Quality of Service of Next Generation Wireless Networks

Ms Colette Consani
Data Network Architecture Laboratory
University of Cape Town

Abstract—Quality of Service (QoS) has always been an important item to the telecommunications industry. Performance of calls are monitored to ensure that every word can be clearly heard. In IP networks, QoS has only become a notable issue since the introduction of Voice over IP (VoIP). In order to achieve a voice QoS level in line with the telecommunications industry, VoIP IP packets have to arrive at their destination in a timely manner with very few packets dropped along the way.

Current wireless IP access networks, such as IEEE 802.11 have failed to entice consumers to use multimedia services since they only support best effort services over a limited coverage area. In order to improve provider revenue and for next generation wireless networks to become successful, QoS should be of crucial concern.

Since consumers desire ubiquitous access to these next generation multi-service heterogeneous networks, we propose a framework for developing QoS mechanisms at a Radio Network Controller (RNC), straddling multiple access technologies, to provide guaranteed QoS.

The proposed framework relies on knowledge of how the MAC and physical layers of the enabling access technologies work. Thus we discuss the development of cross-layer QoS performance models for two current popular wireless access technologies, namely IEEE 802.11e and UMTS.

Index Terms—Cross-layer Performance Model, IEEE 802.11e, Next Generation Network Quality of Service, UMTS.

I. INTRODUCTION

With the increasing popularity of VoIP, QoS for real-time traffic has become an important characteristic for IP-based networks. The use of multiple different services at the same time on a common network introduces the need to implement mechanisms to maintain the QoS for each application.

The term QoS can either refer to qualitative or quantitative characteristics. Qualitative characteristics, such as audio clarity, picture detail, color and audio synchronization for video streaming are characteristics required by the user. But these characteristics are subjective and difficult to measure and analyze. On the other hand we have quantitative QoS characteristics, such as throughput, delay, jitter and error rate. These characteristics are based on the particular technology employed, and usually measurements can be recorded from a real-world network or simulator and compared to the level that a user requires.

One of the main design goals of Next Generation Networks (NGNs) is to ensure end-to-end QoS [1]. End-to-end QoS can be defined as the network related performance of an application as rated by the end user.

The author would like to thank OPNET, Telkom SA, THRIP, NRF and Siemans for supporting this research project.

NGNs will be packet-based networks consisting of a common Core Network (CN) connecting heterogeneous access networks that, in the case of overlapping or adjacent wireless cells, could be centrally controlled by a common RNC. The users of these NGNs will be able to seamlessly roam between the access networks by their access devices transparently requesting vertical handoffs and will be able to access the same services no matter which access technology they are using. It may also be possible for the user to be connected to more than one access network simultaneously. The users will also have the ability of communicating with other users connected to different access network types or even accessing external networks such as the Internet.

The current trend is to have IP as the enabling technology for the core network of next generation networks. The UMTS standard, release 6, seems to be a good candidate technology for the next generation core network since it is composed of an all-IP core, and proposals have been considered to allow vertical handoffs between various technologies such as IEEE 802.11 & UMTS [2][3][4].

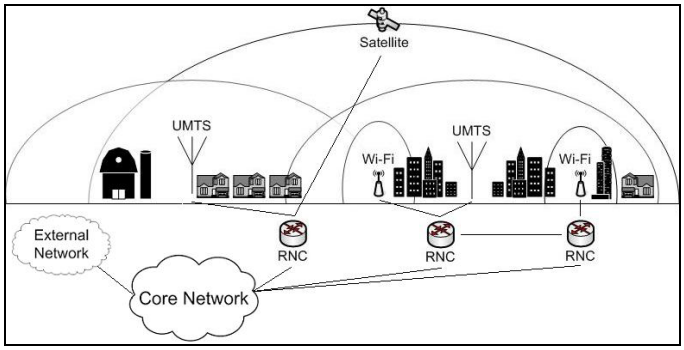

Figure 1: Example Next Generation Wireless Network

In section II we introduce some common approaches to ensuring end-to-end QoS. We go on to describe our proposed framework for QoS mechanisms in heterogeneous networks. In section IV we introduce two common wireless access technologies, namely IEEE 802.11e and UMTS In section V we describe the theory behind developing performance models needed for our framework.

II. END-TO-END QOS

When a link in the end-to-end path becomes congested, the expected QoS, as seen by the end user, may not be maintained.

A popular response to counteract such QoS issues is to avoid congestion in the first place by constructing networks with vast over-provisions of bandwidth [5]. However, when a network boasts a congested wireless link as part of the obligatory end-to-end path, trying to install excess bandwidth is not always a feasible solution. This is because wireless network technologies share a limited wireless

channel, and in many cases critical real-time applications cannot simply trust this uncertain approach of avoiding congestion due to the bursty nature of IP traffic.

The Third Generation Partnership Project (3GPP) suggests the well-known QoS mechanism, Differentiated Services (DiffServ), to be used in the core network (CN) to achieve QoS [6].

DiffServ resides in the network layer, see fig 2, and has been used in wired networks for many years. Thus many existing networks and devices offer support for DiffServ. The protocol scales very well since it does not explicitly reserve bandwidth for each traffic flow. It rather allocates different QoS parameters to different predefined classes and then prioritizes traffic throughout the network based on its class.

The downfall of DiffServ is that it merely offers relative performance, and not absolute performance to a traffic flow. When one or more links of the network become congested, DiffServ may fail to maintain the required QoS for a traffic flow as specified by their Service Level Agreements (SLAs).

In NGNs, the wireless access networks are segments of the system where traffic for multiple mobile users will converge to travel through a relatively low bandwidth link. Even though a DiffServ implementation may provide sufficient QoS services for wired CN, it is suggested that it might not perform adequately for heterogeneous networks containing wireless links [7].

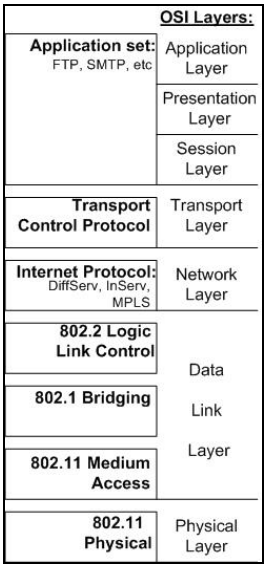

Figure 2: The OSI Layer Model with example layer-specific protocols

In DiffServ merely prioritization is offered. QoS mechanisms for the Medium Access Control (MAC) and physical layers have been neglected because data rates of wired networks are very high and the physical layer's error rate very low.

However, due to the unpredictability of error prone wireless channel and with the capacity of the wireless access networks being the bottleneck in the implementation of an end-to-end QoS mechanism, we should investigate how to complement the DiffServ CN QoS mechanism by adding further DiffServ aware QoS mechanisms at the RNCs in order to protect the QoS of the entire end-to-end network.

Note, that we suggest adding *DiffServ aware* QoS mechanisms. We do this to ensure that end-to-end Classes of Service (CoS) is supported. This makes provision for graceful degradation of services should the network become overloaded. In table 1 we show the classes offered by two popular wireless technologies, IEEE 802.11e and UMTS, and show how mapping between the different classes can take place.

DiffServ	UMTS	802.11e
Expedited Forwarding	Conversational	Voice
Assured Forwarding	Streaming	Video
	Interactive	Best Effort
Best Effort	Background	Background

Table 1: Example of mapping classes between DiffServ, UMTS & IEEE 802.11e.

It is common for networks to employ admission control at the edge of the CN to avoid overloading of the network. However in NGNs that contain wireless access links, admission control for the access networks should also be considered.

As an example, imagine a user that is connected to an uncongested CN and an uncongested IEEE 802.11e access network. If he requests to begin a videoconference, this service would usually be permitted. But in a heterogeneous network, where the overlapping UMTS macro cell is congested, and the probability is high that the user will move out of the coverage area of the 802.11e network, this service should not be allowed by the admission control mechanism since the probability of the call being dropped is too high.

An admission control mechanism decides whether to accept or reject a new connection depending on the impact it will have on the system. It is widely accepted that, from an end user's perspective, disallowing (blocking) a connection request is better than dropping a connection once transmission is in progress [8]

The job of the admission control mechanism is to keep the network load and interference low enough to ensure that no active connections are dropped. However, if the mechanism is not economical enough, too few calls will be admitted which can cause a decrease in the network utilization.

So the main challenge in the design of an efficient admission control scheme is to balance the conflicting requirement of having to maximize network utilization, while still keeping the call dropping probability very low.

Another way to avoid congestion is to manage the network's resources intelligently. Resource management is usually specific to the particular technology employed. However, as we noted earlier, in NGNs a user may be connected to multiple heterogeneous access networks simultaneously. Thus there is a need for heterogeneous access resources to be managed as a common resource and traffic to be allocated dynamically to these resources.

III. Framework for a Generalized QoS Mechanism

In multi-service heterogeneous networks, designing QoS schemes poses new challenges. As we noted above, the

NGN's resource manager will need to take into account multiple different technologies and the corresponding resources that they have to offer before allocating resources to a connection.

The admission control scheme is also more complex. Not only must it consider the impact on the admitting network, it should also consider the impact the connection may have on nearby access networks should a vertical handoff be required.

So we see that QoS mechanisms need to be able to predict the impact of an action on the QoS of active connections. In next generation networks, before proceeding with a proposed admission or transmission, the QoS impact of the action on multiple diverse networks, including the core network, may need to be queried.

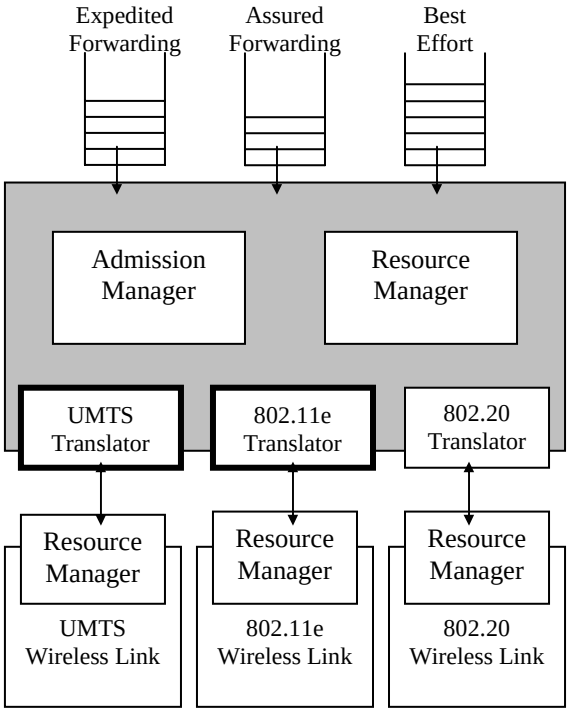

Figure 3: Generalized RNC QoS Framework

In figure 3 is an overview of how a RNC QoS mechanism can be developed. We see that traffic arrives at the RNC, categorized into their DiffServ classes. In order for the Admission Manager (AM) and Resource Manager (RNC-RM) to make informed decisions about possible actions, they communicate with the Resource Manager (T-RM) of the particular technology to find out if there is enough resource to perform the action with effecting the QoS of the other active connections. The request made by the AM or RNC-RM should be identical for all access networks. However, the specific translator of the access network will request the status of the wireless network from the appropriate T-RM, and then translate the reply into a general format that the AM or RNC-RM can understand. In order to generate intelligent replies, the translator needs to understand the workings of the access technology's QoS mechanisms. It should also be able to map the DiffServ categorized traffic to the QoS classes specified by the wireless access technology, as shown in the example in table 1.

Thus we see that in order to develop a generalized RNC QoS mechanism we need to have a good understanding of the QoS offerings of the different access technologies in order to develop the needed translators. In the next two sections we give an overview of IEEE 802.11e and UMTS, and suggest a method to model the behaviour of their QoS mechanisms.

IV. Overview of Wireless Access Technologies

A. IEEE 802.11x

The original 802.11 standard specifies two channel access methods, namely the mandatory Distributed Coordination Function (DCF) and the optional Point Coordination Function (PCF).

The DCF is based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). All stations can access the channel with equal probability and share it according to equal frame rate and not according to equal throughput. This offers no support for priority access to the channel for time-sensitive traffic.

The PCF is an optional mode that was introduced as an initial attempt at supporting time-sensitive traffic flows, using a contention free service. PCF splits the time into a contention-free period (CFP) and a contention period (CP). Only stations polled by the Point Coordinator (PC) may transmit during the CFP.

Even though the PCF can offer some sort of priority to an overloaded station, it cannot differentiate between traffic types or sources. Therefore, it cannot tell which stations have long queues of time-sensitive traffic, and which only hold best-effort traffic. From [9] we know that under high load, 802.11 does not offer an acceptable level of QoS for real-time traffic.

Another problem with PCF is that the PC has to contend with other stations to gain control of the wireless medium. Therefore, the starting time and length of the CFP can vary.

These downfalls led the IEEE to created Task Group e. The 802.11e draft standard [10] presents a new MAC scheme called the Hybrid Coordination Function (HCF). The HCF consists of two parts, namely the mandatory Enhanced Distributed Channel Access (EDCA) and the optional HCF Controlled Channel Access (HCCA).

As in PCF, a CP and CFP together form a superframe. During the CP, access to the wireless medium is managed using EDCF rules. Thereafter a CFP begins, where a HC sends QoS CF-Polls to stations.

EDCA is an extension of DCF and is a contention-bases channel access mechanism that introduces the concept of service differentiation among different Access Categories (ACs), see table 1.

Within the station, each AC has its own queue and backoff mechanism, which is parameterized according to the specific AC. These vary the probability of the AC winning the channel contention. A higher priority AC will have a shorter Arbitration Frame Interface (AIFS), which is the additional time that it has to freeze its backoff counter while the medium is idle before beginning decrementing. Each AC also has its own minimum and maximum contention window (CW) size and persistence factor (PF).

Although EDCA specifies the tools to allow for prioritizing traffic, it does not specify values for AIFS, CW_{min}, CW_{max}, or PF for the different ACs. This allows for flexibility in the QoS offered by the system. However, it also means that the higher layer resource management is dependent on these vendor or operator assigned values.

The optional HCCA scheme offers contention-free channel access by using centralized polling and scheduling algorithms to allocate channel resources to the various ACs of the stations. Unlike the original 802.11, where it was not compulsory to have a PC, HCF must have a centralized Hybrid Controller (HC).

An important difference between the PC and the HC is that the HC has priority over all other stations in the WLAN. Thus the HC does not have to contend for control of the wireless medium, can initiate HCF access at any time, and stations can be guaranteed predictable transmission opportunity times.

As part of a new controlled contention mechanism, the HC maintains a summary of the queue lengths for each AC of each station. This information is sent to the HC by the stations via the new QoS control field during a Controlled Contention Period (CCP). The CCP begins when the HC sends a specific control frame instructing legacy stations remain silent until the end of the CCP. The control frame defines the number of controlled contention opportunities, as well as the subset of ACs that may submit requests for a transmission opportunity during the CFP.

Using the contention summary generated during CCPs, the HC determines which stations, including the AP, will be allocated transmission opportunities during CFP. When a station receives permission to transmit from the HC, the HC does not identify which AC should transmit data. It leaves the decision to the station. By decentralizing this decision HCF defines a scalable solution for maintaining the QoS of ACs.

B. UMTS

Universal Mobile Telecommunication System (UMTS) radio networks are based on Code Division Multiple Access (CDMA). By using codes to separate user data, multiple transmissions can occur during a single time instance.

One of the main proposed advantages of UMTS networks is that they will offer wireless access to multimedia services, over and above the voice calls and www services offered by 2G and 2.5G networks. The various applications, such as speech, video and interactive sessions, can have varying QoS requirements. Applications will be allocated to one of the QoS classes defined in the UMTS standards [11], see table 1 according to their QoS requirements.

UMTS offers a packet-switched service for IP traffic. IP data traffic is inherently bursty, and due to these bursts the system load varies significantly. This causes interference levels in the access network to be highly variable. When interference peaks occur that exceed a level that the power mechanism is able to handle, the Signal-to-Interference-Ratio (SIR) decreases to unacceptable values and bursts of errors occur in the transmission. The Automatic Repeat reQuest (ARQ) procedure, which is usually more effective than Forward Error Control (FEC) in handling long bursts of error, can retransmit the lost frames. But these retransmissions increase congestion and interference for other User Equipment (UE) sharing the same access network, therefore leading to further possible transmission errors and thus retransmissions. Particularly when the access network is close to saturation, this feedback loop of errors causing retransmissions causing errors can make the network behaviour unstable.

Direction	Proposed QoS Traffic Type	Transport Channel	Physical Channel
UL/DL	Streaming	DCH	DPDCH
DL	Interactive/ Background	DSCH	PDSCH
UL	Interactive/ Background	CPCH	PCPCH
DL	Short Interactive	FACH	SCCPCH
UL	Short Interactive	RACH	PRACH
DL	Broadcast Streaming	BCH	PCCPCH

Table 2: Overview of Transport and Physical Channels linked to the logical Dedicated Transport Channel

To maintain an acceptable balance between overall system capacity and the individual application's QoS the limited radio resources, which act as a bottleneck in the system, need to be managed intelligently. The radio interface of UMTS is highly adaptable and many transport channels are introduced which are mapped to different physical channels that can transmit simultaneously. Each one of the channels can have difference combinations of spreading factor, error control and code rate. The bandwidth and content protection required under varying channel conditions and system load can be maintained by dynamically varying these channel parameters.

Coverage and capacity are closely linked in UMTS access networks. Since UE can be located at varying distances from the node B, there is a possibility that the transmission of a user close to the node B will drown out the transmission of a user on the outskirts of the cell. So to ensure a coverage area in accordance with the planning, the system load level should be limited to ensure that the required transmitted power will be lower than the maximum transmitted power allowed and high enough to be received with an acceptable error rate.

But the ability to offer different levels of QoS to simultaneous multi-directional transmissions, while trying to optimize the overall system capacity and maintaining the coverage, makes the radio resource management implementation extremely intricate. From [12] we see that

"the complexity of detailed services definition and system parameters optimization has been moved out of specifications and left to UMTS vendors and operators"

So, similar to 802.11e, the selection of optimal channel parameters for all active physical channels, in order offer the required QoS to the different traffic classes and maintain overall system capacity, is left to the vendors and operators.

V. Modeling the Wireless Access Networks

A. Queuing Theory

Many multimedia applications are sensitive to end-to-end QoS and it is important to understand queue characteristics such as waiting time, service time, queue length and obviously queue blocking probability. It is well known that due to the limited capacity and error-prone nature of wireless networks, as the number of active connections increases, the network performance degrades dramatically. Thus to allow a QoS mechanism to evaluate whether the end-to-end QoS can be upheld if an admission is allowed, it is necessary to investigate the queuing model for wireless networks to obtain a relation between offered load and performance metrics such as average delay, variance of delay (jitter) and average throughput.

Each access network can be modeled as a queuing system that is characterized by an arrival process and a service time distribution. The distribution of the frame service time can be approximated by a discrete probability distribution because the smallest time unit in both UMTS and 802.11e is a time slot, although the duration of the two are different.

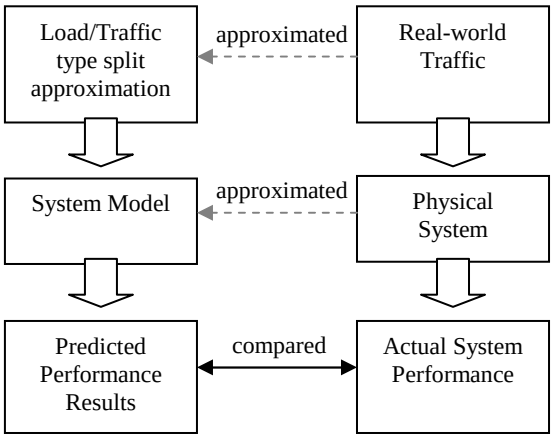

Fig 4: Relationship between real-world & queuing model

Understanding the nature of the traffic is critical for a good system design. Since the heterogeneous NGNs are not commonly implemented as yet, it is still unknown what the load and traffic type split in the system will be. If we were to develop approximate distributions for the traffic using traffic data from other systems such as the Internet or GPRS networks, we will very likely find that this model will be out of date very soon. The traffic on the NGNs will very likely change dramatically during the first few years that they become popular, since new services will come into existence and become more prevalent, while other services will become extinct.

Thus, we suggest that as an initial approximation, the IP traffic should be modeled as multiple classes with different arrival rates and packet sizes, estimated by the Poisson and Geometric distributions respectively. This traffic estimation is used widely in research, and has the benefit that the equations are relatively easy to manipulate.

B. System Models

In wired networks the QoS that is offered is mainly dependant on the way that resources are allocated. The remaining capacity of a network is usually easy to calculate and physical errors are negligible, making it relatively easy to predict QoS by studying the technology's MAC layer.

However, in wireless networks physical layer errors are not negligible and resource management strategies are not statically defined by the specifications. This implies that in order to predict the QoS offered to a transport layer connection, a cross-layer performance model of the technology should be developed to evaluate the various parameterizations for the QoS-enabling mechanisms.

To develop cross-layer performance models we consider the performance of individual connections. Given the current system load and the type of traffic that the connection is carrying, we want to be able to predict the delay and throughput this connection can expect.

As we discussed in the previous section, a system can be modeled using a queue where there are packets arriving at the IP layer that get transmitted to the receiver according to a certain service time. This service time is the time it takes for the MAC layer to process and schedule non-colliding frame access to the medium and to transmit the bits, taking into account that a frame may need to be transmitted multiple times due to channel transmission errors.

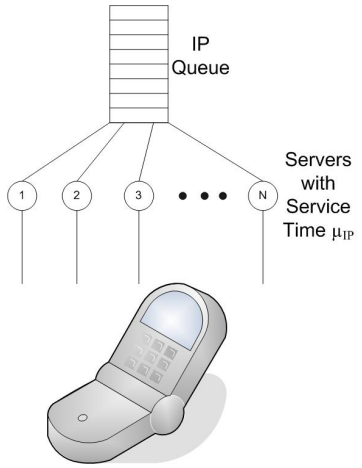

Figure 5: Queue with 1 to N servers with service time μ_{IP}

As seen in figure 5, a queuing system could have 1 to N servers. In 802.11e there would only be one server because there is only one common wireless channel to transmit on. In UMTS, there could either be one, if the traffic were transmitted on a Downlink Shared Channel (DSCH), or multiple servers if the traffic type we are considering is transmitted over multiple Dedicated Channels (DCHs).

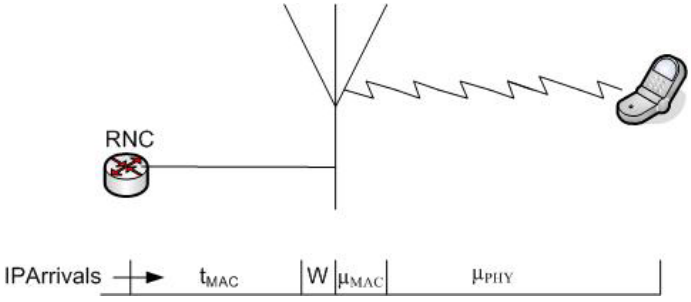

Figure 6: Components of the IP packet service time μ_{IP}

Above we can see an informal breakdown of the IP packet service time μ_{IP}. In 802.11e μ_{IP} is mainly determined by the backoff mechanism that determines the MAC service time, μ_{MAC}, and is dependent on the congestion state of the system. Retransmissions due to bit errors caused by fading can also cause the μ_{IP} to vary slightly.

In UMTS, the μ_{IP} is also affected by fading. However the main component that affects μ_{IP} is not congestion as in 802.11e. It is the interference caused by other flows transmitting simultaneously that results in bit errors and therefore retransmissions. This interference is due to the fact that there are many physical channels available to give almost one physical channel to each connection. However, if the DSCH is used, or we run out of channel codes for DCHs, the physical channel needs to be shared, and congestion comes into play.

VI. Conclusion

In this paper we introduce some common approaches to ensuring end-to-end QoS. The need for a RNC QoS mechanism in heterogeneous NGNs containing wireless links is identified, and a proposed framework is described. The framework motivates the development of cross-layer performance models of wireless access technologies. We introduced two common wireless access technologies, namely IEEE 802.11e and UMTS.

We note that in the development of an 802.11e cross-layer performance model, even though physical layer bit errors need to be accounted for, the most important layer is the MAC layer. The performance of 802.11e is congestion limited, and the backoff mechanism or scheduler is situated in the MAC layer.

In UMTS, congestion could come into play if the DSCH is being shared, or if there is a shortage of codes. However, generally the performance is limited at the physical layer by the amount of interference in the access network, cause by other users transmitting simultaneously on other channels.

So we see that although different technologies are performance limited at different layers, by developing cross-layered performance models up to the network layer, the translators proposed in our framework will be able to predict the performance implications of proposed actions.

Acknowledgment

I would like to thank Dr Hutchison and Ian Saunder for their inspiration and steadfast support to this work.

References

[1] "NGN 2004 Project Description", Version 3, 12 February 2004, www.itu.int/ITU-T/studygroups/com13/ngn2004/

[2] Agilent Technologies, "UMTS Network and Service Assurance", www.iec.org

[3] "Requirements on 3GPP system to Wireless Local Area Network (WLAN) interworking", TS 22.234, www.3gpp.org

[4] V.K. Varma, K.D. Wong, K Chua, F. Paint, "Integration of 3G Wireless and Wireless LANs", Guest Editorial, *IEEE Communications Magazine*, Nov 2003.

[5] U. Jain, Y. Yokoyama, A. Kumar, "Study of Factors Influencing QoS in Next Generation Networks", Ericsson Inc, http://whitepapers.zdnet.co.uk/0,39025945,60130431p-39001014q,00.htm

[6] J. Chen, V.C.M. Leung, "Improving End-to-End Quality of Services in 3G Wireless Networks by Wireless Early Regulation of Real-time Flows", in Proc. IEEE PIMRC'03, Beijing, China, September 2003 http://citeseer.ist.psu.edu/672558.html

[7] S. V. de Vasconcellos, J.F. de Rezende, "Using Differentiated Services in 3G Cellular Networks", http://citeseer.ist.psu.edu/554320.html

[8] T. Janevski, B. Spasenovski, "Admission Control for QoS Provisioning in Wireless IP Networks", in Proc European Wireless Conference 2002, http://docenti.ing.unipi.it/ew2002/proceedings/136.pdf

[9] T. Kawata, S. Shin, A.G. Forte, H. Schulzrinne, "Using Dynamic PCF to Improve the Capacity for VoIP Traffic in IEEE 802.11 Networks", in *IEEE WCNC*, March 2005, http://www1.cs.columbia.edu/~ss2020/papers/dpcf_wcnc.pdf

[10] IEEE 802.11e Draft Standard/D13.0, January 2005, http://shop.ieee.org

[11] "QoS Concept and Architecture", TS 23.107, http://www.3gpp.org/ftp/Specs/1999-10/for-itu/23107-300.pdf

[12] F. Borgonovo, A. Capone, M. Cesana, L. Fratta, "Performance evaluation of UMTS packet services", ST Journal of System Research, vol. 1, Issue 1, February 2004, http://www.st.com/stonline/press/magazine/stjournal/vol01/art04.htm

C Consani obtained her BSc and BSc(Hons) from the University of Cape Town, in 2002 and 2003 respectively. She is currently working towards her MSc with the DNA laboratory.