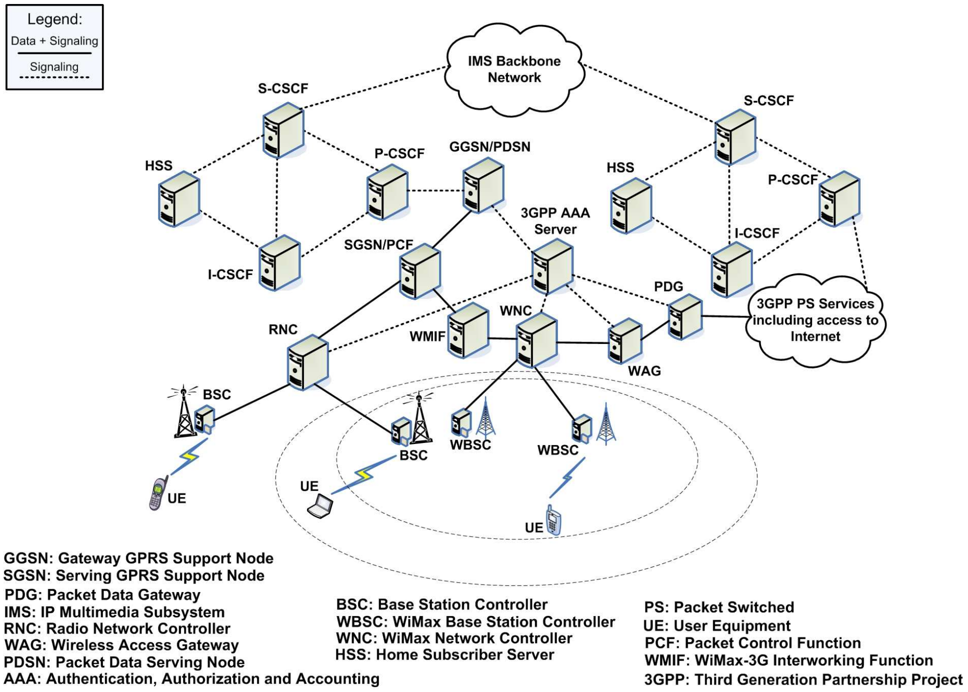

Legend:
Data + Signaling
Signaling
IMS Backbone Network
S-CSCF
HSS
P-CSCF
GGSN/PDSN
I-CSCF
SGSN/PCF
3GPP AAA Server
S-CSCF
HSS
P-CSCF
I-CSCF
PDG
3GPP PS Services including access to Internet
RNC
WMIF
WNC
WAG
BSC
BSC
WBSC
WBSC
UE
UE
UE
GGSN: Gateway GPRS Support Node
SGSN: Serving GPRS Support Node
PDG: Packet Data Gateway
IMS: IP Multimedia Subsystem
RNC: Radio Network Controller
WAG: Wireless Access Gateway
PDSN: Packet Data Serving Node
AAA: Authentication, Authorization and Accounting
BSC: Base Station Controller
WBSC: WiMax Base Station Controller
WNC: WiMax Network Controller
HSS: Home Subscriber Server
PS: Packet Switched
UE: User Equipment
PCF: Packet Control Function
WMIF: WiMax-3G Interworking Function
3GPP: Third Generation Partnership Project

Fig. 2. The TCWC Interworking architecture.

The connected WiMax and 3G ANs can be owned or operated by either the same or different service providers. The WAG of the WiMax network connects to the proxy-call session control function (P-CSCF) server in the IMS network via the packet data gateway (PDG). In general, there is a separate WAG for each AN. For IMS networks controlled by different operators, each network has separate serving-call session control function (S-CSCF) and interrogating-call session control function (I-CSCF) servers. In order to provide ubiquitous access, the two service providers should have a service level agreement (SLA) for IMS session establishment between the two ANs. An IMS backbone network connects IMS networks that are owned by different operators. If the WiMax and 3G networks are owned by the same operator, the packet data gateway (PDG) and gateway GPRS support node (GGSN) (packet data serving node (PDSN) in case of 3G CDMA) are connected to the same P-CSCF server.

The PDG provides 3GPP IP access to external IP networks. In the TCWC architecture, a UE is identified by multiple IP addresses. For example, when a UE in WiMax accesses IMS or 3GPP PS services, the UE is identified by two IP addresses - a local IP address and a remote IP address. In the WiMax AN, the local IP address identifies the UE and is used for packet delivery. The UE's local IP address may be translated by network address translation (NAT) before transmitting the packet from the UE to another IP network, including public land mobile networks (PLMNs). In the external network, which the WiMax is accessing via PDG, the remote IP address identifies the UE and is used to encapsulate data packets transmitted from the UE to the PDG tunnel.

A tunnel from the UE to the PDG carries PS-based service traffic in 3GPP IP Access. A single tunnel may carry the data for more than one IP flow and for different services. Individual IP flow and service traffic separation may not be possible at intermediate nodes due to possible IP header data encryption within these tunnels. However, QoS can still be assured if the WiMax UE and PDG deploy a differentiated service (DS) mechanism and appropriately mark (color) the DS field in the external IP header according to the QoS requirement for a particular traffic flow. The PDG assigns a remote IP address to the WiMax UE, registers the WiMax UE's local IP address, and binds the UE local IP address with the UE remote IP address. The PDG also performs the encapsulation/decapsulation of packets since the PDG is the terminating/originating point of tunnel between the UE and PDG. The WAG collects per tunnel accounting information (e.g. byte count, elapsed time, etc.) and sends this charging information to the 3GPP authentication, authorization, and accounting (AAA) server [28].

The TCWC architecture has many advantages including authentication, authorization and accounting (AAA) reuse, mobility management, and the QoS handling infrastructures in 3G cellular networks. The 3GPP system [28] provides WiMax network authentication. The TCWC architecture provides 3G services to WiMax users with guaranteed QoS and seamless mobility. In addition, extensions to the TCWC architecture can support relay-based WiMax networks, which would further provide QoS assurances for cellular transmission, particularly at cell edges [29]. Constant QoS level assurances are not feasible in the near future due to bandwidth differentials between different wireless access technologies. However,

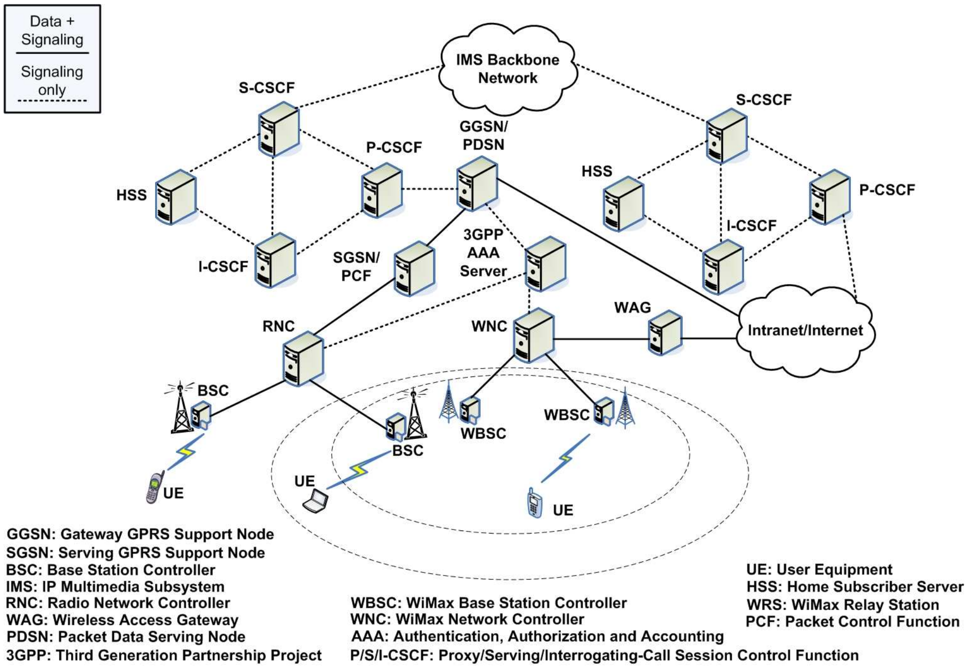

Data + Signaling
Signaling only
IMS Backbone Network
S-CSCF
P-CSCF
GGSN/ PDSN
HSS
I-CSCF
SGSN/ PCF
3GPP AAA Server
S-CSCF
HSS
I-CSCF
P-CSCF
WAG
RNC
WNC
Intranet/Internet
BSC
BSC
WBSC
WBSC
UE
UE
UE
GGSN: Gateway GPRS Support Node
SGSN: Serving GPRS Support Node
BSC: Base Station Controller
IMS: IP Multimedia Subsystem
RNC: Radio Network Controller
WAG: Wireless Access Gateway
PDSN: Packet Data Serving Node
3GPP: Third Generation Partnership Project
WBSC: WiMax Base Station Controller
WNC: WiMax Network Controller
AAA: Authentication, Authorization and Accounting
P/S/I-CSCF: Proxy/Serving/Interrogating-Call Session Control Function
UE: User Equipment
HSS: Home Subscriber Server
WRS: WiMax Relay Station
PCF: Packet Control Function

Fig. 3. The LCWC Interworking architecture.

QoS support ensures user service in accordance with a QoS profile and IMS application requirements. TCWC architecture QoS assurances can be provided via an appropriate QoS conversion mapping between 3G and WiMax QoS classes. QoS conversion mapping between WiMax and 3G UMTS is given in [4] and [10].

The TCWC architecture has several disadvantages. The TCWC architecture imposes direct exposure of the 3G core network interfaces to the WiMax network, which introduces security challenges. In addition, interworking function implementation requires extensive efforts, especially for the WiMax ANs not owned by the 3G operators. For seamless TCWC architecture operation, WiMax UEs must implement the 3G protocol stack on their standard network cards, which increases the non-recurring engineering (NRE) design cost for UEs. Furthermore, modification of the 3G core network nodes (i.e. SGSN/PCF and GGSN/PDSN) is required to handle the increased load caused by the direct injection of WiMax traffic.

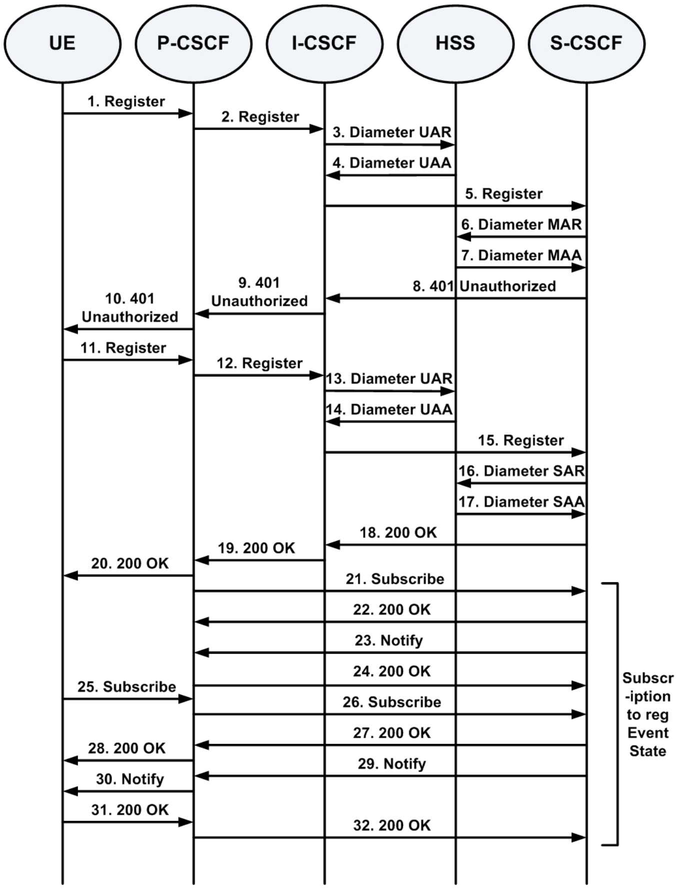
4.2 LCWC: A Loosely Coupled WiMax-3G Interworking Architecture

Figure 3 depicts our proposed LCWC interworking architecture which integrates WiMax and 3G wireless cellular networks based on a loosely coupled paradigm with IMS support. The dotted lines in Figure 3 represent 3G and WiMax base station coverage areas. Different ANs (3G networks and WiMax in our case) can be owned by different service providers (or the same operator). The WiMax WAG connects to the P-CSCF server in IMS via the Internet. In general, each AN has its own separate WAG and S-CSCF and I-CSCF servers. For IMS session establishment between two ANs, the two service providers should have a SLA with each other. The WAG and PDSN are connected to the same P-CSCF server if the same operator owns the WiMax and 3G networks.

The WAG is a gateway via which the data to/from the WiMax AN can be routed to/from an external IP network. The WiMax AN consists of WiMax base stations, which are controlled by the WBSC. Several WBSCs are controlled by one WNC. The WNC is connected to the WAG to provide WiMax users with 3GPP packet-switched (PS) and IMS services. Since the LCWC interworking architecture integrates 3G and WiMax networks based on the loosely coupled paradigm, the WiMax AN connects to the Internet or Intranet via the WAG, and through this connection, the UE can access the IMS network's CSCF servers.

In the LCWC architecture, the WiMax AN does not directly connect to 3G network elements, such as the SGSNs and GGSNs. The LCWC architecture has distinct signaling and data paths for the WiMax AN. The inter-operability with the 3G network requires mobile-IP functionality and SIP support to handle mobility across networks, and authentication, authorization, and accounting (AAA) services in the WiMax AN's WAG. This support is necessary to interwork with the 3G's home network AAA servers. The 3GPP system [28] provides WiMax network authentication.

The LCWC architecture's main advantage is that the LCWC architecture enables independent deployment and traffic engineering in WiMax ANs. In addition, this architecture utilizes standard IETF-based protocols for AAA and mobility in the WiMax network. Furthermore, since no interworking functions are required, LCWC



UE: User Equipment
HSS: Home Subscriber Server
Diameter UAR: Diameter User-Authentication-Request
Diameter UAA: Diameter User-Authentication-Answer
Diameter MAR: Diameter Multimedia-Authentication-Request
Diameter MAA: Diameter Multimedia-Authentication-Answer
Diameter SAR: Diameter Server-Assignment-Request
Diameter SAA: Diameter Server-Assignment-Answer
P/I/S-CSCF: Proxy/Interrogating/Serving-Call Session Control Function

Fig. 4. The IMS registration process (adapted from [30], [31].)

architecture deployment is less complex than TCWC architecture deployment. However, the main disadvantage of the LCWC architecture is that the LCWC architecture has no QoS guarantees because traffic must pass through the Internet (where QoS is difficult to assure).

5 The IMS Registration and Session Setup Procedures

In order to provide background for IMS registration and session setup analysis, we briefly describe the IMS registration and session setup procedures [30], [31]. For brevity, we limit our discussion to relevant IMS registration and session setup steps (see [30], [31] for further details). The IMS registration and session setup step numbers correspond to the numbers in Figure 4, and Figure 5, respectively.

5.1 IMS Registration Procedure

The IMS registration is a mandatory procedure in which the IMS user requests authorization to use the IMS services and consists of the following steps (Figure 4): 1) The IMS registration begins with a user equipment (UE) (a generic term for either SN or CN) SIP REGISTER request sent to the P-CSCF. 2) The P-CSCF forwards the SIP REGISTER request to the I-CSCF in the user's home network. 3) The I-CSCF sends a Diameter User-Authentication-Request (UAR) to the home subscriber server (HSS) for authorization and determination of S-CSCF (serving-call session control function) already allocated to the user. 4) The HSS authorizes the user and responds with a Diameter User-Authentication-Answer (UAA). 5) The I-CSCF forwards the SIP REGISTER request to the S-CSCF. 6) The S-CSCF sends a Diameter Multimedia-Authentication-Request (MAR) message to the HSS for downloading user authentication data. The S-CSCF also stores its uniform resource indicator (URI) in the HSS. 7) The HSS responds with a Diameter Multimedia-Authentication-Answer (MAA) message with one or more *authentication vectors*. 8) - 10) The S-CSCF creates a SIP 401 Unauthorized response with a challenge question that the UE SN must answer. 11), 12), & 15) The UE answers the challenge question in a new SIP REGISTER request response. 16) If authentication is successful, the S-CSCF sends a Diameter Server-Assignment-Request (SAR) to inform the HSS that the user is registered and the HSS can download the user profile. 17) The HSS replies with a Diameter Server-Assignment-Answer (SAA). 18) - 20) The S-CSCF sends a 200 OK message to inform the user of successful registration. The subscription to a *reg event* state provides the user with his/her IMS network registration status. 25) & 26) The UE sends a *reg event* SUBSCRIBE request to the P-CSCF, which then proxies the request to the S-CSCF. 27) & 28) The S-CSCF sends a 200 OK after accepting the *reg event* subscription. 29) & 30) The S-CSCF also sends a NOTIFY request containing registration information in extensible markup language (XML) format. 31) & 32) The UE finishes the subscription to the *reg event* state process by sending a 200 OK message. Note that steps 21) - 24) represent the *reg event* state subscription process for the P-CSCF and follow the same procedure as in steps 25) - 32).

5.2 IMS Session Setup Procedure

After the IMS SN UE is registered, the SN can initiate IMS session establishment with another registered IMS CN UE. The following steps outline the ISM session establishment procedure (Figure 5): 1) The SN initiates the IMS session establishment procedure by sending a SIP INVITE request to the SN's P-CSCF. 2) The P-CSCF responds by sending 100 Trying provisional response to the SN. 3) The P-CSCF forwards the INVITE request to the originating home network's S-CSCF. 5) The S-CSCF forwards the INVITE request to the appropriate terminating home network's I-CSCF. 7) To obtain the S-CSCF's address allocated to the user, the I-CSCF queries the HSS with a Diameter Location-Information-

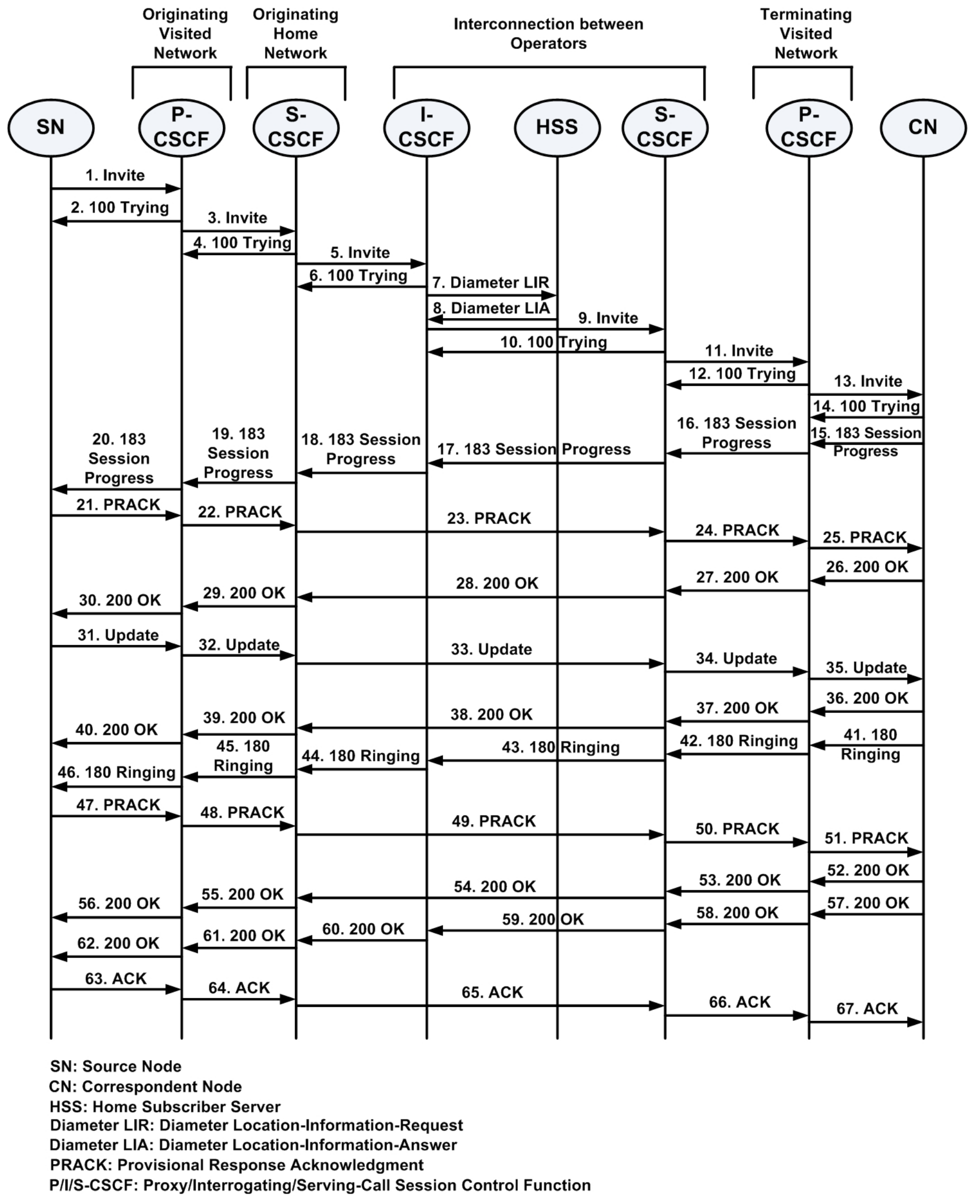


Fig. 5. The IMS session setup process (adapted from [30], [31].)

Request (LIR) message. 8) The HSS provides the S-CSCF's address in the Diameter Location-Information-Answer (LIA) message. 11) & 13) The S-CSCF forwards the SIP INVITE request to the CN via the CN's P-CSCF (assigned during CN registration). 15) - 20) The CN responds by sending the session description protocol's (SDP) provisional response 183 Session Progress to the SN, which informs the SN of the CN's supported and desired codecs for the session . 21) - 25) The SN acknowledges the provisional response 183 Session Progress with the SDP PRACK request containing modified codec information (if necessary, depending on the SN's supported and desired codecs). 26) - 30) The CN responds to the SDP PRACK request with a 200 OK. 31) - 35) The SN sends the UPDATE request to the CN after the SN's network resource reservation process. 36) - 40) The CN sends the CN's local network resource reservation status to the SN via the SDP 200 OK message. 41) - 46) When the CN UE rings, the CN sends the 180 Ringing provisional response to the SN. This response traverses the same CSCF servers that the INVITE request traversed. 47) - 51) When the SN receives the 180 Ringing response, the SN generates a locally stored ring-back tone to indicate to the caller that the CN UE is ringing and sends a PRACK request to the CN. 52) - 56) The CN sends the 200 OK response to the PRACK request. 57) - 62) The CN sends the 200 OK response to the INVITE transaction after the callee accepts the session. 63) - 67) The SN sends the ACK request to the CN, thus completing the IMS session establishment procedure.

6 Delay Analysis Model for the IMS Signaling Procedures

In this section, we present our delay analysis model for the IMS signaling procedures with an emphasis on the IMS registration process. Our SIP signaling delay analysis begins after the general packet radio service (GPRS) attach procedure and the packet data protocol (PDP) context activation procedure in the 3G and WiMax networks. These procedures are required to obtain an IP address. Next, we present our SIP message link layer analysis for various channel rates. We then evaluate the effects of different WiMax-3G interworking architectures on the IMS signaling analysis. Even though our proposed analytical modeling paradigm is broadly applicable to any networking multimedia communication protocol, in this section we describe the details pertaining to IMS signaling.

The IMS registration and session setup signaling delay is comprised of three components:

$$\overline{D} = D_t + D_p + D_q \tag{1}$$

where $\overline{D}$, D_t, D_p, and D_q denote the total average IMS signaling delay, average transmission delay, average processing delay, and average queueing delay, respectively. In the following subsections, we describe these components in detail.

6.1 Transmission Delay

The transmission delay is the delay incurred during signaling message transmission and is affected by message size, channel bandwidth, and propagation delay (we incorporate propagation delay into the transmission delay). The propagation delay is the delay incurred due to signaling message propagation between nodes and is affected by distance between nodes and wireless/wired channel characteristics. Our transmission delay model only considers the wireless link transmission delays because we can assume that the wired link transmission delay is negligible due to high available bandwidth and low bit error rates (BERs) [32]. We model the wireless link transmission delay with and without RLP using TCP for the transport layer protocol [33]. Our TCP analysis for WiMax is crucial as IEEE 802.16m evaluation methodology documents TCP layer throughput (and/or delay) metric as a mandatory performance measurement criterion in addition to the PHY and MAC layer throughput measurement [27].

The average delay for successful TCP segment transmission with no more than N_{TCP} retransmission trials and without RLP $D_{TCPnoRLP}$ is [34]:

$$\begin{aligned} D_{TCPnoRLP} &= (K-1)\tau + \frac{D}{(1-q^{N_{TCP}})(1-2q)} \\ &+ \frac{1-q}{1-q^{N_{TCP}}} D \left[\frac{q^{N_{TCP}}}{1-q} - \frac{2^{N_{TCP}+1} q^{N_{TCP}}}{1-2q} \right] \end{aligned} \tag{2}$$

where K is the number of frames per packet, τ is the inter-frame time, D is the end-to-end frame propagation delay over the radio channel, q denotes the packet loss rate without RLP, and N_{TCP} indicates the maximum allowable TCP retransmissions in case of packet loss.

The average delay for successful TCP segment transmission with no more than N_{TCP} retransmission trials and with RLP $D_{TCPwithRLP}$ is [34]:

$$\begin{aligned} D_{TCPwithRLP} &= D_{RLP} + \frac{2Dr(1-r)}{1-r^{N_{TCP}}} \\ &\times \left[1 + \frac{4r\left(1-(2r)^{N_{TCP}-2}\right)}{1-2r} - \frac{r\left(1-r^{N_{TCP}-2}\right)}{1-r} \right] \end{aligned} \tag{3}$$

where D_{RLP} denotes the packet delay when RLP is used and r denotes the RLP packet loss rate.

The IMS registration procedure, including subscription to the *reg event* state, consists of eight message exchanges between the UE and the IMS network's P-CSCF server (Messages 1, 10, 11, 20, 25, 28, 30, and 31 in Figure 4). Whereas 3G networks improve frame error rate (FER) with RLP, WiMax networks do not use RLP due to higher available bandwidth [35]. The IMS registration transmission delay in 3G networks $D_{t-imsreg-3g}$ is:

$$D_{t-imsreg-3g} = 8 \times D_{TCPwithRLP} \tag{4}$$

The IMS registration transmission delay in WiMax networks $D_{t-imsreg-wimax}$ is:

$$D_{t-imsreg-wimax} = 8 \times D_{TCPnoRLP} \tag{5}$$

The IMS session setup procedure's transmission delay can be modeled in a similar fashion [14].

6.2 Processing Delay

The processing delay is the delay incurred during packet encapsulation and decapsulation at the network layer. We model the processing delays for network nodes in the IMS signaling path. The main processing delay for IMS databases is the address lookup delay. IMS databases can store user records based on IP address in address tables. When a user record is queried, the IMS databases search address tables for the queried user among all network users N (this assumption is valid since we consider only one tier of a complete network infrastructure). Address table lookup time can be reduced using larger cache line sizes for multi-way search or an adaptive multiple-column binary search method for longer IPV6 addresses [36]. Our models assume that the IMS database uses the adaptive binary search method.

The HSS stores the Private User Identity and the collection of Public User Identities assigned to a user [31]. The HSS uses the Private Identity as an index to retrieve the International Mobile Subscriber Identity (IMSI) and the user profile [37]. Hence, the address lookups can be performed on the Private/Public user identity. We consider the possibility of storing IP addresses in the HSS for completeness since the IMS UE is a device that has IP connectivity and is able to request an IP address from the network (e.g. SIP Phone, personal computer

(PC), personal digital assistant (PDA)). Our processing delay analysis is equally applicable to both alternatives (whether the lookup is performed on IP addresses or the Private/Public user identity) because the processing delay is dependent on the identity length.

We assume a fixed processing delay d_{p-ed} for packet encapsulation and decapsulation for the IMS network nodes that do not perform an IMS database lookup. This assumption does not affect our model's accuracy as the processing delay accounts for a very small fraction of the total average delay [12].

The IMS database HSS processing delay d_{p-hss} in nanoseconds is the sum of the address lookup delay and the fixed processing delay d_{p-ed}:

$$d_{p-hss} = d_{p-ed} + 100\left(\log_{k+1} N + \frac{L}{S}\right) \text{ ns} \tag{6}$$

where L denotes the IP address (or Public/Private User Identity) length in bits (e.g. L is 32 or 128 for IPv4 and IPv6, respectively). We note that the Public/Private User Identity length is not constant and varies for different users (depending upon the user name), but we can assume that the length is equal to L for a typical case. Appropriate identity lengths depend on actual implementation details and can be substituted in (6). S is the machine word size in bits and k is a system-dependent constant. The adaptive binary search method for the address lookup attributes the $\log$ factor. The 100 ns multiplication factor accounts for the fact that the address lookup time increases for each memory access [36].

The IMS registration processing delay $D_{p-imsreg}$ is:

$$\begin{aligned} D_{p-imsreg} &= 4d_{p-sn} + 10d_{p-pcscf} + 6d_{p-icscf} \\ &\quad + 4d_{p-hss} + 8d_{p-scscf} \end{aligned} \tag{7}$$

where d_{p-sn}, $d_{p-pcscf}$, $d_{p-icscf}$, d_{p-hss}, and $d_{p-scscf}$ denote the unit packet processing delay at the SN, P-CSCF, I-CSCF, HSS, and S-CSCF, respectively, and integer coefficients denote the number of IMS registration signaling messages processed at respective nodes (see Figure 4). Thus, a node's processing delay is modeled by counting the number of messages a node receives.

The IMS session setup processing delay can be modeled similarly as:

$$\begin{aligned} D_{p-imssetup} &= 7d_{p-sn} + 24d_{p-pcscf} + 24d_{p-scscf} \\ &\quad + 6d_{p-icscf} + d_{p-hss} + 5d_{p-cn} \end{aligned} \tag{8}$$

where d_{p-cn} denotes the unit packet processing delay at the CN and integer coefficients denote the number of IMS session establishment signaling messages processed at respective nodes (see Figure 5).

6.3 Queueing Delay

The queueing delay is the delay incurred due to packet queueing at network nodes. Our queueing delay model includes all network nodes involved in the IMS signaling procedures. A queue's total packet queueing delay is the summation of the queueing delay at each node the packet traverses between the SN and the CN. The queueing delay at a node depends upon the number of currently queued packets at that node. We model the SN and IMS network nodes (P-CSCF, I-CSCF, S-CSCF, and HSS) with M/M/1 queues and a Poisson process signaling arrival rate. Calculating the average queueing delays using the M/M/1 queueing model is a valid assumption since the M/M/1 model gives sufficiently accurate results for WiMax and 3G networks [15], [19]. The rationale behind these assumptions is that the SN and the IMS network nodes perform dedicated tasks and thus their service rate can be captured by exponential distribution [35].

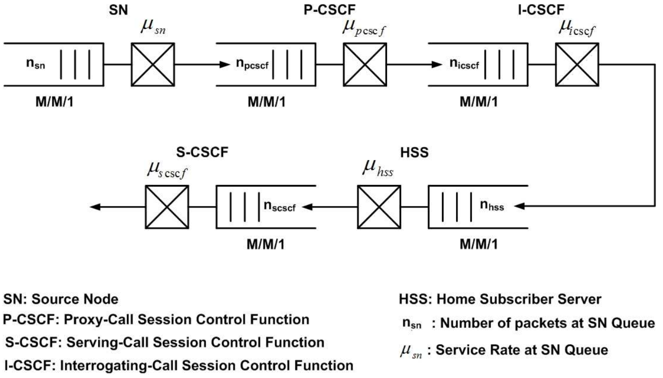

SN
μ_{sn}
n_{sn}
M/M/1
P-CSCF
μ_{pcscf}
n_{pcscf}
M/M/1
I-CSCF
μ_{icscf}
n_{icscf}
M/M/1
S-CSCF
μ_{scscf}
n_{scscf}
M/M/1
HSS
μ_{hss}
n_{hss}
M/M/1
SN: Source Node
P-CSCF: Proxy-Call Session Control Function
S-CSCF: Serving-Call Session Control Function
I-CSCF: Interrogating-Call Session Control Function
HSS: Home Subscriber Server
n_{sn} : Number of packets at SN Queue
μ_{sn} : Service Rate at SN Queue

Fig. 6. Queueing network for IMS registration process.

For an M/M/1 queue to be in an equilibrium state, the input and output Poisson processes must have equal arrival and departure rates given by λ [38]. For a queueing network with M/M/1 queues in tandem, if the first queue's input process is Poisson, the next stage's M/M/1 queue's input process is also Poisson, and so on [39]. Figure 6 depicts *queueing network* of M/M/1 queues in tandem for the IMS registration process. Thus, the IMS registration queueing delay $D_{q-imsreg}$ is:

$$\begin{aligned} D_{q-imsreg} &= 4E[w_{sn}] + 10E[w_{pcscf}] + 6E[w_{icscf}] \\ &\quad + 4E[w_{hss}] + 8E[w_{scscf}] \end{aligned} \tag{9}$$

where integer coefficients denote the number of IMS registration signaling messages received at respective nodes and $E[w_{sn}]$, $E[w_{pcscf}]$, $E[w_{icscf}]$, $E[w_{hss}]$, and $E[w_{scscf}]$ denote the packet queueing delay at the SN, P-CSCF, I-CSCF, HSS, and S-CSCF, respectively (see Figure 4). The packet queueing delay at the SN is [39]:

$$E[w_{sn}] = \frac{\rho_{sn}}{\mu_{sn}(1-\rho_{sn})} \tag{10}$$

where $\rho_{sn} = \lambda_{e-sn}/\mu_{sn}$ denotes the SN queue's utilization, μ_{sn} denotes the SN queue's service rate, and λ_{e-sn} denotes the *effective arrival rate* (in packets per second) at the SN queue. Thus, $\lambda_{e-sn} = \sum_{i \in N_{sn}} \lambda_i$, where N_{sn} denotes the number of active sessions at the SN, which includes the current IMS signaling session. A network node's effective arrival rate λ_e can be calculated from that node's utilization. λ_e and the associated queueing time for other network nodes can be modeled in a similar fashion. The IMS session setup queueing delay can be modeled following an analogous approach:

$$\begin{aligned} D_{q-imssetup} \quad &= \quad 7E[w_{sn}] + 24E[w_{pcscf}] + 24E[w_{scscf}] \\ &\quad + 6E[w_{icscf}] + E[w_{hss}] + 5E[w_{cn}] \end{aligned} \tag{11}$$

where integer coefficients denote the number of IMS session establishment signaling messages received at respective nodes and $E[w_{cn}]$ denotes the packet queueing delay at the CN (see Figure 5).

In order to generalize our proposed queuing model, we reanalyze our system using general packet arrival and service time distributions. The IMS registration and session setup queueing network consists of G/G/1 queues in tandem. The expected packet queueing delay at the CN using the G/G/1 model can be given as:

$$\begin{aligned} E[w_{cn}] \quad &= \quad \frac{\lambda^2_{e-cn}(\sigma^2_{u-cn} + \sigma^2_{v-cn}) + (1-\rho_{cn})^2}{2\lambda_{e-cn}(1-\rho_{cn})} \\ &\quad - \frac{\nu^{(2)}_{h-cn}}{2\nu_{h-cn}} \quad : \quad \rho_{cn} < 1 \end{aligned} \tag{12}$$

where λ_{e-cn} denotes the effective arrival rate at the CN queue, $\rho_{cn} = \lambda_{e-cn}/\mu_{cn}$ denotes the CN queue's utilization (μ_{cn} denotes the CN queue's service rate), σ^2_{u-cn} denotes the variance of packet inter-arrival times at the CN queue, and ν_{h-cn} and ν^2_{h-cn} denote the first and second moments of the CN queue *idle period* I_{cn}, respectively. The expected value of the CN queue idle period (the period of time when there are no packets in the queue) can be given as:

$$E[I_{cn}] = \frac{1-\rho_{cn}}{\lambda_{e-cn}p_{0-cn}} \quad : \quad \rho_{cn} < 1 \tag{13}$$

where p_{0-cn} is the probability that a packet arrives when the CN queue is empty. The expected packet queueing delays for other network nodes (SN, P-CSCF, I-CSCF, S-CSCF, HSS) can be written similar to (12). The expected packet queueing delays using G/G/1 queues can substituted into (9) and (11) to obtain the G/G/1 queueing network delay for the IMS registration and IMS session setup processes, respectively.

Our proposed G/G/1 model is rigorous and captures all the distributions of packet arrival and service times. Our M/M/1 analysis ((9), (10), and (11)) is a special case of G/G/1 where packet arrival and service times are Markovian [40].

A priority based M/G/1 model could be used for the CN (a special case of G/G/1 where packet arrival time is Markovian with general packet service time) with the assumption that while the IMS network nodes perform dedicated jobs (and thus have an M/M/1 model), the CN may be busy with a variety of other messages aside from SIP messages, and thus may have a general service time distribution [35]. The expected queueing delay at the CN queue using the M/G/1 model can be given as:

$$E[w_{cn}] = \frac{1}{\mu_{cn}} + \frac{\lambda_{e-cn}}{2(1-\rho_{cn})} S^{(2)}_{cn} \quad : \quad \rho_{cn} < 1 \tag{14}$$

where $S^{(2)}_{cn}$ denotes the second moment of packet service time and the remaining terms have the same meaning as defined above for (10) and (12).

6.4 Total Delay

The total IMS registration delay for a 3G network is:

$$\overline{D}_{imsreg-3g} = D_{t-imsreg-3g} + D_{p-imsreg} + D_{q-imsreg} \tag{15}$$

The total IMS registration delay for a WiMax network is:

$$\begin{aligned} \overline{D}_{imsreg-wimax} \quad &= \quad D_{t-imsreg-wimax} + D_{p-imsreg} \\ &\quad + D_{q-imsreg} \end{aligned} \tag{16}$$

The equations governing the total IMS session setup delay can be similarly derived. It is important to note that if the SN and CN have not registered with the IMS network, then they must undergo the IMS registration process before session establishment.

6.5 SIP Message Application and Link Layer Analysis in WiMax and 3G Networks

In this subsection, we analyze the application layer SIP message sizes and associated link layer frames after SigComp-based compression. SigComp can reduce SIP message sizes by as much as 88% with negligible compression and decompression overhead. In our analysis, we use compression rates of 55% and 80% for initial SIP messages (such as INVITE, REGISTER) and subsequent SIP messages (200 OK, SUBSCRIBE, NOTIFY, 401 UNAUTHORIZED, etc.), respectively [1], [13]. Using these compression rates, the SIP message size for INVITE is 810 bytes; REGISTER is 225 bytes; 183 SESSION PROGRESS, PRACK, 100 TRYING, 180 RINGING, and UPDATE is 260 bytes; ACK is 60 bytes; and all subsequent SIP messages are 100 bytes.

We calculate the number of frames per packet K for different 3G and WiMax channel rates using a 3G transmission model [34] and the model's extension for WiMax. For the 3G network, we consider 19.2 kbps and 128 kbps channel rates and for the WiMax network, we consider 4 Mbps and 24 Mbps channel rates. We choose these particular channel rates based on commonly available channel rates but other channel rates result in similar trends for IMS registration and IMS session establishment [35]. For the 4 Mbps WiMax network, we assume quaternary phase shift keying (QPSK) modulation and a 1/2 convolutional code rate. For the 24 Mbps WiMax network, we assume 64-QAM (Quadrature Amplitude Modulation) and a 3/4 convolutional code rate [26]. For the 3G network, we assume an RLP frame duration and inter-frame time τ of 20 ms [34]. For the WiMax network, we assume a frame duration and inter-frame time of 2.5 ms, which is independent of the channel rate [29]. For the

TABLE 2
Number of frames per packet K for various 3G (19.2 and 128 kpbs) and WiMax (4 and 24 Mbps) channel rates

SIP Message	Size (Bytes)	19.2 kbps	38.4 kbps	128 kbps	4 Mbps	24 Mbps
SIP INVITE	810	17	9	3	1	1
SIP REGISTER	225	5	3	1	1	1
183 SESSION PROGRESS	260	6	3	1	1	1
SIP 180 RINGING	260	6	3	1	1	1
SIP PRACK	260	6	3	1	1	1
SIP 100 TRYING	260	6	3	1	1	1
SIP UPDATE	260	6	3	1	1	1
SIP 200 OK	100	3	2	1	1	1
SIP SUBSCRIBE	100	3	2	1	1	1
SIP NOTIFY	100	3	2	1	1	1
SIP 401 UNAUTHORIZED	100	3	2	1	1	1
SIP ACK	60	2	1	1	1	1

19.2 kbps 3G network channel rate, each frame consists of $19.2 \times 10^3 \times 20 \times 10^{-3} \times \frac{1}{8} = 48$ bytes. For the SIP REGISTER message, K = $\lceil \frac{225}{48} \rceil = 5$. Following the same methodology, Table 2 shows the K values for all the relevant IMS registration and session setup messages. Our transmission delay analysis carefully considers these K values for all signaling messages exchanged on the wireless link.

6.6 Case Study: CDMA2000 Evolution Data Only Wireless Transmission

We analyze the CDMA2000 Evolution Data Only (EV-DO) standard and associated wireless link transmission analytical model as a specific case for generic 3G networks. The presented model is based on the existing 3GPP2 standards and incorporates the characteristics of the EV-DO wireless channel as well as transport layer protocols [41], [42]. EV-DO, and its enhanced version EV-DO's Rev. A (EV-DO rev. A), is widely adopted as the 3G high-speed wireless data standard [43]. EV-DO operates at various UE/base station negotiated data transmission rates and frame length combinations based on the wireless channel condition. To minimize FER, EV-DO rev. A reduces transmission rates and frame lengths as the channel interference increases. (For brevity in the remainder of this paper, we refer to EV-DO rev. A as EV-DO.)

The frame retransmission mechanism in the EV-DO standard is based on RLP. The forward link (FL) is time division multiplexed and divided into time slots of duration 1.667 ms or 600 slots/second. In an FL traffic channel, variable transmission rates are achieved via data rate control (DRC) values. Each DRC is associated with a value pair consisting of physical layer frame length and number of slots per frame. For example, the DRC 1 format has a frame length of 1024 bits and 16 slots per frame, resulting in a transmission rate of $(1024/16) \times 600 = 38400$ bps. The DRC variable transmission rate is achieved by changing underlying communication system parameters such as modulation schemes, coding gains, preamble size, etc. In a 4-slot interlacing scheme, one particular slot is used to transmit data associated with a given frame (i.e. each fourth slot allocated to an active terminal is separated by three slots used by other terminals). This interleaving scheme improves system throughput by using a hybrid automatic repeat request (HARQ) scheme. When HARQ is used, *early termination* enables the receiver to decode the complete physical layer frame before all nominal slots are received. The reverse link (RL) operation is similar to the FL operation.

The fragmentation of packets into frames is crucial in the RLP layer model. For a transport layer packet of M bits received for transmission at the EV-DO layer, the packet will be divided into k frames depending upon the physical layer frame payload size α (in bits) and the associated overhead δ (136 bits per frame):

$$k = \left\lceil \frac{M}{\alpha - \delta} \right\rceil \tag{17}$$

From the transport layer perspective, a successful packet transmission takes place if all the frames of a packet are transmitted successfully either on the first attempt or on the subsequent RLP retransmissions. The mean RLP delay is [42]:

$$\begin{aligned} E\left(D_{RLP}\right) &= \frac{1}{P_s} \sum_{j=0}^{k} \binom{k}{j} (1-p)^{k-j} \\ &\times \left(p(1-p)^2\right)^j E\left(D_{RLP}|k,j\right) \end{aligned} \tag{18}$$

where P_s is the probability of successful transmission of a packet with k frames, p is the FER (the probability that a transmitted frame is lost), j denotes the number of frames initially lost and then recovered by retransmission, and $E\left(D_{RLP}|k,j\right)$ denotes the expectation of RLP packet delay for a k frame packet given j RLP retransmissions. If the number of transmission attempts

TABLE 3
Number of frames per packet k, forward link slot span $\widehat{S}_{FL}$ and reverse link slot span $\widehat{S}_{RL}$ for SIP signaling messages in CDMA2000 EV-DO DRC 1

SIP Message	Size (Bytes)	k	$\widehat{S}_{FL}$	$\widehat{S}_{RL}$
SIP INVITE	810	8	5	2
SIP REGISTER	225	3	6	2
SESSION PROGRESS	260	3	6	2
SIP 180 RINGING	260	3	6	2
SIP PRACK	260	3	6	2
SIP 100 TRYING	260	3	6	2
SIP UPDATE	260	3	6	2
SIP 200 OK	100	1	16	5

at the transport layer is N, the transport level packet mean delay D_{TLP} is [42]:

$$\begin{aligned} D_{TLP} &= \frac{1}{1-q^N}\sum_{j=0}^{N-1}(1-q)q^j \\ &\quad \times \big(D_{RLP}+(2^j-1)RTO\big) \end{aligned} \qquad (19)$$

where RTO denotes the transport layer protocol's retransmission time-out value and q denotes the packet loss rate.

For SIP-based IMS registration and session establishment procedures, we consider DRC 1, which corresponds to poor channel condition with a physical layer frame payload size of 1024 bits. We calculate the number of frames k for different SIP messages. For the SIP INVITE message with a compressed size of 810 bytes, $k = \lceil (810 \times 8)/(1024 - 136) \rceil = 8$. For large k values, we use the effective FL slot span $\widehat{S}_{FL} = \lfloor \bar{S}_{FL}+1 \rfloor$. For single frame packets ($k = 1$), we use $\widehat{S}_{FL} = S_{FL}$. For all other k values, we determine $\widehat{S}_{FL}$ by extrapolating linearly between these two values (i.e. the values obtained for packets with large number of frames and single frame packets). S_{FL} and $\bar{S}_{FL}$ denote the nominal FL slot span and the average FL slot span, respectively. For large and medium k, we use the effective RL slot span $\widehat{S}_{RL} = 2$. For $k = 1$, we use $\widehat{S}_{RL} = 5$. For the remaining k values, $\widehat{S}_{RL}$ can be determined using extrapolation between the two values (i.e. the values obtained for packets with a large number of frames and single frame packets). Table 3 shows the number of frames k, effective FL slot span $\widehat{S}_{FL}$ due to early termination gain, and effective RL slot span $\widehat{S}_{RL}$ for selected SIP messages for DRC 1 with a 38.4 kbps channel rate.

6.7 Interworking Architectures and the IMS Signaling Delay

Our delay analysis is equally valid for all WiMax-3G interworking architectures supporting IMS services. Different architectures cause the IMS registration signaling messages (sent from the IMS terminal (UE) to the first point of contact with the IMS network) to flow between different architecture specific nodes. More specifically, for different architectures, different network nodes will be along the path between the UE (SN or CN) and the P-CSCF. To aggregate total delay, these differences require specific modeling of the network nodes involved. For the TCWC architecture, the total delay from the SN to the P-CSCF in a 3G network includes the delays incurred at the base station controller (BSC), radio network controller (RNC), SGSN/PCF, and GGSN/PDSN. It should be noted that the SIP messages are transmitted on wireless links from the SN to the BSC while the GGSN will be hard-wired to the P-CSCF. Similarly, delays incurred at the PDG, WAG, WNC, and WBSC should be added to the signaling delay from the P-CSCF to the CN in a WiMax network. For the LCWC architecture, delays incurred at the BSC, RNC, SGSN/PCF, and GGSN/PDSN constitute the additional delay from the SN to the P-CSCF in a 3G network. The delays incurred at the Internet, WAG, WNC, and WBSC constitute the incremental delay from the P-CSCF to the CN in a WiMax network.

7 Numerical Delay Analysis

In this section, we present the numerical results for the delay analysis of SIP-based signaling for IMS sessions in WiMax-3G interworking architectures.

7.1 Parameter Values

We present numerical results, which reflect the results obtained from an actual prototype implementation of 3G, WiMax, and IMS infrastructures with parameter values selected carefully from standard literature. Our analyzed network consists of two 3G base station controllers (BSCs) and three WiMax BSCs. The 3G BSC cell radius is 1000 m and WiMax BSC cell radius is 700 m. The user density is 0.001 per square meter for both networks [6], [7], [44]. These cell radii and user densities specify the number of users in the 3G cellular network, N_{mn1} = 5000 and the WiMax network, N_{mn2} = 3000.

For the transmission delay calculation, the frame error probability p can be obtained from the FER. The end-to-end frame propagation delay D for both the 19.2 kbps and 128 kbps 3G channels is 100 ms. For the 4 Mbps and 24 Mbps WiMax channels, D is 0.27 ms and 0.049 ms, respectively [35]. Both the frame duration T and the inter-frame time τ for the 3G and WiMax networks is assumed to be 20 ms [34] and 2.5 ms, respectively, and is independent of the channel bit rate [29]. The maximum RLP retransmissions n and maximum number of TCP retransmissions N_{TCP} are assumed to be 3 [15], [34], [45].

For the lookup processing delay calculation, we assume an address length L of 32 bits (corresponding to IPv4 and/or assuming the same constant Public/Private user identity length for user identity lookups) and a processor machine word size S of 32 bits (for 32-bit machines). However, numerical results can be obtained for IPv6 and 64-bit machines by setting L and S equal to 64. We assume a system dependent constant value k equal to 5 [36]. The unit packet processing delay for SGSN/PCF, GGSN/PDSN, and the Internet is assumed to be 8×10^{-3} seconds. The unit packet processing delay for the remainder of the network nodes is assumed to be 4×10^{-3} seconds [6], [7]. This constant processing delay assumption does not invalidate our results as the processing delay constitutes a very small percentage of the total average delay and result trends remain similar even if we assume a variable processing delay at each node [12].

For the queueing delay calculation, we assume a service rate μ of 250 packets/sec at all nodes [6]. The signaling and data traffic from other network sources constitutes the *background utilization* at network nodes. Since the HSS must handle network traffic from different ANs, the HSS's background utilization is assumed to be 0.7. We assume background utilization of 0.5 for the SGSN/PCF and GGSN/PDSN and 0.7 for the Internet. The background utilization is assumed to be 0.4 for the remaining nodes. We base these background utilization assumptions on the average traffic load estimates at respective nodes. However, our model is applicable to any other background utilization values as background utilization fluctuates during different periods of the day. These background utilization values also determine a network node's *effective arrival rate* λ_e. First, the arrival rate due to background utilization λ_{bg} is calculated as $\lambda_{bg} = \rho_{bg} \cdot \mu$ where ρ_{bg} is a node's background utilization and μ is the node's service rate. The effective arrival rate is simply the sum of the arrival rate due to background utilization and the signaling arrival rate for the session λ_s, i.e. $\lambda_e = \lambda_{bg} + \lambda_s$. Finally, λ_e is used to calculate the *effective utilization* ρ_e of a node as $\rho_e = \lambda_e/\mu$ which is then ultimately used to calculate a packet's expected waiting time at a node's queue using (10).

7.2 Channel Rate Effects on IMS Signaling Delay

Our first experiment analyzes the IMS registration and session setup signaling delay for 3G network channel rates of 19.2 kbps and 128 kbps and WiMax network channel rates of 4 Mbps and 24 Mbps. The frame error probability rate p and the IMS signaling arrival rate λ in packets per second are fixed at 0.02 and 9, respectively.

Figure 7 shows IMS registration signaling delay in seconds versus varying channel rates. The figure shows that for 3G networks, the IMS registration signaling delay decreases with increased channel rate, while for the WiMax network, the IMS registration signaling delay remains nearly constant with increased channel rate. Furthermore, the delay for WiMax networks is considerably less than for 3G networks. These results are due to high WiMax channel rates, which reduce transmission delay effects.

Figure 8 depicts the IMS session setup delay when the SN is in a 3G network and the CN is in a WiMax network for different combinations of 3G and WiMax channel rates. It can be noticed that the IMS session setup delay is greatly affected by the 3G channel rate (IMS session setup delay decreases considerably as the 3G channel rate increases), whereas the IMS session setup delay is negligibly affected by changing the WiMax channel rate.

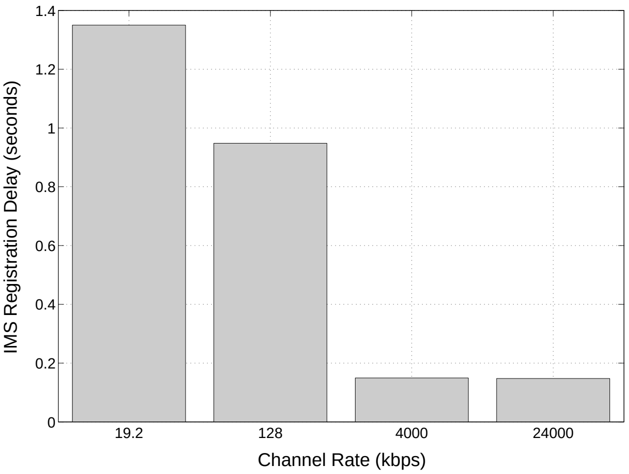


Fig. 7. IMS registration signaling delay for various channel rates for a fixed signal arrival rate λ and frame error probability p.

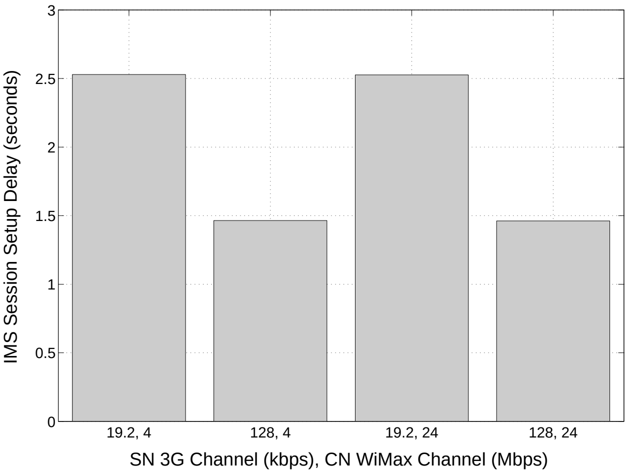


Fig. 8. IMS session setup delay for various channel rates when the SN is in a 3G network and the CN is in a WiMax network for a fixed signaling arrival rate λ and frame error probability p.

7.3 Arrival Rate Effects on the IMS Signaling Delay

Our second experiment analyzes the effects of varying the IMS signaling arrival rate λ in packets per second on the IMS registration and session setup signaling delay. The frame error probability p is fixed at 0.02. Results are calculated for arrival rates λ of 4, 9, 15, 21, and 24 packets per second. The 3G and WiMax networks use 128 kbps and 24 Mbps channel rates, respectively. These results also analyze the effects of interworking architectures on

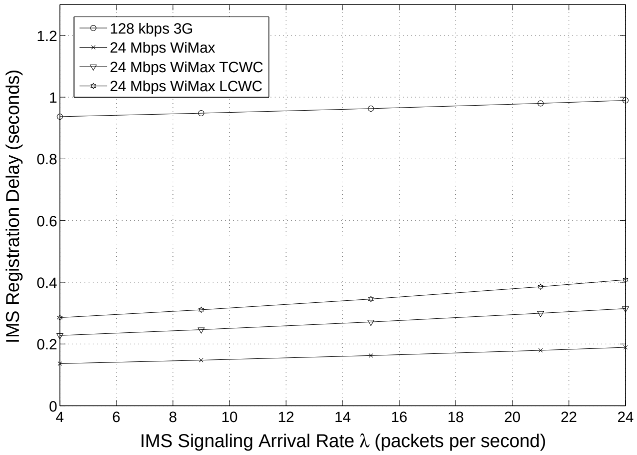


Fig. 9. The effects of changing the arrival rate λ on IMS registration delay for a 128 kbps 3G network and a 24 Mbps WiMax network for fixed frame error probability p.

the IMS registration signaling delay for different arrival rates.

Figure 9 shows IMS registration delay in seconds versus varying IMS signaling arrival rates for a 128 kbps 3G network and a 24 Mbps WiMax network for TCWC and LCWC architectures. The figure shows that the IMS registration signaling delay increases gradually with increasing arrival rate. This increase is in accordance with the queueing theory phenomenon where increased arrival rates result in increased network queue sizes and increased packet queueing time. Figure 9 also shows that the IMS registration signaling delay in the TCWC architecture is lower than in the LCWC architecture. The effects of changing the signaling arrival rate on IMS session establishment follows a similar trend and are omitted for brevity. However, there is a greater increase in the IMS session setup delay with increasing arrival rates as compared to the IMS registration delay.

7.4 Frame Error Probability Effects on IMS Signaling Delay

Our third experiment analyzes the effects of frame error probability p on the IMS signaling delay. The arrival rate λ is fixed at 9 packets per second. We examine frame error probability p values of 0.01, 0.02, 0.05, 0.1, and 0.2 for channel rates of 128 kbps and 24 Mbps for 3G and WiMax networks, respectively.

Figure 10 depicts the IMS registration signaling delay in seconds versus varying frame error probabilities for a 128 kbps 3G network and a 24 Mbps WiMax network for TCWC and LCWC architectures. The figure shows that the IMS registration signaling delay in 3G networks increases gradually as frame error probability increases, whereas the frame error probability has negligible effect on the IMS registration signaling delay for WiMax networks. The IMS registration delay for 3G networks is the same for both TCWC and LCWC interworking architectures due to identical additional network nodes in the path from the UE to the P-CSCF. However, for the WiMax network, the IMS registration delay for the TCWC architecture is less than the LCWC architecture due to different network nodes along the path from the UE to the P-CSCF.

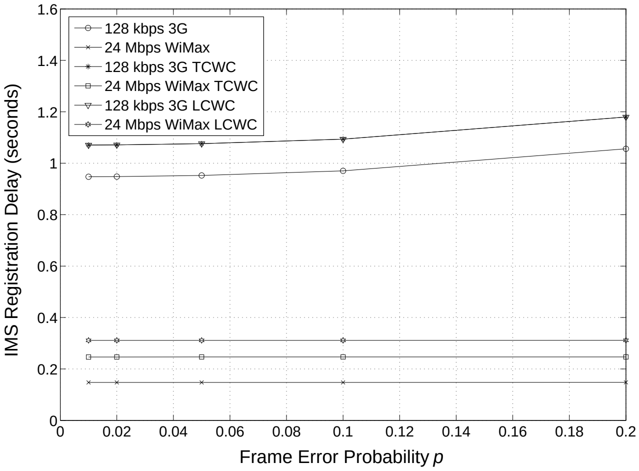


Fig. 10. The effect of varying frame error probability p on the IMS registration signaling delay for 128 kbps 3G and 24 Mbps WiMax networks with a fixed signaling arrival rate λ.

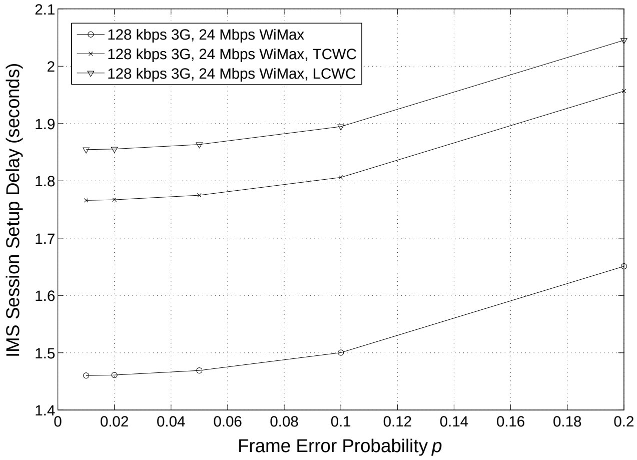


Fig. 11. The effect of varying frame error probability p on the IMS session setup delay when the SN is in a 128 kbps 3G network and the CN is in a 24 Mbps WiMax network for a fixed signaling arrival rate λ.

Figure 11 depicts the IMS session setup delay in seconds versus varying frame error probabilities when the SN is in a 128 kbps 3G network and the CN is in a 24 Mbps WiMax network for TCWC and LCWC architectures and a fixed signal arrival rate λ. Results show that the IMS session setup delay increases slowly with increasing frame error probability. Varying the frame error probability affects IMS session setup delay more than the IMS registration process because IMS session setup requires a larger number of exchanged signaling messages. The results also show that the IMS session setup delay in the TCWC architecture is lower than the delay in the LCWC architecture.

7.5 Numerical Results Summary and Analysis

In this subsection, we provide a summary and in depth analysis of our numerical results. Results show that the 3G channel rate has the most affect on the IMS signaling delay ((3) and (4)). Furthermore, the IMS signaling delay in WiMax networks is significantly lower than in 3G networks due to the high WiMax channel rates, and thus the transmission delay becomes negligible ((2) and (16)). Varying the WiMax channel rate has negligible effect on the IMS signaling delay because of our assumption that the inter-frame time and frame duration is independent of the WiMax channel rate. If we relax this assumption and assume that the inter-frame times and frame durations are dependent on the WiMax channel rate, the resulting IMS signaling delay will vary.

Increasing signaling arrival rate increases the IMS signaling delay due to the queueing theory phenomenon that an increase in arrival rate results in an increase in the number of queued packets, thus causing an increase in packet waiting time ((9) and (10)). Increased arrival rate affects the IMS session setup delay more than the IMS registration delay because IMS session establishment requires a larger number of exchanged messages as compared to IMS registration signaling (Figure 4 and Figure 5). In general, increasing arrival rates have the most impact on signaling protocols that require a large number of exchanged messages and will result in network congestion.

Overall, the IMS signaling delay increases with increased frame error probability, but these affects are more pronounced in 3G networks than in WiMax networks. Increased frame error probability has less affect on WiMax networks due to lost packet retransmission at high channel rates. Increased frame error probability has a more pronounced affect at lower channel rates due to more costly lost packet retransmissions.

Results show that the WiMax-3G interworking architectures contribute significantly to the total IMS signaling delay. Therefore, careful consideration must be taken in WiMax-3G interworking architecture design so as to minimize the negative effects on IMS signaling. The interworking architecture's effects are more prominent for IMS session establishment signaling than for the IMS registration signaling due to a larger number of exchanged messages in the IMS session establishment signaling as compared to the IMS registration signaling.

Results reveal that the IMS signaling delay in the TCWC architecture is always lower than in the LCWC architecture because in the LCWC architecture, the IMS signaling delay is mostly affected by the Internet dynamics. The Internet's utilization and packet waiting times vary considerably over time. The delays used in our LCWC interworking architecture analysis assume a fixed specific Internet utilization, however utilization may increase or decrease drastically with heavy or light traffic, respectively. In general, the IMS signaling delay in LCWC interworking architectures is never less than in TCWC interworking architectures. The TCWC interworking architecture offers predictable signaling delay because Internet dynamics are not involved. These results verify our assertion that TCWC interworking architectures can support QoS guarantees for network traffic flows.

Thus, we conclude that a tightly coupled paradigm can more tightly restrict IMS signaling delays to reasonable limits. However, tightly coupled architecture deployment requires more effort than loosely coupled architecture deployment, and hence a definite tradeoff exists between performance efficiency and implementation cost. Moreover, we conclude that SIP-based signaling is well suited for IMS registration and IMS session establishment procedures because acceptable signaling delays (in most cases) enable network operators to provide reasonable QoS support. Furthermore, our results support increasing WiMax network coverage to provide higher data rates as well as lower IMS signaling delays and WiMax-3G interworking architecture deployment with IMS infrastructure support.

7.6 Numerical Results Verification

In order to verify our numerical results presented in the previous subsection, we consider a specific 3G model closely following the 3GPP standards. However, due to particular dependencies, we do not consider several components. Since the processing delay is highly dependent on system parameters, we do not consider the processing delay in our verification. In addition, queueing delay is dependent on network conditions such as congestion and network source arrival rates, which can fluctuate over time. Moreover, the processing and queueing delays are nearly independent of the specific transmission delay model and these delays will remain constant for different 3G transmission delay models.

We compare the transmission delay results for a generic 3G model (used in our presented numerical results) and a model based on the CDMA2000 EV-DO rev. A standard with DRC 1, which corresponds to poor channel condition. Table 4 shows the transmission delays for the EV-DO (both FL and RL) and the generic 3G model for selected SIP messages. These values reveal that the individual SIP message transmission delays correspond closely for both models. The EV-DO RL transmission delay is less than the EV-DO FL transmission delay because of the additional three subframe interlacing scheme used in RL and because we have omitted the RL subframe synchronization delay [42].

Table 5 depicts the total IMS registration and IMS session establishment delay values when both the SN

TABLE 4
CDMA2000 forward link delay D_{EV}^{FL}, CDMA2000 reverse link delay D_{EV}^{RL} and 3G delay D_{3G} for SIP signaling messages

SIP Message	D_{EV}^{FL} (ms)	D_{EV}^{RL} (ms)	D_{3G} (ms)
SIP INVITE	249.6	236	260.9
SIP REGISTER	117.6	92.1	140.3
SESSION PROGRESS	117.6	92.1	140.3
SIP 180 RINGING	117.6	92.1	140.3
SIP PRACK	117.6	92.1	140.3
SIP 100 TRYING	117.6	92.1	140.3
SIP UPDATE	117.6	92.1	140.3
SIP 200 OK	111	95.7	120.2
SIP SUBSCRIBE	111	95.7	120.2
SIP NOTIFY	111	95.7	120.2
401 UNAUTHORIZED	111	95.7	120.2
SIP ACK	111	95.7	100.1

TABLE 5
IMS registration and IMS session setup delays for the generic 3G and CDMA2000 EV-DO models

IMS Signaling Procedure	D_{EVDO} (ms)	D_{3G} (ms)
IMS Registration	819.4	1001.8
IMS Session Setup	2776.7	3367.2

and CN are in a 3G/CDMA2000 EV-DO wireless system for the generic 3G model and the EV-DO model. These results verify the correctness of our numerical results because the IMS registration and IMS session establishment transmission delays for both models closely correspond even though the EV-DO model's transmission delays are consistently less than the generic 3G model transmission delays. It is important to note that DRC 14 transmission delays (corresponding to good channel condition) would be much lower than DRC 1 transmission delays. Thus, we can conclude that the generic 3G and WiMax models give an upper bound on the IMS registration and IMS session establishment delays when compared to the specific 3G and WiMax models. Intuitively, specific 3G and WiMax models can operate at various data transmission rates and frame length combinations based on the wireless channel conditions to minimize the FER, which decreases the total delay.

In addition, we verify our numerical results using simulation. Our IMS session setup delay corresponds closely with the numerical and simulation results presented in [42] and correspond closely with the results obtained from a detailed WiMax simulation model implemented with the network simulator 2 (ns-2) [26]. For completeness, we describe the ns-2 WiMax implementation. The ns-2 WiMAX module focuses on the WiMax MAC protocol. The ns-2 WiMax module implements the IEEE 802.16 point-to-multipoint (PMP) mode (which allows one WBSC to service multiple UEs concurrently) and WiMax features such as CS, CPS, and PHY (Section 3).

The PHY layer of the ns-2 WiMax module implements OFDMA. The ns-2 `Traffic Generating Agent` (TGA) is an application level traffic generator that generates voice over IP (VoIP), MPEG, FTP, and HTTP traffic. The TGA traffic is classified into five different types of WiMax service: the UGS, rtPS, ertPS, nrtPS, and BE, each with an associated priority (Table 1). The TGA packets are transferred to different types of priority queues according to their service types by using the CS layer SFID-CID mapping mechanism. The data packets in these queues are treated as MSDUs and are passed to the WiMAX module in a round robin manner.

The MAC management component initiates the ranging process to enter the WiMAX system or to transmit the MSDUs according to the scheduled time obtained from the UL-MAP (Section 3). The ns-2 `Network Interface` adds a propagation delay and then broadcasts the MSDUs using the air interface. The ns-2 `Channel` object uses the `WirelessPhy` class. The WiMAX module also receives packets via the air interface from other nodes. The WiMax module determines whether or not the received packet is a control packet. The MAC management object takes the corresponding action in case of a control packet, otherwise the MAC management object passes the packet to the `Link Layer` (LL) object after defragmentation. The LL in turn passes the packet to the TGA.

[26] provides simulation parameter details. The simulation results revealed that the WiMax MAC delay increased with increased WiMax UEs (and hence increased arrival rates). The results from the publicly available WiMax module for ns-2 [46] are also close to our presented results, however for completeness we highlight results for WiMax QoS differentiation. [47] showed via ns-2 simulations that the throughput and delay vary for each service class (Table 1) as the number of UEs increases. The throughput and delay values for UGS are not affected by an increasing number of users. However, other classes (specifically the BE service) are significantly affected by an increasing number of UEs. Similarly, packet loss rate for the UGS service class remains almost unaffected, whereas the other service classes progressively lose more packets with an increasing number of UEs. Thus, these ns-2 simulations verify our presented numerical results.

Our numerical results closely reflect the 3G, WiMax, and IMS prototypes due to careful selection of parameters' values from the literature. This obviates the need for rigorous and demanding simulation for numerical results verification. Also, our numerical analysis is highly flexible and network operators can specialize the parameter values to reflect the actual network implementation to obtain network implementation specific results. Additionally, our analytical model is beneficial in initial network design stages enabling network engineers to obtain upper bound estimates on signaling delays.

8 Future Research Directions

In this section, we give future research directions and propose several interesting research problems related to our work.

Our analysis and numerical results (Section 7) reveal that the IMS registration procedure requires substantial network resources and significant associated delay. For an IMS UE (terminal) moving at high speeds and frequently crossing several core IMS networks, the IMS registration process for each IMS network would consume substantial network resources. This use of network resources may be unjustified particularly if the IMS UE is idle and not involved in any active IMS sessions. Thus, there is a need for intelligent techniques to reduce the IMS registration overhead.

We propose a *lightweight IMS registration* strategy that dictates when IMS UEs should register to a new IMS core network as they move away from the originally registered core IMS network. We propose the addition of a new component, the *Local IMS Register* (LIR), to the core IMS network. The LIR would record visiting IMS user Public/Private User Identities in a local LIR database. The LIR would serve as a local anchor between the S-CSCF server and the IMS UE, and would eliminate the redundant registration cost for the complete IMS registration process each time the IMS UE moves to a region covered by a different core IMS network.

We suggest that the IMS registration process should be configured dynamically for IMS users according to the IMS user's activity profile. This dynamic IMS registration process selection would result in IMS users undergoing different IMS registration procedures depending on the specific type of IMS services used. For users utilizing high QoS IMS services, and thus require low IMS session setup time, an IMS UE may undergo a full IMS registration procedure whenever the user enters a different core IMS network. On the other hand, for users utilizing low QoS IMS services, and thus can tolerate larger IMS session setup time, it may not be efficient to undergo the full IMS registration process, but rather would be preferable to use a lightweight IMS registration process.

In our proposed lightweight IMS registration process, the IMS UE would send a "SIP Local IMS Register" message to the LIR. The SIP Local IMS Register message would contain the address of the old S-CSCF server. The LIR may also inform the old S-CSCF server that the IMS UE currently resides in the LIR's coverage area, and thus any subsequent calls to the IMS UE should be directed by the original S-CSCF server to the LIR, which would relay the data or signaling messages to the IMS UE. It would be interesting to verify that our proposed lightweight IMS registration using a cost analysis procedure [48] to quantify improvements in performance and reduction in network resources and signaling cost as compared to the standard IMS registration procedure.

In a WiMax-3G interworking architecture, each 3G cell may or may not overlap WiMax cells. IMS users that cannot be accommodated in the WiMax cells due to traffic overflow are transferred to 3G cells. It would be interesting to calculate the IMS users residence time distribution, the handoff traffic, the expected channel occupation time, and the IMS session incompletion probability for WiMax. These calculations require further research to investigate the IMS session WiMax to 3G handoff rules to maximize performance and minimize IMS session incompletion probabilities in WiMax-3G interworking architectures.

Since both the WiMax tightly and loosely coupled architectures have associated advantages and disadvantages, we propose a hybrid tightly and loosely coupled WiMax-3G interworking architecture (Hybrid Coupled WiMax-Cellular (HCWC)). The HCWC architecture would route the signaling and data traffic either through a tightly or loosely coupled path. The decision to route the signaling and data traffic to a particular path can be formulated as an optimization problem with an objective function to minimize delay and/or cost. The optimization constraints can be the tolerable delay specified by the user (for the IMS registration and session setup processes) and the network cost constraints. The route optimization in HCWC would result in a robust WiMax-3G interworking architecture capable of delivering the desired service in all network conditions. However, the downside of the HCWC would be the additional cost and complexity of hybrid network formation.

The HCWC route optimization problem can be extended to perform *dynamic optimizations* (an optimization that adapts to changing network conditions) using dynamic profiling [49], [50]. The profiling modules at network nodes would gather profiling statistics, such as queue utilization, wireless channel condition, and packet loss. The profiling modules would transmit these statistics to the optimization module to perform the dynamic optimization based on the profiling statistics [50]. To the best of our knowledge, no previous work addresses dynamic optimizations for WiMax-3G interworking architectures with IMS support.

Although our WiMax-3G interworking architectures (specifically TCWC) are intended to provide QoS support mechanisms, QoS is not feasible without a rigorous admission control mechanism. In the future, it may be interesting to model a semi-Markov decision process (SMDP) based joint WiMax-3G session admission controller [51] subject to QoS constraints for multiple traffic classes. A joint call admission controller that is cognizant of the state of the WiMax and 3G networks (i.e. the number of sessions for each traffic class in the two networks) would be feasible in our proposed TCWC architecture.

Finally, an analytical model derivation for WiMax transmission delay that closely follows WiMax specifications and the analysis of IEEE 802.16 physical layer adaptive modulation capability and multi-rate data encoding capability for real-time IMS applications is the focus of our future work.

9 Conclusion

In this paper, we analyzed the SIP-based IMS registration and session setup signaling delay in 3G and WiMax access networks. We also analyzed the effects of novel WiMax-3G interworking architectures on the IMS signaling delay. Our numerical analysis revealed that the tightly coupled architectures have lower IMS signaling delays than loosely coupled architectures. It can be concluded that a tightly coupled system is more appropriate for restricting the IMS signaling delays to acceptable limits. However, the tightly coupled architecture deployment requires more effort than the loosely coupled architecture deployment, and hence a definite tradeoff exists between performance efficiency and implementation cost. Numerical data analysis indicated that the IMS registration and session setup signaling delay in WiMax networks is much less than the IMS registration and session setup signaling delay in 3G networks. Our numerical results encourage the deployment of WiMax-3G interworking architectures with the IMS infrastructure support.

Acknowledgments

The initial part of this work was supported by Bell Canada and the Natural Sciences and Engineering Research Council of Canada (NSERC) under the supervision of Dr. Vincent Wong.

References

[1] R. Price, C. Bormann, J. Christoffersson, H. Hannu, Z. Liu, and J. Rosenberg, "Signaling Compression (SigComp)," RFC 3320, January 2003.

[2] P. Taaghol, A. K. Salkintzis, and J. Iyer, "Seamless Integration of Mobile WiMAX in 3GPP Networks," *IEEE Communications Magazine*, vol. 46, no. 10, pp. 74–85, October 2008.

[3] "WiMAX Forum Announces Launch of Global Roaming Program," in *WiMAX Forum*, 2009. [Online]. Available: http://www.wimaxforum.org/node/432

[4] F. Xu, L. Zhang, and Z. Zhou, "Interworking of WiMax and 3GPP Networks based on IMS," *IEEE Communications Magazine*, vol. 45, no. 3, pp. 144–150, March 2007.

[5] G. Ruggeri, A. Iera, and S. Polito, "802.11-Based Wireless LAN and UMTS Interworking: Requirements, Proposed Solutions and Open Issues," *Elsevier Computer Networks*, vol. 47, no. 2, pp. 151–166, February 2005.

[6] C. Liu and C. Zhou, "An Improved Interworking Architecture for UMTS-WLAN Tight Coupling," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC'05)*, Atlanta, Georgia, March 2005.

[7] ——, "HCRAS: A Novel Hybrid Internetworking Architecture between WLAN and UMTS Cellular Networks," in *Proc. of IEEE Consumer Communications and Networking Conference (CCNC'05)*, Las Vegas, Nevada, January 2005.

[8] H. Mahmood and B. Gage, "An Architecture for Integrating CDMA2000 and 802.11 WLAN Networks," in *Proc. of IEEE Vehicular Technology Conference (VTC'03)*, Orlando, Florida, October 2003.

[9] Q. Nguyen-Vuong, L. Fiat, and N. Agoulmine, "An Architecture for UMTS-WIMAX Interworking," in *Proc. of 1st International Workshop on Broadband Convergence Networks (BcN'06)*, Vancouver, Canada, April 2006.

[10] D. Kim and A. Ganz, "Architecture for 3G and 802.16 Wireless Networks Integration with QoS Support," in *Proc. of International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine'05)*, Orlando, Florida, August 2005.

[11] H.-T. Lin, Y.-Y. Lin, W.-R. Chang, and R.-S. Cheng, "An Integrated WiMAX/WiFi Architecture with QoS Consistency over Broadband Wireless Networks," in *Proc. of IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, Nevada, January 2009.

[12] A. Munir and V. Wong, "Interworking Architectures for IP Multimedia Subsystems," *ACM/Springer Journal on Mobile Networks and Applications*, vol. 12, no. 5, pp. 296–308, December 2007.

[13] M. Melnyk and A. Jukan, "On Signaling Efficiency for Call Setup in all-IP Wireless Networks," in *Proc. of IEEE International Conference on Communications (ICC'06)*, Istanbul, Turkey, June 2006.

[14] A. Munir, "Analysis of SIP-based IMS Session Establishment Signaling for WiMax-3G Networks," in *Proc. of IEEE International Conference on Networking and Services (ICNS'08)*, Gosier, Guadeloupe, March 2008.

[15] H. Fathi, S. Chakraborty, and R. Prasad, "Optimization of SIP Session Setup Delay for VoIP in 3G Wireless Networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 9, pp. 1121–1132, September 2006.

[16] F. Xu, L. Zhang, and Z. Zhou, "Interworking of Wimax and 3GPP networks based on IMS," *IEEE Communications Magazine*, vol. 45, no. 3, pp. 144–150, March 2007.

[17] N. Rajagopal and M. Devetsikiotis, "Modeling and Optimization for the Design of IMS Networks," in *Proc. of IEEE Annual Simulation Symposium (ANSS'06)*, Huntsville, Alabama, April 2006.

[18] A. Anzaloni, M. Listanti, I. Petrilli, and D. Magri, "Performance Study of IMS Authentication Procedures in Mobile 3G Networks," in *Proc. of ACM International Conference on Wireless Communications and Mobile Computing (IWCMC'07)*, Honolulu, Hawaii, August 2007.

[19] W. Wu, N. Banerjee, K. Basu, and S. Das, "SIP-based Vertical Handoff between WWANs and WLANs," *IEEE Wireless Communications*, vol. 12, no. 3, pp. 66–72, June 2005.

[20] P802.16Rev2/D9, "Part 16: Air Interface for Broadband Wireless Access Systems (Revision of IEEE Std 802.16-2004 and consolidates material from IEEE Std 802.16e-2005, IEEE Std 802.16-2004/Cor1-2005, IEEE Std 802.16f-2005, and IEEE Std 802.16g-2007)," The IEEE 802.16 Working Group on Broadband Wireless Access Standards, January 2009.

[21] F. Wang, A. Ghosh, C. Sankaran, P. J. Fleming, F. Hsieh, and S. J. Benes, "Mobile WiMAX Systems: Performance and Evolution," *IEEE Communications Magazine*, vol. 46, no. 10, pp. 41–49, October 2008.

[22] A. Lera, A. Molinaro, and S. Pizzi, "Channel-Aware Scheduling for QoS and Fairness Provisioning in IEEE 802.16/WiMAX Broadband Wireless Access Systems," *IEEE Network*, vol. 21, no. 5, pp. 34–41, September-October 2007.

[23] S. Ahmadi, "An Overview of Next-Generation Mobile WiMAX Technology," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 84–98, June 2009.

[24] 802.16m 09/0034r1, "IEEE 802.16m System Description Document (SDD)," The IEEE 802.16 Task Group, September 2009. [Online]. Available: http://wirelessman.org/tgm/core.html#09_0034

[25] A. Altaf, M. Y. Javed, and A. Ahmed, "Security Enhancements for Privacy and Key Management Protocol in IEEE 802.16e-2005," in *Proc. of IEEE ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD)*, Phuket, Thailand, August 2008.

[26] J. Chen, C. Wang, F. Tsai, C. Chang, S. Liu, J. Guo, W. Lien, J. Sum, and C. Hung, "The Design and Implementation of WiMax Module for ns-2 Simulator," in *Proc. of ACM Workshop on ns-2: the IP Network Simulator (WNS2'06)*, Pisa, Italy, October 2006.

[27] B.-H. Kim, J. Yun, Y. Hur, C. So-In, R. Jain, and A.-K. Al Tamimi, "Capacity Estimation and TCP Performance Enhancement over Mobile WiMAX Networks," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 132–141, June 2009.

[28] 3GPP, "3GPP System to Wireless Local Area Network (WLAN) Interworking; System Description," TS 23.234 (v7.2.0), June 2006.

[29] C. Hoymann, K. Klagges, and M. Schinnenburg, "Multihop Communication in Relay Enhanced IEEE 802.16 Networks," in *Proc. of 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Helsinki, Finland, September 2006.

[30] 3GPP, "Signalling flows for the IP multimedia call control based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3 (Release 5)," TS 24.228 (v5.15.0), September 2006.

[31] G. Camarillo and M.-A. Garcia-Martin, *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds*. John Wiley and Sons, 2004.

[32] J. Kurose and K. Ross, *Computer Networking, A Top-Down Approach Featuring the Internet*. Addison-Wesley, 2003.

[33] 3GPP, "Radio Link Protocol (RLP)for circuit switched bearer and teleservices (Release 7)," TS 24.022 (V7.0.0), June 2007.

[34] S. Das, E.Lee, K. Basu, and S. Sen, "Performance Optimization of VoIP Calls over Wireless Links using H.323 Protocol," *IEEE Transactions on Computers*, vol. 52, no. 6, pp. 742–752, June 2003.

[35] N. Banerjee, W. Wu, K. Basu, and S. Das, "Analysis of SIP-based Mobility Management in 4G Wireless Networks," *Elsevier Computer Communications*, vol. 27, no. 8, pp. 697–707, May 2004.

[36] B. Lampson, V. Srinivasan, and G. Varghese, "IP Lookups Using Multiway and Multicolumn Search," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 324–334, June 1999.

[37] C.-M. Huang and J.-W. Li, "Efficient and Provably Secure IP Multimedia Subsystem Authentication for UMTS," *The Computer Journal*, vol. 50, no. 6, pp. 739–757, October 2007.

[38] E. Gelenbe and G. Pujolle, *Introduction to Queueing Networks*. John Wiley and Sons, 1998.

[39] J. Medhi, *Stochastic Models in Queueing Theory*. Academic Press, An imprint of Elsevier Science, 2003.

[40] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*. John Wiley and Sons, 2008.

[41] Q. Bi, P. Chen, Y. Yang, and Q. Zhang, "An Analysis of VoIP Service Using 1xEV-DO Revision A System," *IEEE Journal on Slected Areas in Communications*, vol. 24, no. 1, pp. 36–45, January 2006.

[42] M. Melnyk, A. Jukan, and C. Polychronopoulos, "A Cross-Layer Analysis of Session Setup Delay in IP Multimedia Subsystem (IMS) with EV-DO Wireless Transmission," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 869–881, June 2007.

[43] 3GPP2, "cdma2000 High Rate Packet Data Air Interface Specification," 3GPP2 C.S0024-B (v2.0), March 2007.

[44] S. Mohanty and J. Xie, "Performance Analysis of a Novel Architecture to Integrate Heterogeneous Wireless Systems," *Elsevier Computer Networks*, vol. 51, no. 4, pp. 1095–1105, March 2007.

[45] H. Fathi, S. Chakraborty, and R. Prasad, "On SIP Session Setup Delay for VoIP Services over Correlated Fading Channels," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 1, pp. 286–295, January 2006.

[46] NIST, "Advanced Network Technologies Division: Seamless and Secure Mobility," in *National Institute of Standards and Technology*, Gaithersburg, Maryland, September 2009. [Online]. Available: http://www.antd.nist.gov/seamlessandsecure/download.html

[47] P. Neves, F. Fontes, J. Monteiro, S. Sargento, and T. M. Bohnert, "Quality of service differentiation support in WiMAX networks," in *Proc. of IEEE International Conference on Telecommunications (ICT)*, St. Petersburg, Russia, June 2008.

[48] J. Ho and I. F. Akyildiz, "Local Anchor Scheme for Reducing Signaling Costs in Personal Communications Networks," *IEEE/ACM Transactions on Networking*, vol. 4, no. 5, pp. 709–725, October 1996.

[49] S. Sridharan and S. Lysecky, "A First Step Towards Dynamic Profiling of Sensor-Based Systems," in *Proc. of IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, San Francisco, California, June 2008.

[50] A. Munir and A. Gordon-Ross, "An MDP-based Application Oriented Optimal Policy for Wireless Sensor Networks," in *Proc. of IEEE/ACM International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, Grenoble, France, October 2009.

[51] 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*, vol. 6, no. 1, pp. 126–139, January 2007.

Arslan Munir received his B.S. degree from the University of Engineering and Technology (UET), Lahore, Pakistan, in 2004, and his M.S. degree from the University of British Columbia (UBC), Vancouver, Canada, in 2007. He is currently working towards his Ph.D. degree in Electrical and Computer Engineering at the University of Florida (UF), Gainesville, Florida, USA. From 2007 to 2008, he worked as a software development engineer at Mentor Graphics in the Embedded Systems Division. He was the recipient of many academic awards including the Gold Medals for the best performance in Electrical Engineering and academic Roll of Honor. He served as a TPC member for IEEE ICNS'09 and ICNS'10 and is an active reviewer for various IEEE/ACM conferences including ISVLSI, GLSVLSI, HPCA, CODES+ISSS, and LCN. His current research interests include embedded systems, dynamic optimizations, and wireless networks. He is a student member of IEEE.

Ann Gordon-Ross is an Assistant Professor of Electrical and Computer Engineering at the University of Florida, Gainesville, Florida, USA. She received her B.S. and Ph.D. in Computer Science from the University of California, Riverside in 2000 and 2006, respectively. She is a member of the NSF Center for High-Performance Reconfigurable Computing (CHREC) at the University of Florida. Her research interests include embedded systems, computer engineering, low-power design, reconfigurable computing, platform design, dynamic optimizations, hardware design, real-time systems, computer architecture, and multi-core platforms.