

Notes on Quality of Service and Performance Evaluation in 4G All-IP Networks

Y. Koucheryavy, D. Moltchanov
Institute of Communication Engineering
Tampere University of Technology
Tampere, Finland
{yk, moltchan, harju}@cs.tut.fi
<http://www.cs.tut.fi/tlt>

Abstract—Up to date the Internet and wireless communications have been considered as separate technologies because of different types of traffic they were intended for. Currently, with commercial launch of third generation (3G) networks, the convergence of these technologies is becoming reality. The purpose of our paper is to overview recent developments and show directions of further work in performance evaluation of fourth generation IP-based networks (4G All-IP). This will involve defining appropriate QoS models, as well as addressing traffic modeling and engineering problems, and internetworking issues among mobile wireless networks and between these and the current Internet. We consider both current traffic modeling and performance evaluation techniques and give reasons why these techniques cannot be applied to forthcoming 4G All-IP networks. Based on these we show the need for new types of traffic models for 4G All-IP networks, which should be able to capture three parts as solid one: mobility of users, teletraffic nature of applications and unreliable medium properties. We also give a brief review of performance evaluation of future 4G All-IP networks.

4G All-IP networks, performance evaluation, wireless networks, mobile networks, radio access networks

I. INTRODUCTION

The developing of mobile communication systems shifts approximately in ten years cycles. Nowadays, with the commercial launch of third generation (3G) networks upon us, we have to face the question of what a fourth generation (4G) scenario might or should be. Rather than linking (but not excluding) 4G to something like “yet more bandwidth” the trend is to see 4G as a new paradigm for wireless networks, mobile and fixed, where network and service characteristics of reconfigurability and interworking networks, adaptable services, interoperability so on are the order of the day.

For instance such systems must be flexible enough to provide all those services which are provided by wired networks today without any limitations, and provide them in an always best connected (ABC) access strategy, e.g. providing the users with the most suitable connection for the given environment, satisfying as best as possible the QoS requirements of the service and the user at a price which is

right for both. Such features of future networks can be based on the multi-access networks support and should be transparent to the user. One of the possible and promising solutions for 4G systems is to use reconfigurable mobile networks (including ad-hoc networks). Such networks are envisioned as being highly dynamic self-organizing and self-configuring networks, which will not necessarily require fixed infrastructure (as in the Ad Hoc networking case) and may take on the characteristics of an autonomous system in the Internet. In Ad Hoc networks all nodes are capable of movement and theoretically may be interconnected arbitrary. These features make Ad-Hoc networking an attractive addition for the purposes of mobile communication, improving its pervasiveness, potential QoS benefits including adding more options and possibilities for an ABC access strategy implementation, while raising new challenges in many R&D areas of mobile networking (traffic & network engineering, security, business models etc.).

However, one of the first major challenges is to address the QoS question. In “All-IP networking” 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.

While 4G networks are not as well defined as 3G, it is almost clear that these networks will use IP protocol as end-to-end transport technology. The major motivation is to have a common service platform and common transport technology for future composite “Internet mobile” network. 4G All-IP based mobile network will be able to provide enormous benefits for users, operators, service and content providers like interactive multimedia services, global mobility, service portability, interoperability, support of different terminals and equal service quality as in current Internet.

Nowadays, researchers are stressing to combine new access technologies like wireless LANs (WLAN), with 3G cellular systems. We are also expecting that this combination would provide plenty of new services offered to nomadic users. Therefore, 4G All-IP networks will likely to combine broadband radio access technologies like HIPERLAN/2 and wireless LAN (802.11b) with 3G wireless access technologies like universal mobile telecommunication system (UMTS) terrestrial radio access network (UTRAN) and GSM evolution

RAN (GERAN). Such combination will enable seamless IP-based services for users in hot-spot areas.

In addition to enabling broadband wireless access to the Internet, 4G All-IP architecture has to satisfy the requirements of QoS-aware applications. However, those characteristics of wireless links, such high and correlated error rate and low bandwidth of typical wireless channels, need to be addressed before a wireless Internet service would be deployed.

Current multimedia applications like audio, video and data are characterized by much higher bit rates compare to those produced by compressed speech information or even by high-quality pulse code modulation (PCM) codecs, which have been used for a long period of time in both fixed and mobile networks. These new services require development of new methods of traffic theory, optimization and design. For example, it is necessary to develop a theory of multi-dimensional distributed queues with priority service for integrated traffic in multi-point radio channels with multiple accesses. Among others, the special attention should be paid to the traffic modeling - it provides the starting point in theoretical analyses of QoS experienced by application.

Our paper is organized as follows. Section 2 deals with 4G All-IP networks peculiarities. Current 3G networks architecture and development of 4G All-IP networks are considered there. In Section 3 QoS issues of both 3G and 4G All-IP networks are considered. In Section 4 the attention is paid traffic modeling in current and future networks. We overview there both mobility and traffic nature issues as well as define an integrated traffic models. Performance evaluation issues are considered in Section 5. The conclusions are given in the last section.

II. FUTURE 4G ALL-IP NETWORKS

A. 3G Network Architecture

The 3G network architecture known in Europe as Universal Mobile Telecommunication System (UMTS) is an evolution of GSM network. The network is mainly consists of two parts (subnetworks): a UMTS terrestrial radio access network (UTRAN) and core network (CN). Further the CN is logically and physically divided into circuit switched CN (CSCN) and packet switched CN (PSCN). Simplified network architecture is shown in Fig. 1. There are a number of UMTS releases of standards each of which covers different RAN and CN transmission technologies. For example, release R99 of the UMTS system supports WCDMA access technology at the air interface and ATM transport technology whereas release R00 supports two RAN technologies.

Since the main objective of R00 is to use the same CN for the two RAN both available types of RAN GERAN (GSM-EDGE RAN) and UTRAN have to be connected to the same CN.

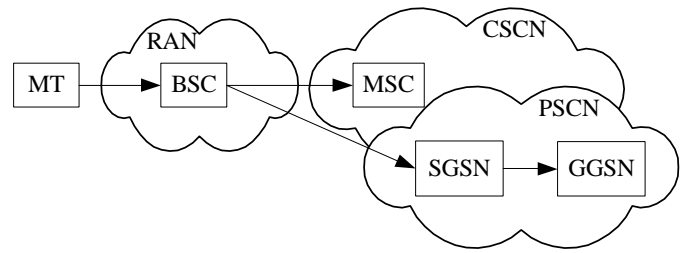


Fig. 1. Simplified 3G network architecture.

B. 4G All-IP Network Architecture

As far as 4G All-IP networks are not as well defined as 3G we outline here basic features which are to be implemented in these systems.

The scope of an All-IP mobile network focuses on both core networks and access networks. In 4G All-IP networks packet switched (PS) CN and circuit switched (CS) CN are supposed to be replaced by single IP based core network (PSCN). Therefore, the whole end-to-end path between mobile terminal and service access point is assumed to be IP based, all mobile terminals are to be fully IP capable and all services are IP based.

A clear separation between radio access network (RAN) and CN in 3G networks has already led the network to multi-access environment (Fig. 2). The main objectives of the RAN are to provide an access technology at the air interface and to hide all access specific features from the CN. Therefore, RAN comprises all functions that enable a user to access services. Because of that, the CN has a little impact on introduction of new RAN, and can evolve independently of it. In order to enable seamless IP-based services for users in hot spot areas and on the move, at this stage it is needed to study a system architecture that combines broadband wireless access technologies (HIPERLAN/2, IEEE 802.11) with RANs (UTRAN, GERAN), since WLANs were primarily developed to match requirements of non real-time services like file transfers, remote access to LAN and other while UTRAN and GERAN are mainly targeted on high quality voice communications. Note that the main reason to consider such heterogeneous environment is that it is almost impossible to define a RAN that combines all advantages of different RANs.

In addition to multi-access environment, it is proposed that 4G All-IP networks will have a layered network infrastructure with at least two hierarchical levels. It is proved that such architecture provides a high flexibility for current monoservice cellular network and expected that it will be able to perform well in a multiservice All-IP environment.

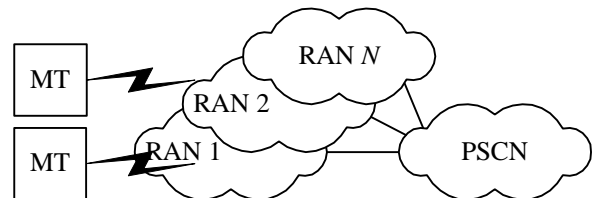


Fig. 2. Multiaccess 4G network architecture.

In accordance with layered network infrastructure, there will be cells of different size (picocells, microcells, macrocells) each of which servicing users with different mobility patterns or users producing an overflow traffic. Layers with picocells or microcells are able to provide a high capacity with high bandwidth in hot-spot areas. They can serve slow mobility users with relatively high traffic demands. Macrocells will serve users with high mobility patterns (highway users). Different ANs can be considered as different levels network hierarchy.

C. Reconfigurable Access Networks

Reconfigurable access networks also known as ad-hoc networks are valuable topic of ongoing research. These networks do not have a fixed configuration and all stations are assumed to be mobile. Wireless links between them are assumed to be configured on demand. Within the 4G All-IP framework ad-hoc networks can be considered as a special structure of RAN.

It is supposed that ad-hoc configuration of RAN is able to provide plenty of benefits to service providers especially in areas of low population. However, both concept of ad-hoc networking and basic protocols for these networks are still under heavy development and, therefore, we can expect that at least the first evolution of 4G-All-IP network will not use such types of access networks.

D. Always Best Connected

Always best connected (ABC) is a concept that allow users of 4G All-IP networks to choose the most suitable RAN which is the best fit for applications from QoS point of view. There is a common agreement that ABC should be implemented as an inherent feature of future 4G All-IP networks. ABC is straightly related to vertical (intersystem) handover, which is one of the fundamental issue of 4G All-IP networks. Indeed, by proper implementation of vertical handover it is possible to achieve seamless services performance over multi-access AN in 4G All-IP networks.

III. QoS IN 4G ALL-IP NETWORKS

A major weakness of the current IP networks is that they do not provide a capability to support a wide range of QoS guarantees. This problem will even worsen as wireless services become more popular and have to be supported by the Internet. It provides best-effort service only which, in turn, does not satisfy real-time services nature.

In the past IETF has proposed two end-to-end QoS frameworks: Integrated Services (IntServ, [1]) and Differentiated Services (DiffServ [2]). IntServ approach is based on connection admission control (CAC) procedures and uses explicit resource reservation technique. IntServ can provide deterministic QoS guarantees and requires signaling protocol to inform network elements about the necessary resource reservation. On the other hand, DiffServ employs quite different approach. DiffServ defines packets marking procedures to distinguish between packets with different QoS requirements. It provides probabilistic guarantees to aggregated traffic flows inside the network and uses a sort of CAC

algorithms, which is based on service level agreements (SLAs) between subscribers and service providers or between two service providers. Whereas, IntServ and DiffServ have received a considerable attention from researchers, few studies on supporting a wireless IntServ and DiffServ have been published.

The DiffServ framework is aimed on traffic aggregates [3,4] and no per-flow state needs to be maintained in the core routers while IntServ, when used with RSVP can provide a deterministic QoS guarantees to applications. Based on the above trends, it can be identified that the integration of wireless systems and IntServ or DiffServ can be promising research direction.

Packet data services will play a major role in 4G All-IP wireless multimedia services. A key component of packetized services is to ensure end-to-end QoS requirements. To achieve that four traffic classes has been defined by 3GGP. They are conversational, streaming, interactive and background classes. The conversational and streaming traffic classes should be utilized to carry real-time traffic flows (VoIP, VoD). The interactive and background classes are targeted to non real-time traditional Internet applications (web browsing, telnet, e-mail, FTP).

However, a lot of examples can be readily provided to claim against such rough differentiation. Those applications, which fall into one traffic class in accordance with 3G classification, can actually have significantly different QoS requirement. For example, consider a streaming class, which cover a broad range of video transmission services. If we are dealing with video-on-demand service we are not forced to guarantee low delay and low loss service since those packets, which were lost or delayed by network, can be retransmitted by the sender upon request. However, concerning the videotelephony we should provide both strict delay and strict loss since we are dealing with more conversational service instead of streaming one.

A. QoS Frameworks for 4G All-IP networks

Since the straight connection between 4G All-IP network and public Internet is expected, to implement QoS framework firstly we have to consider those QoS frameworks, which are already available for IP-based fixed networks. Here we consider two most obvious QoS approaches in 4G All-IP network. These are:

- IntServ-DiffServ approach [1];
- DiffServ approach [2].

Let us consider the first one. Since the number of users within the well-provisioned cell coverage area is statistically limited, we can expect that the IntServ can be successfully implemented on a RAN scale. All network elements in this architecture have to be RSVP-aware and should maintain per-source states. Note that as far we are concerned to RAN, which is the special case of arbitrary access network, the amount of RSVP soft states will not be too large.

Along with IntServ implementation within the RAN we can also expect that some sort DiffServ differentiation can be

implemented within the backbone Internet routers serving enormous number of connections. In this case certain border routers have to be located between IntServ and DiffServ parts 4G All-IP networks. Therefore, an adequate mapping between 3G-like service classes IntServ TSpecs and DiffServ codepoints should be provided within the appropriate network entities.

IETF's RFC 2998 covers an implementation of end-to-end QoS within the heterogeneous IntServ-DiffServ environment where within access network and in backbone one IntServ and DiffServ frameworks and implemented correspondently.

In accordance with second approach all routers along the whole end-to-end path are assumed to be DiffServ capable. Therefore, there can be more than one domain and proper service level agreements (SLA) should be established between them. In this case the mobile terminal will be able to use a host-marking procedure in according to DiffServ specifications and therefore can differentiate applications properly. Moreover, when a desired QoS level is not possible to keep unchanged, a renegotiation process should be implemented.

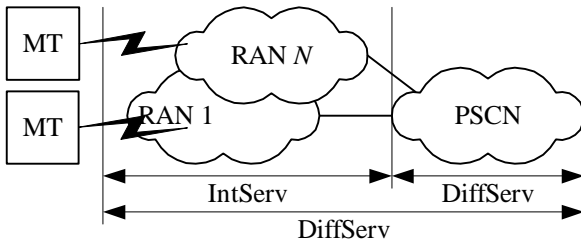


Fig. 3. Two QoS scenarios in 4G All-IP network.

However, we have to note that IP QoS mechanisms, such as DiffServ and IntServ have been specified having wired networks in mind. Currently, the challenge in the QoS area is to add IP QoS in the reservation mechanisms of radio network technologies and to combine the two approaches as one consistent end-to-end resource reservation model.

B. Applications specifics

The multimedia services are growing in popularity. It is primarily caused by several reasons: advances in compression technology, high-bandwidth storage devices and high-speed access networks. It would be beneficial to support such applications in 4G All-IP networks. Moreover, current applications, like multimedia ones, may have a lot of traffic flows involved in communication. These flows may sometimes have different traffic characteristics and require different QoS guarantees.

Applications emerging from current Internet are becoming capable of defining the required QoS level. At the same time, the end-users become more demanding in terms of service level obtained from the service provider. Traffic generated by current and future applications will require strict and different QoS guarantees in terms of bandwidth, losses, delay and jitter.

It is well known that voice traffic can be characterized by strict time constraints, constant or probably varying low bit rates and additionally can tolerate some losses. Videotraffic can be considered as highly variable bit rate application with strict

delay and loss requirements. Plain data traffic is loss sensitive and needs no delay guarantees. Different QoS guarantees for each of these classes should be provided.

IV. LOAD BEHAVIOR MODELS

The mechanisms responsible for QoS provision should predict the future state of the 4G All-IP networks based on both user mobility and user traffic parameters. Such mechanisms have to incorporate a so-called connection admission control (CAC) algorithm, which denies and allows the admission of new calls based on the network states.

Monitoring of real network can be done using specific software. However, in those cases when real implementation of considered network is not exist, an adequate traffic models that emulate the behavior of users should be used instead. From this point of view it is crucial to develop packet level traffic model of various types of traffic, which are assumed to be carried by 4G networks. It can serve as basis of efficient QoS management and network control.

Note that both the addition of new user services and specific user behavior will impose a lot of different and often hardly to satisfy requirements on 4G All-IP network management. Current modeling techniques, developed for fixed networks, do not take into account these new challenges and, therefore, we have to tackle with different modeling approaches.

A. Mobility Models

There are plenty of mobility models presented in literature. The main purpose of these models is to predict user movement within the certain area. Most of available models are connected with and come from the successful solutions of paging problems in 2G networks. Therefore, such models have a restricted representation of teletraffic part while the mobility part receives a lot of attention and often modeled as quite complex stochastic process.

B. Teletraffic Models

As far as there are a number of applications in current IP networks all of them should be supported in 4G All-IP networks.

Each particular application can be described by its traffic characteristics. These characteristics should be measured from real traffic traces and then used to parameterize a mathematical model. Note a fact well known from fixed network traffic modeling that most applications need to be characterized mathematically using different traffic models. In the case of general traffic models like Markovian arrival process (MAP) or batch MAP (BMAP) which are able to model a wide variety of real traffic sources we have to parameterize them differently depending on traffic characteristics generated by certain application. Therefore, modeling of teletraffic is quite complex even if we do not consider other specific features of nomadic users.

C. Integrated Traffic Models

Traffic models, which were developed for wired networks cannot be adequately used in wireless networks since both mobility of users and unreliability of transmission medium were not considered there. Those ones, in which the mobility is considered do not incorporate an adequate teletraffic model of the traffic source or use a quite simplified approach concerning these features of traffic sources. Therefore, an adequate model in wireless environment should take into account both mobility of users and teletraffic characteristics of various applications.

The only example of such traffic model, which can be found in literature, is a sequence of works presented by Pacheco *et.al.* [5]. Based on some general assumption they considered a combination of mobility and teletraffic models. Their model is based on combination of two MAP processes one of which describes the mobility of user while another specify the teletraffic features of applications. Despite the obvious benefits, their model is quite complex and, therefore, restricts its analytical use in performance evaluation of 4G All-IP networks.

Based on abovementioned considerations, there is an increasing need to construct new models of user traffic for 4G All-IP networks. These traffic models can further be applied to investigate a QoS provided for each application using a certain analytic or simulation techniques.

V. PERFORMANCE EVALUATION

Analysis of QoS degradation in 4G All-IP environments is a very challenging task. Firstly, the channel characteristics of a wireless link have unpredictable time-varying behavior due to shadowing, multipath fading, etc. Moreover, the movement of nomadic user from cell to cell introduces another uncertainty. The movement causes a handover mechanism, and therefore, the access point of the mobile host to the fixed networks changes. This results in the establishment of the new path in the fixed part of the RAN and sometimes in CN. Note that this new path can, of course, have quite different delay, loss and bandwidth characteristics compared to old one. In addition, the ABC concept allows the users to change RAN dynamically and, therefore, introduces an additional uncertainty.

A. Packet Level QoS

In most cases in cellular networks QoS has been traditionally characterized through a set of call level QoS parameters like call blocking, handover call blocking etc. Obviously, the QoS provisioning in 4G All-IP networks will not be restricted to circuit switched voice calls or constant bit rate (CBR) connections. Moreover, all calls including voice ones will be IP-based and will require quite different and often variable bit rates. Therefore, instead of call level QoS parameters we have to consider IP level QoS parameters like packet loss, packet delays, delay jitter etc.

Analysis of literature has shown that as opposed to call level performance analysis of 2G and 3G systems up to date there were no research activities in the area of packet level performance analysis of 4G All-IP networks. However, it is

supposed that this topic will be given a considerable attention in next decade.

It is well known that in wired networks packet losses are primarily due to buffers overflow within the routers since the error rate of transmission medium is very small (less than $10E-6$). Therefore, in case of wired network analysis packets losses caused by error at the physical layer can be neglected. At the contrary, analyzing packet level performance of 4G All-IP networks we have to primarily take into account those packet losses which are caused by error rate of wireless transmission medium.

B. Wireless Part of the RAN

We can divide an arbitrary RAN in two separate parts: wireless part and fixed one. They are different and need different means of analysis.

An important property of the wireless links is that its performance is often affected by atmospheric conditions. Provisioning of QoS is therefore difficult because of relatively low bandwidth and very high and correlated bit error rate (BER) of the wireless channel. To overcome these problems, a radio link protocol with a suitable error control mechanism has to be used between the mobile terminal and the access entity. Such protocol eliminates the influence of high and correlated bit error rate using techniques like forward error correction (FEC), automatic repeat request (ARQ) or a combination of them (hybrid ARQ). However, the choice of the protocol at the wireless link and, in general, its operational parameters strictly depend on both wireless link error parameters and time. Therefore, in order to choose the wireless link layer protocol and to define its parameters properly, in addition to integrated traffic model, we have to introduce an error model of the wireless links of those networks, which are planned to use in 4G All-IP networks.

There are a number of wireless link models proposed in recent literature starting from pioneered work of Gilbert. The most convenient way to deal with wireless link is to assume that the bit errors occur with constant probability p for every bit in received sequence. However, later it was shown by Gilbert that bit errors are correlated rather than independent and, therefore, such simple model cannot adequately describe the real wireless link.

It was Gilbert [6] who first proposed to use a Markov chain to describe a real bit correlations in wireless links. His model reflect the correlation link to the some extent and for this purpose requires two states of discrete-time Markov chain one of which is error-free ($p_0=0$) and another is associated with bit error probability $p_1>0$. Later, this model has been extended by Elliot [7] who allowed the error free state of the Gilbert model to have probability of bit error more than zero ($p_0>0$). The model became more flexible in terms of its application area capable of modeling more sophisticated error sequences and correlations. However, such model still has only two states of Markov chain and, therefore, the range of bit error correlations are limited. The next expansion came from Fritchman [8], who allowed an arbitrary number of error free states of Markov chain in Gilbert model and, therefore, the range of bit-error correlations was again significantly expanded. The last

extension was made in middle 90th when Fang verified [9] a versatile Markov based wireless link model by allowing an arbitrary number of error states.

All abovementioned wireless link models were used in performance evaluation of link layer protocols. However, all of these studies were based on steady-state behavior of Markov chain modeling the bit errors. Such analysis, of course, gives us some basic ideas regarding the performance of link layer protocols. However, it is rarely occurs in practice in mobile networks serving the movable users since the mobility patterns and session times differ significantly from user to user and application to application. Therefore, in order to predict the network performance and to adequately predict the QoS degradation experienced by nomadic users we have to integrate the wireless link model into the user integrated traffic model of 4G All-IP users.

Therefore, in general, the time varying nature of wireless link also depends on user mobility, cannot be neglected and considered independently of user traffic model in future 4G All-IP networks. The one possible way to do that is to analyze the mobility patterns of current mobile users. For example, for highly movable users we can expect some periodicity in cell boundaries crossing and, therefore, we can predict to the some extent the sequence of state visited by Markov error model. The other possibility is to incorporate the error model into the integrated traffic model. Mathematically such model can depend on or be incorporated into integrated traffic model of 4G All-IP mobile user. Finally, the performance evaluation of 4G All-IP networks can be graphically represented as shown in Fig. 4.

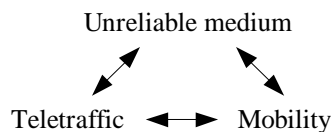


Fig. 4. User traffic model in 4G All-IP network.

C. Fixed Part of the RAN

While changing the path at the wireless part of the RAN, the traffic path within the fixed part of the RAN can also change due to user mobility. Therefore, the network has to be able to characterize the performance parameters of new path and to inform mobile terminals about it.

This can be done by using the classic teletraffic methods presented in literature and taking into account all possible paths of user traffic. Form such point of view it is crucial to provide

high-level QoS-based description of isolated IP networks or domains. For example, it can be useful to characterize the IP network in terms of loss and end-to-end delay. If the used QoS framework is DiffServ it is possible to characterize the end-to-end QoS parameters based on service level agreements.

The means how this information can be delivered to mobile terminal depend on QoS framework implementation in 4G All-IP networks.

D. Performance Evaluation of the CN

Core network performance parameters such as probability of loss and probability of certain delay within the network element or end-to-end performance parameters can be computed using the existing techniques which were successfully applied in fixed QoS-aware networks like ATM or DiffServ-enabled Internet domains.

VI. CONCLUSIONS

In this paper we have identified requirements induced on user traffic model in developing 4G All-IP networks. We have also considered the performance evaluation issues of these networks. We have concluded that performance evaluation of both fixed part of the RAN and core network can be based on existing techniques but requires certain new definitions and performance measures like IP domain's end-to-end delay etc.

REFERENCES

- [1] R. Braden, D. Clark, S. Shenker, "Integrated Services in the Internet Architecture: an Overview," RFC 1633, June 1994.
- [2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services," RFC 2475, December 1998.
- [3] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, "Assured Forwarding PHB Group," RFC 2597, June 1999.
- [4] B. Davie, A. Charny, J. Bennett, K. Benson, J. Le Boudec, W. Courtney, S. Davari, V. Firoiu, D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)," RFC 3246, March 2002.
- [5] N. Antunes, A. Pacheco, R. Rocha, "An integrated traffic model for multimedia wireless networks," *Computer Networks*, Vol. 38, pp. 25-41, 2002.
- [6] E. Gilbert, "Capacity of a burst-noise model," *Bell.Syst. Tech. J.*, Vol. 39, pp. 1253-1265, 1960.
- [7] E. Elliott, "Estimates of error rates for codes on burst-noise channel," *Bell.Syst. Tech. J.*, pp. 1977-1997, 1963.
- [8] B. Fritchman, "A binary channel characterization using partitioned Markov chain," *IEEE Trans. Inform. Theory*, Vol. 13, pp. 221-227, 1967.
- [9] H. Wang, "On verifying the first-order Markovian assumption for a Rayleigh fading channel model," in *Proc. ICUPC'94*, pp. 160-164, 1994.