Cross-layer black-box approach to performance evaluation of next generation mobile networks

Y. Koucheryavy¹, D. Moltchanov¹, J. Harju¹, G. Giambene²

Abstract—Up to date the Internet and wireless communications have been evolved as separate technologies because of different types of traffic they were intended for. Nowadays, with commercial launch of third generation (3G) networks, the convergence of these technologies is getting clear. The adoption of the IP protocol for future mobile environment and extension of packet-based services to the air interface calls for advanced performance evaluation methods. The aim of this paper is to overview recent developments and outline directions of further work in performance evaluation of NG All-IP. We consider both current traffic and wireless channel modelling techniques and give reasons why these techniques are not applicable for future IP based mobile systems. Then, we state the need for new approaches to performance evaluation of NG All-IP networks that should be able to capture simultaneously three parts: mobility of users, traffic nature and unreliable transmission medium.

Keywords—NG All-IP, performance evaluation, cross-layer approach, traffic modelling, wireless channel modelling

I. INTRODUCTION

THE developing of mobile communication systems shifts approximately in ten years cycles. Nowadays, with commercial launch of third generation (3G) networks, we face the question of what a next generation (NG) systems should be. Rather than linking NG to something like 'yet more bandwidth', the trend is to see NG as a new paradigm for wireless networks, mobile and fixed, where network and service characteristics of reconfigurability, interworking and adaptable services are the order of the day.

NG systems should provide intelligent, adaptable and personalized services by encompassing all wireless systems from public to private, operator-driven to ad-hoc, broadband to personal area, 2G systems to 3G systems. They should be flexible enough to support all those services of today’s fixed networks without any limitations, and to provide them in an Always Best Connected (ABC) access strategy [1], so that the users are granted with the most suitable connection for a given environment, satisfying as much as possible the quality of service (QoS) requirements of the service and the user at a price which is right for both.

While NG networks are not well defined, there is a common agreement that these networks will rely on IP protocol as an end-to-end transport technology. The motivation is to introduce a unified service platform and transport facilities for future composite mobile Internet known as NG All-IP network.

In addition to broadband wireless access to the Internet, NG All-IP networks has to satisfy requirements of QoS-aware applications. Indeed, this is an inherent problem for many service types even in fixed networks. Wireless and mobility add their own QoS problems on top of this inherent IP flaw. Characteristics of wireless links, such high and correlated bit error rate and limited bandwidth of wireless channels have to be addressed before wireless Internet services will be commercially deployed. To facilitate development of these networks novel methods of teletraffic theory, optimization and design must be developed.

The intention of this paper is to clarify whether performance evaluation methods available to date can be successfully applied to forthcoming NG All-IP systems. Within such framework we consider both current traffic modelling and wireless channel modelling approaches available in literature. Indeed, survey of literature have shown that due to circuit-switched technology utilized by 2G systems, traffic models designed for conventional 2G mobile systems are primarily concerned with capturing session level parameters of traffic sources while network layer traffic demands were assumed to be constant during whole duration of a session [2], [3], [4], [5]. Therefore, those traffic models are not appropriate when mobile systems operate over IP protocol as an end-to-end transport technology.

Concerning wireless channel modelling techniques one might notice that most studies performed so far were devoted to performance evaluation of link layer protocols, while there were no studies targeted on IP layer performance evaluation. Therefore, those wireless channel models developed have to be extended to the network layer. Indeed, parameters of link layer’s error concealment protocols depend on both user’s traffic demands and wireless channel characteristics. Therefore, model of user’s traffic demands and wireless channel model should be merged together rather than considered independently.

The rest of the paper is organized as follows. In Section II we briefly consider an evolution of current mobile systems to NG All-IP networks. Then, in Section III we consider an end-to-end service architecture in NG All-IP networks. Performance modelling methods are outlined in Section IV. After that we propose a unified cross-layer approach for performance evaluation of these networks. Conclusions are drawn in the last section.

¹Authors are with the Network and Protocol group, Institute of Communication Engineering, Tampere University of Technology, Tampere, Finland, http://www.cs.tut.fi/lt/, e-mail: {y.k.moltchan,harju}@cs.tut.fi. ²Author is with Dipartimento di Ingegneria dell’Informazione, University of Siena, Siena, Italy, e-mail: giambene@unisi.it.
II. EVOLUTION TOWARDS NG ALL-IP NETWORKS

As mobile wireless communication systems evolve from supporting circuit-switched, voice-centric systems to future NG All-IP networks, the reliable and efficient support of heterogeneous multimedia services becomes significantly important [6]. In this scenario, system design becomes an important task, especially for what concerns the provision of effective solutions for the ubiquitous access to Internet and the QoS guarantee for different traffic classes. The system must be optimized to improve the performance, and therefore, QoS perceived by users and the efficient use of network resources, which is an important prerogative for making revenue for operators and for allowing the realization of services at affordable costs.

NG system should be based on new optimization paradigms in accordance with two different and complementary approaches, as outlined below [7].

A. Vertical approach

NG system optimization calls for vertical design of the air interface protocol stack. Such cross-layer approach requires interfaces between different platforms across the layers, which exchange control information beyond the standard ISO/OSI structure. Cross-layer interfaces can be within, between or beyond adjacent abstraction layers. Although interfaces between adjacent layers are in general preferable, there can be the need for efficient and direct interaction between non-adjacent layers; in general, a layer should be aware of the other layers of the protocol stack. In fact, OSI layer 3 (e.g., IP) and above often need direct interfaces to OSI layer 2, e.g., for handover support. Another example concerns transmission parameters (e.g., transmission mode, channel coding and link layer retransmissions) that must be related to application characteristics (e.g., type of information, source coding, etc.), network characteristics, user preferences and context of use. Moreover, layers 2 should be aware of higher layer (i.e., 3 and 4) behaviors in order to take decisions on traffic management. In general, we can refer to the air interface protocol architecture shown in Fig. 1, with the proposed interactions among the different layers and with a middleware.

B. Horizontal approach

In the NG scenario, different wireless technologies will operate to allow the best access to the users, depending on their locations, mobility characteristics, applications, user profile, etc. This is in accordance with the ABC paradigm, and therefore it is necessary that the use of the resources in the different radio access networks (RAN) be globally coordinated by means of a resource brokerage function. Such intelligence is centralized and allocates sessions to RANs or switches them from one to another, when some conditions are met.

III. END-TO-END VIEW OF NG ALL-IP NETWORKS

A. End-to-end service architecture in All-IP network

Initially, a major weakness of IP-based networks is that, they do not provide QoS guarantees to their applications. Indeed, only the best effort service was supported in public domains, so that inadequate QoS levels were obtained by emerged in the middle of 90th real-time multimedia services. However, currently, these networks are developing towards QoS-aware networks.

To maintain QoS in fixed IP networks, IETF has proposed two QoS frameworks. These are Integrated Services (IntServ, [8]) based on connection-oriented resource reservation principle, and Differentiated Services (DiffServ, [9]) based on service differentiation approach. IntServ, based on connection admission control (CAC) procedures, can potentially provide deterministic QoS guarantees and requires a signaling protocol in order to inform network elements about necessary resource reservation. On the other hand, DiffServ employs a different approach. In order to distinguish between flows with different QoS requirements DiffServ specification defines packets marking procedures. It provides probabilistic QoS guarantees to aggregated traffic flows and uses CAC algorithm, which is based on Service Level Agreements (SLAs) between subscribers and service providers or between two service providers.

Both IP QoS mechanisms have been specified having a wired networks in mind. Whereas these mechanisms had already received a lot of attention from researchers, only few studies on supporting IntServ or DiffServ at the wireless interface have been published. One may note that QoS problem in IP networks will get even worse when multimedia services will become more popular and have to be extended to the air interface.

Let us consider an end-to-end client-server service configuration in NG All-IP networks as shown in Fig. 2. One may note that in principle, there can be a number of different access networks (AN) in NG All-IP network, however, only two ANs are shown in the figure. These are fixed AN and RAN. In accordance with All-IP architecture, RAN should take on certain configuration and consists of a number of IP-capable routers, part of them are actually combined with radio access equipment. These nodes are first routers in uplink direction when mobile terminals access the network. In what follows, they are called first fixed access points (FFAP). According to client-service ar-
chitecture under consideration, the service, a user wish to access, is located in the fixed part of IP based network. In this case, packets generated by client are routed via air interface to FFAP, then through gateway node between RAN and fixed IP network, and via a number of routers located in fixed part of the IP network.

To evaluate QoS expectations that a user application may experience in considered client-server service configuration, we have to take into account all those packet losses and delays that may occur at each point up to destination and in backward direction.

It is well-known, that packet losses may occur due to both bit errors in transmission medium and network nodes’ congestion situations. Given the current evolution of fixed transmission medium, it is well understood that packet losses caused by bit errors at physical layer can be neglected. Due to these reasons, it is always claimed that packet losses in fixed networks occur due to buffer overflows caused by congestion of network nodes. Therefore, to evaluate performance of user applications in fixed networks it is sufficient to evaluate those packet losses caused by buffer overflows.

One the contrary, dealing with wireless networks we cannot neglect those packet losses caused by bit errors in wireless transmission medium. Moreover, it is well-known that these errors may contribute a lot in end-to-end performance degradation of packetized user application. Indeed, they have completely different nature compared to what we dealt with in fixed networks so far. Due to these reasons, the wireless interface between mobile terminal and FFAP becomes an unexplored ‘weak point’ in future end-to-end IP QoS architecture.

B. End-to-end protocol model

To determine what kind of performance degradation at the wireless interface we should expect, let us consider an end-to-end protocol model of service architecture considered in previous subsection, as shown in Fig. 3.

One might note that since the error rate of fixed transmission medium is very low, we can assume that delay caused by physical and link layer is constant and equal to the sum of propagation time over the fixed transmission medium and time to decode and packetize these bits at the physical and link layer before sending them to IP layer.

Taking look at the wireless interface we have to note that abovementioned constant delay assumption, which are common for fixed networks, no longer holds. Indeed, error concealment procedures may significantly increase the delay of transmission applying various error correction techniques like Forward Error Correction (FEC), Automatic Repeat Request (ARQ) or combinations of them. Additionally, these link layer protocols may allow some errors propagate to IP layer. These errors may also cause retransmissions through underlying layers. During long periods of worse quality of wireless channels both errors and delays can be very high, and therefore, may significantly affect the performance of packetized user applications. Therefore, dealing with end-to-end performance evaluation of NG All-IP networks, in addition to IP level peculiarities at the wireless interface we have to take into account behavior of underlying layers. Therefore, an extension of QoS guarantees to the wireless interface introduces problems that should be solved before required QoS of user services will be achieved. These new challenges require development of new cross-layer integrated methods of performance evaluation at the wireless interface.

There are a lot of publications related to performance evaluation of fixed networks (see [10], [11], [12], [13], [14], [15], [16] among others). On the contrary, there are only a little work devoted to packet level performance evaluation of packetized user applications running over the wireless channels. To date the only throughout study presented in [17], where authors consider the convergence of wireless channel modelling and queuing analysis.

IV. PERFORMANCE MODELLING OF AIR INTERFACE

Performance evaluation of QoS experienced by applications at the air interface is a very challenging and complicated task due to a number of factors. Firstly, wireless channels have highly time-varying behavior due to shadowing, multipath fading, atmospheric conditions and other effects. Secondly, the nomadic behavior of users introduces another uncertainty. It may cause a handover procedure between different cover-

Fig. 3. End-to-end protocol model of NG All-IP network.
A. Models of user’s traffic demands

Due to circuit-switched technology utilized by 2G systems, a user of a mobile system is granted a CBR channel during the cell dwell time (time for which a particular user stays in a certain cell). When a free channel is not available, a user is forced to leave the system. Handover calls between cells are served almost similarly: the handover call to a new cell is assigned one of the free channels during the cell dwell time. If all channels in the new cell are busy, the handover call is terminated. Therefore, the QoS provided to users by 2G mobile systems is typically limited to session level parameters like new call blocking, handover call blocking etc. Correspondingly, traffic models designed for 2G systems are primarily concerned with capturing session level parameters of traffic sources while network layer traffic demands were assumed to be constant during whole duration of a session.

In NG All-IP networks all calls including voice ones will be IP based. Hence, dealing with NG All-IP networks in addition to session level QoS parameters we have to consider IP level QoS characteristics like packet loss, packet delays, delay jitter etc. Thus, novel traffic models taking into account both session and packet level traffic demands are needed.

Due to the fact that users of NG All-IP mobile networks may not occupy constant bandwidth during the whole duration of a session, these systems may potentially benefit from statistical multiplexing at the network layer. In this case, a small fraction of users that employ bandwidth-greedy applications may produce a bottleneck at the wireless link. Therefore, the focus of user teletraffic modelling should be extended to include single traffic models of various applications at both session and packet levels.

A survey of literature (see [2], [3], [4], [5] among others) has shown that traffic models designed for conventional 2G systems do not take into account mobility behavior of single users. Most of them assume a stationary behavior of large populations of users and try to capture session level parameters by distributions like exponential one or its mixtures. Those models, while were shown to be relatively fair predictors of stationary characteristics, are not concerned with per-user granularity, and therefore, are not appropriate for traffic modelling purposes in NG All-IP networks, where the mobility behavior can significantly affect session level parameters, and as a result, packet level characteristics.

To illustrate the influence of user mobility behavior on its traffic demands let us define two classes of user mobility behavior with following two states:

- Fixed;
- Nomadic.

Assume that when the mobile user is in the fixed state it uses network services with probability $p^F$, and with probability $(1 - p^F)$ stays idle. Similarly, when the mobile user is on move it accesses service with probability $p^N$, and does not use the network with complementary probability $(1 - p^N)$. Indices $F$ and $N$ stand for fixed and nomadic states, respectively.

In what follows we explain why it is essential to assume that call level traffic demands depend on the mobility of the user, which leads to the case $p^F \neq p^N$.

To date voice service has been a dominating application in conventional 2G systems. Due to physiological tests the aural impression of information does not require too much attention (up to 20%), and therefore, a moving user is aware of the situation around even when moving (driving, walking, etc.). Therefore, to date one could fairly assume $p^F = p^N$.

The situation changes with the introduction of multimedia services. Indeed, the visual information gives us more than 80% percent of overall information. Moreover, the combination of aural and visual information dominates with overall percentage up to 97%, and therefore, significantly affects our attention. When the aural information is presented in conjunction with other media it is fair to assume that moving users will not use such services frequently, since they require too much attention to concentrate on perceived information. It leads to the case $p^F \neq p^N$ and dependence between mobility and traffic demands is evident. Since the calling activity of the user influences its traffic demands, such dependence may affect the QoS provided to the user.

Additionally, in some countries there are restrictions imposed on the usage of mobile phones when driving. It also leads to the case $p^F \neq p^N$ even for conventional monomedia applications.

Due to abovementioned reasons dealing with traffic modelling in NG All-IP networks we have to consider three dependent components: mobility behavior of the user, session level teletraffic demands and packet level teletraffic demands. Mobility of the user implicitly define user traffic demands at the session level, and therefore, traffic demands at the packet level.

**Mobility models.** There are a number of mobility models available in literature. The main purpose of those models is to predict the movement of users
within a certain area. Most of those models are related to and came from solutions of paging problems in 2G mobile systems. Due to their nature, such models have a very simplified representation of teletraffic part while the mobility part receives main attention and often modelled by complex stochastic process.

**Teletraffic models.** As far as there are a number of applications in current IP networks, all of them should be supported by NG All-IP networks. It is known that each application is characterized by its own traffic characteristics. These characteristics should be measured from real traffic traces and then used to parameterize a model.

It is recognized from traffic modelling in fixed network that most applications can be represented using different traffic models. However, in those cases when general traffic model is used, we have to parameterize it differently depending on the traffic characteristics of certain application. The future trend must be to use a general models like Markovian Arrival Process (MAP) or Batch MAP (BMAP) or their discrete-time counterparts in order to represent the teletraffic part in NG All-IP networks. In addition to unification purposes, such method should potentially provide more capabilities to represent teletraffic part of user's traffic model in NG All-IP networks [18], [19].

**Cross-layer integrated models.** Teletraffic models, developed for wired networks cannot be effectively used in wireless environment since mobility of the user is not taken into account. Those models, developed for 2G networks have an oversimplified structure of teletraffic part. Therefore, an adequate and accurate model for the NG All-IP environment should explicitly take into account both mobility of users and teletraffic characteristics of various applications at per-user granularity. These models have to be further applied to investigate the QoS level provided for each application using analytic or simulation techniques.

To date only few studies that simultaneously consider user's mobility behavior and its traffic demands are available. Particularly, we can refer to the study presented by Antunes et al. [20]; on the basis of general assumptions regarding teletraffic and mobility of the user, they propose to use a combination of two MAPs one of which describes user's mobility while the other specifies the teletraffic characteristics of applications. Despite obvious advantages, their model is quite complex, and therefore, the application area is limited.

In [21], authors developed an integrated cross-layer traffic model for NG All-IP networks. Basically, it consists of two different parts: mobility model and teletraffic model. The mobility behavior of the user is captured by a Markov chain with finite state space, while the teletraffic characteristics at the call level are represented by a Markov chain with two states. Both models are integrated, so that the whole model can be seen as a doubly stochastic process while does not lead outside one-dimensional Markov chain framework. This is due to the special integration of both parts, which makes a teletraffic part a function of the mobility one of the composite model. They also show how to extend the proposed model to capture packet level characteristics of CBR and VBR traffic types. While analytical tractability is retained, the main application area is simulation studies.

**B. Models of the wireless channels**

An important property of wireless channels is that their performance is affected by a number of factors. It is always claimed that most important are:

- Multipath fading;
- Shadowing.

Multipath fading is caused by multiple path propagation of the wave between transmitter and receiver. Due to unknown number of possible paths between transmitter and receiver multipath fading is complex stochastic process that depend on the mobility of the user.

Shadowing is caused by the loss of the line on sight between transmitter and receiver due to shadowing of the propagating wave by large obstacles. While moving between shadowers, the received signal power varies in accordance with alternating interruptions and release of the line of sight between the receiver and the transmitter. Size and location of obstacles are random, and therefore, the occurrence and duration of interruption are also random. One may note that the movement of the user between shadowers is mobility-dependent process.

The mobility behavior of the user plays an important role in both abovementioned factors. It is well-known that mobility of the user affects both multipath fading and shadowing. However, in addition, movement of the user introduces another fading factor which is proportional to the distance between the transmitter and receiver. Indeed, during the duration of the session the user may change its location many times, and therefore, power of received signal varies in accordance with distance between transmitter and receiver. In what follows we call it *constant fading.*

One might not that the signal-to-noise ratio (SNR) experienced by mobile terminal is complex mobility-dependent stochastic process resulting in three fading components each of which significantly influence performance of the wireless channel.

There are a number of wireless channel models proposed in literature. State-of-the-art models capturing fast fading characteristics of wireless channels are based on Markov chains where the state space of the models is designed to capture histogram of relative frequencies of received SNR level. These models are also able to capture autocorrelation properties found in empirical SNR traces. To date only few studies dealing with capturing slow fading phenomenon of wireless channels are available. One could refer to the study performed by Huo [22] who employs a one dimensional stochastic process to represent a movement of the user between shadowers. To date there were no studies aimed to capturing constant fading character-
istics of wireless channels caused by change of distance between transmitter and receiver.

Basically, all abovementioned models primarily capture only one fading component which is dominant for a given type of wireless channel. However, it rarely happens that the performance of wireless channel is affected by only one fading component. Given a movement of a user, one should expect an existence of all fading components when the influence of each component cannot be neglected.

We have to note that all abovementioned models do not concern with IP layer at which user’s traffic demands are usually represented. Indeed, those models give only basic ideas how SNR changes over time depending on fading characteristics of wireless channels, while they do not clarify how many IP packets would be lost or retransmitted.

One should note that often the bit error rate (BER) of the wireless channels is the complex but deterministic function of received SNR [17]. Therefore, provisioning of IP layer QoS for wireless channels is therefore difficult due to high and correlated BER caused by changes in received SNR. To overcome these problems, a radio link protocol (RLP) with a suitable error control mechanism has to be used at the link layer. The choice of error concealment strategy and its parameters depends on both wireless channel characteristics and the nature of the traffic source. Therefore, wireless channel model at IP layer can be represented as a complex function of the following components: user’s mobility behavior, SNR, user’s application, coding scheme used and the physical layer and error correction techniques at the link layer.

C. Cross-layer black-box approach

Considering components that influence performance of packetized applications running over the wireless channels one may note that mobility of the user affects both teletraffic characteristics and quality of the wireless channel. Additionally, wireless channel model and user’s traffic model are complex functions of a number of environmental characteristics like user preferences, link layer error concealment procedure, applications, landscape etc. In this case, the analysis of QoS expectations that a user application may experience running over the wireless channel is sophisticated task which involves a number of independent stochastic factors. Therefore, both quality of wireless channel and user’s traffic demands should not be considered independently of each other.

One of the possible ways to evaluate QoS expectations is to integrate wireless channel model to teletraffic model. This integration have to be performed at IP layer. However, it leads to high complexity of resulting model. Indeed, the resulting model in addition to cross-layer structure of each counterpart should reflect all interdependence (correlation) properties between different parts of two models.

Another approach is to take a so-called ‘black-box’ approach. In accordance with this approach we assume that the we are given a set possible parameters at the input of the system, i.e. set of possible applications with corresponding session level and packet level characteristics, set of various mobility behaviors of a user, set of physical layer’s coding schemes, set of link layer’s error correction methods with appropriate sets of their parameters. Using these parameters we have to implement a testbed and collect an appropriate performance measures at the output of the system. These performance measures include IP level performance parameters as well as other higher level performance measures like perceived quality obtained by the user.

Fig. 4. Cross-layer black-box approach.

In this case the testbed is seen as a complex translator parameters at the input of the system to corresponding output characteristics. Therefore, one can tune certain input parameters like physical layer’s coding schemes or link layer’s error correction methods to optimize them to match the desired output.

V. Conclusion

In this paper we considered conventional traffic modelling and wireless channel techniques and gave reasons why they cannot be applied to forthcoming NG All-IP wireless networks. We performed analysis of interactions between mobility behavior of the user, its traffic demands and the quality of the wireless channel. Based in this we outlined cross-layer structure of novel traffic models and wireless channel models which should be developed to advance current mobile systems to truly end-to-end IP based wireless networks.

We also proposed a cross-layer black-box approach for qualitative and quantitative evaluation of QoS expectations of user’s applications running over the wireless interface. The major advantage of proposed approach is that it allows to avoid complexity of cross-layer traffic and wireless channel modelling methods at IP layer.

REFERENCES

mance and channel holding time characterization in wire-


mance and channel holding time characterization in wire-


