Formalizing Set of Multiservice Models for Analyzing Pre-emption Mechanisms in Wireless 3GPP Networks

Konstantin Samouylov, Irina Gudkova(✉), and Ekaterina Markova

Peoples' Friendship University of Russia, Moscow, Russia
{ksam,igudkova,emarkova}@sci.pfu.edu.ru
http://aipt.sci.pfu.edu.ru/en

Abstract. Users of wireless 3GPP LTE networks are provided with a wide range of multimedia services with varying QoS requirements; due to this fact a problem of an effective network resources' distribution arises and, consequently, a task of the optimal RAC schemes development. According to the international standards, two types of services are defined within LTE networks – GBR services and non-GBR services. The GBR services generate streaming traffic and non-GBR services – elastic traffic the bit rate of which can dynamically change depending on the cell load. Also, the service priorities differ and are organized with the help of different mechanisms, e.g. service interruption mechanism and mechanism of bit rate degradation. The paper proposes a formal unique description of RAC schemes that is used to develop an example set of models realizing three possible pre-emption based scenarios in multiservice wireless networks.

Keywords: Wireless network · Radio admission control · RAC · Guaranteed bit rate · GBR · Non-GBR · Service interruption · Bit rate degradation

1 Introduction

Mobile 4G (4th generation) networks based on LTE (Long Term Evolution) technology [1] provide a wide range of multimedia services with varying quality of service (QoS) requirements. This is one of the most important trends in the modern telecommunication systems and networks modernization. A new classification of multiservice traffic identification has been proposed in LTE networks for reflecting the varying demands of users. According to the 3GPP (3rd Generation Partnership Project) consortium recommendations (TS 36.300, TS 23.401, TS 23.203), fifteen types of services are singled out in LTE. They are characterized by different service priorities and network resource requirements: guaranteed bit rate (GBR) services and non-GBR services.

The reported study was funded by RFBR according to the research project No. 16-37-00421 mol_a.

V. Vishnevsky and D. Kozyrev (Eds.): DCCN 2015, CCIS 601, pp. 61–71, 2016.
DOI: 10.1007/978-3-319-30843-2_7

The GBR services are real time services, e.g. voice telephony, video telephony, or real time gaming, for which a minimum bit rate value is specified, i.e. guaranteed bit rate. Nevertheless, when free resources are available in the network, the instantaneous bit rate can exceed the minimum specified one, but can not exceed some threshold value known as the maximum bit rate (MBR).

The non-GBR services are those for which no minimum (guaranteed) bit rate value is specified, the instantaneous bit rate can vary depending on the cell load. It is determined with the help of the so-called aggregate maximum bit rate (AMBR), allowing to differentiate services by the service priority levels. So, the bit rate cannot exceed the maximum value specified for user equipments (UE-AMBR) or determined by network characteristics (access point name AMBR, APN-AMBR). Examples of such services include e-mail, web browsing, or interactive gaming.

Various radio resource management (RRM) mechanisms are used to guarantee QoS. One of these mechanisms is the radio admission control (RAC) [2,4–10] aimed at admitting or rejecting user requests for service taking into account the limited frequency bands and varying QoS requirements. For this purpose, the different priority levels (allocation and retention priority, ARP) are assigned to multimedia services and define the pre-emption RAC algorithms within the corresponding RAC schemes. According to 3GPP TS 23.203: "The range of the ARP priority level is 1 to 15 with 1 as the highest level of priority. The pre-emption capability information defines whether a service data flow can get resources that were already assigned to another service data flow with a lower priority level. The pre-emption vulnerability information defines whether a service data flow can lose the resources assigned to it in order to admit a service data flow with higher priority level. The pre-emption capability and the pre-emption vulnerability can be either set to 'yes' or 'no'." In accordance with this definition of ARP, it is evident that in case of the lack of resources already occupied by lower priority services a new arrived higher priority request could be accepted, at best, through service bit rate degradation (partial pre-emption) [4,6,7,9,10], or at worst, through service interruption (full pre-emption) [4,7,8]. Other mechanisms such as reservation [5,6], threshold [5], and probabilistic management [5,9] could also be applied to realize the priority-service discipline.

Since the number of services LTE users are interested in varies as well as the services themselves within this specified range varies, more than 200 RAC schemes could be set up considering all possible pairwise influences of users on each others due to various pre-emption algorithms. Taking into account this fact and the annual growth of mobile traffic, it seems that the increasing need for the development of a general model considering the overall possible priority mechanisms could be transformed into the need for a set of specific models reflecting the predefined required services and mechanisms. Nevertheless, a unique notation is needed to describe at least, firstly each service within a RAC scheme, and secondly, the pre-emption mechanism applied to each pair of services. This is exactly the main purpose of the paper, i.e. to propose a such notation (Sect. 2) and to illustrate its usage with an example set of specific models of RAC schemes (Sects. 3 and 4).

2 Formal Description of RAC Models

Thereby, each RAC scheme is characterized by the number $K \in \{1, \ldots, 15\}$ of services (Table 1), as well as by the pre-emption scenario that is realized with the use of bit rate degradation and service interruption mechanisms.

Table 1. Characteristics of LTE service types (TS 23.203)

QCI	Resource type	Priority level	Examples of services	
1		2	Conversational voice	
2	GBR	4	Conversational live streaming video	
3		3	Real time gaming	
4		5	Non-conversational buffered streaming video	
65		0.7	Mission critical user plane push to talk voice	
66		2	Non-mission-critical user plane push to talk voice	
5		1	IMS signalling	
7		7	Voice, live streaming video, interactive gaming	
6	Non-GBR	6	Buffered streaming video, TCP-based applications (e.g., www, e-mail, chat, FTP, P2P file sharing, progressive video, etc.)	for multimedia priority services subscribers
8		8		for premium subscribers
9		9		for non-privileged subscribers
69		0.5	Mission critical delay sensitive signalling	
70		5.5	Mission critical data	

Note that for GBR services increasing or decreasing bit rate from a GBR value to a MBR one does not affect the mean service time. At the same time, changing bit rate for non-GBR services results in changing the mean transfer time. So, two types of traffics can be singled out: streaming traffic [11,12] that is characterized by fixed values of bit rate and time and elastic traffic [2] that is characterized by a fixed value of file size and a variable data transfer time [3]. Traffic can be transferred in two modes – unicast mode or multicast mode. For the last one two disciplines have been analyzed [4]: T1 – multicast session is closed when the first user who has opened the session leaves it, – and T2 – multicast session is closed when the last user leaves the session. T stands for the so-colled "transparent" service discipline in queuing theory.

Each of fifteen services could be described with the help of the following parameters:

$$\mathcal{S} =< \text{priority level, resource type, data transfer mode} > .$$

Assume that any RAC scheme is defined by the following mechanisms: bit rate degradation (partial pre-emption) and service interruption (full pre-emption). If k is the pre-emption capable service and k' is the pre-emption vulnerable service, than the pre-emption algorithm could be defined by the following matrix: $\mathbf{P} = (\mathbf{p}(k,k'))_{k,k'=1,\ldots,K} = (d(k,k'), i(k,k'))_{k,k'=1,\ldots,K}$, where

$$d(k,k') = \begin{cases} 1, \text{users of service } k' \text{ will be degraded,} \\ \quad \text{when a request for service } k \text{ arrives,} \\ 0, \text{otherwise,} \end{cases}$$

$$i(k,k') = \begin{cases} 1, \text{users of service } k' \text{ will be interrupted,} \\ \quad \text{when a request for service } k \text{ arrives,} \\ 0, \text{otherwise.} \end{cases}$$

3 Example Set of Markov Models of RAC Schemes

Let us illustrate the proposed formal description by an example set of models namely three Markov models (Fig. 1) with different traffic types, namely: unicast streaming, multicast streaming, and elastic traffics, as well as some combinations of different pre-emption mechanisms – first of all, bit rate degradation and service interruption.

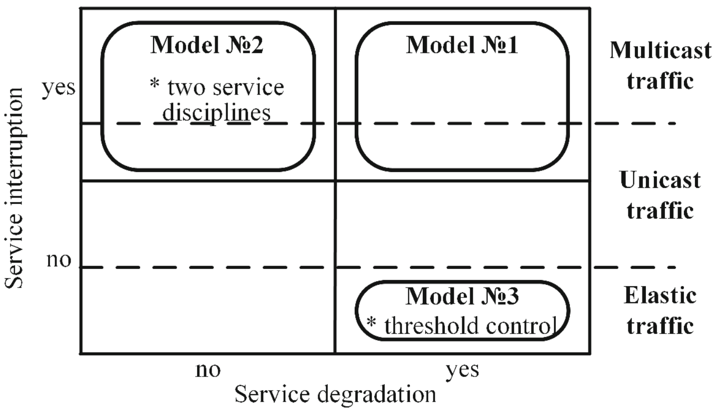

Fig. 1. Set of the RAC schemes.

Let us consider an LTE cell having the capacity of C bps. The incoming requests for service generate unicast, multicast, and elastic traffics and arrive as

Poisson processes. Resource occupancy times by unicast and multicast traffics as well as elastic file sizes are exponentially distributed. Bit rates are $d = 1$ and b bps for unicast and multicast traffics and e bps corresponds to the MBR for elastic traffic. The system states of models will be described by vectors with the following components: n – number of unicast users, m – state of a multicast connection ($m = 1$, if at least one user is being served; $m = 0$, otherwise), u – number of transferred elastic files.

3.1 Model of RAC Scheme with Multicast Traffic Degradation

Let us consider the first model (Fig. 2) [4]. Suppose that users are provided with two services that generate streaming traffic. The service that has a higher priority level ($\mathcal{S}_1$) (e.g. video conference) generates multicast traffic, and the service that has a lower priority level ($\mathcal{S}_2$) (e.g. video on demand) generates unicast traffic. The number of resources allocated for the establishment of a multicast connection is adaptively changed along a certain specified value set of $b_1 > ... > b_k > ... > b_K$ bps and is determined in accordance with the number of free resources – $\max\{b_k,\ k = 1, ..., K:\ b_k \le C - n\}$.

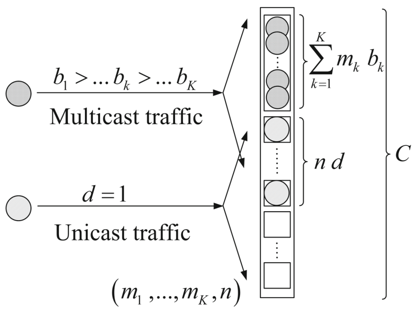

Fig. 2. Model of RAC scheme with multicast traffic degradation.

The RAC scheme is realized by two mechanisms of priority admission control – degradation of multicast bit rate and interruption of unicast users. So, the scheme could be defined in the following manner:

$$K = 2,\ \mathcal{S}_1 =< 4,\ \text{GBR},\ \text{multicast} >, \mathcal{S}_2 =< 5,\ \text{GBR},\ \text{unicast} >,$$

P	1	2
1	(0, 0)	(0, 1)
2	(1, 0)	(0, 0)

The admission control is realized in two stages. The first stage is the degradation of the number of resources occupied by multicast traffic down to the minimum value of b_K bps. The second stage is the interruption of $b_K - (C - dn)$

unicast users in case of the lack of resources. The state of a multicast connection is described as $m_k \in \{0,1\}$, $k = 1, ..., K$: $m_k = 1$ if connection is established and occupies b_k bps, $m_k = 0$ if connection does not occupy b_k bps. The system state space of the corresponding Markov model has the following form

$$\mathcal{X}_1 = \{(\mathbf{m}, n): \quad \mathbf{m} = \mathbf{0},\ n = 0, \ldots, C,\ \ \mathbf{m} = \mathbf{e}_1,\ n = 0, \ldots, C - b_1, \\ \mathbf{m} = \mathbf{e}_k,\ n = C - b_{k-1} + 1, \ldots, C - b_k,\ k = 2, \ldots, K\}.$$

3.2 Model of RAC Scheme with Unicast Traffic Interruption

Let us consider the second scheme of the set (Fig. 3) with two multicast traffic disciplines – T2 (e.g. real time gaming $\mathcal{S}_1$), and T1 (e.g. video conference $\mathcal{S}_2$), – and unicast traffic (e.g. video on demand $\mathcal{S}_3$).

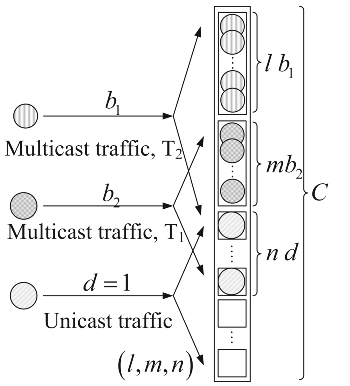

Fig. 3. Model of RAC scheme with unicast traffic interruption.

The RAC scheme realizes the mechanism of unicast traffic interruption in case of the lack of resources required for the establishment of a multicast T1 or T2 connection; the admission control is defined by the following services' parameters and matrix:

$$K = 3, \mathcal{S}_1 =< 3,\ \text{GBR, multicast} >,$$
$$\mathcal{S}_2 =< 4,\ \text{GBR, multicast} >, \mathcal{S}_3 =< 5, \text{GBR, unicast} >,$$

P	1	2	3
1	(0, 0)	(0, 0)	(0, 1)
2	(0, 0)	(0, 0)	(0, 1)
3	(0, 0)	(0, 0)	(0, 0)

Let us denote as $l \in \{0, 1\}$ the state of a multicast connection with discipline T2, as b_1 and b_2 bps the number of resources necessary for the establishment of multicast T1 and T2 connections, respectively. Then the state space has the following form

$$\mathcal{X}_2 = \{(l, m, n) \in \{0,1\} \times \{0,1\} \times \{0,1,...,C\} : b_1 m + b_2 l + dn \leq C\}\,.$$

3.3 Model of RAC Scheme with Elastic Traffic Degradation

Unlike the previous sections where we have proposed the models with streaming traffic, let us construct a scheme with two types of services that generate elastic traffic (Fig. 4), and with MBR thresholds $e_1 > e_2$ bps [10].

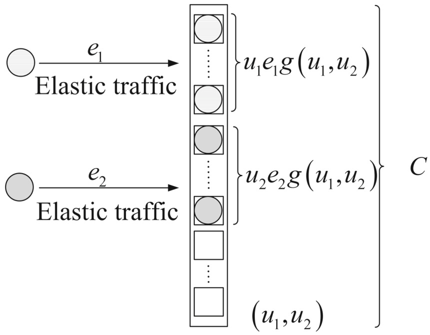

Fig. 4. Model of RAC scheme with elastic traffic degradation.

The number of resources occupied by files can dynamically vary from a maximum to a minimum value that is necessary to guarantee requirements for the mean transfer time and is realized through a threshold U of the number of transferred files. In case of low cell load, the maximum number or resources is occupied for files transfer. If the load rises so that it is no longer possible to guarantee a MBR for each transferred file, the instantaneous bit rate is degraded proportionally to an individual MBR, in accordance with some coefficient of bit rate degradation. Because of this, the RAC scheme and corresponding state space of the Markov model have the following forms:

$$K = 2, \mathcal{S}_1 =< k,\ \text{non-GBR, elastic} >, \mathcal{S}_2 =< k,\ \text{non-GBR, elastic} >,$$

P	1	2
1	(0, 0)	(1, 0)
2	(1, 0)	(0, 0)

$$\mathcal{X}_3 = \{\mathbf{u} \geq \mathbf{0} : \ \ u_1 + u_2 \leq U\}\,.$$

4 Numerical Analysis of Performance Measures

Let us consider a LTE network cell with the peak capacity equal to 100 Mbps. For our experiment, we have chosen the capacity of the cell with streaming or elastic traffic according to the data provided by the Cisco Systems company [13]. For example, in 2015, 58 % of the global mobile traffic is video traffic and 25 % is data traffic. It means that the capacity of the cell with streaming or elastic traffic is equal to 58 Mbps or to 25 Mbps respectively.

Based on Cisco Systems report, as well as on the requirements of the program Skype [14], which are recommended for providing video conference service to seven and more people, we can obtain initial data for numerical analysis (Table 2). In view of fact that the model No.1 and the model No.3, namely models with multicast and elastic traffic degradation, are the most general ones out of the models considered in Sect. 2, let us consider the numerical analysis for these models' performance measures – namely, mean bit rate of multicast streaming traffic and mean coefficient of the bit rate degradation for elastic traffic.

Table 2. Example of initial data for numerical analysis

Parameter	Example of service	Traffic type	
		Streaming	Elastic
C		59 Mbps	38 Mbps
d	video on demand	2 Mbps	–
b	video conference	8 Mbps, 6 Mbps, 4 Mbps	–
e_1	service within 4G network	–	6 Mbps
e_2	service within 3G network	–	3 Mbps

According to the forecasts by Cisco Systems, mobile applications providing video services will generate most of global mobile data traffic by 2018, namely about 69 %. However, not only in the future, but also in the beginning of 2012, mobile video represents more than half of global mobile data traffic. In this regard, let us consider an example of a single cell supporting video conference and video on demand services. One of the main trends of 5G networks is the need to save resources to allocate for high quality video services. This is possible through the implementation of MBMS (Multimedia Broadcast and Multicast Service) subsystem.

To illustrate the behavior of the models' performance measure – mean bit rate $\overline{b}$ – let us summarize the initial data of the numerical example in the Table 3. Having such initial set $\{8\text{ Mbps},\ 6\text{ Mbps},\ 4\text{ Mbps}\}$ of bit rates, we can distinguish seven variants of bit rate sets $(K; b_1, ..., b_K)$: $(1;\ 4)$, $(1;\ 6)$, $(1;\ 8)$, $(2;\ 6, 4)$, $(2;\ 8, 4)$, $(2;\ 8, 6)$, $(3;\ 8, 6, 4)$ (we omit units of measurement). The Fig. 5 shows also the combination of bit rates sets into three groups but already according to maximum bit rate b_1.

Table 3. Numerical data for model with multicast traffic degradation

Traffic type	Service	Traffic rate	Initial set of requirements	Resource occupation
Unicast	video on demand	90 %	2 Mbps	2 h
Multicast	video conference	10 %	8 Mbps, 6 Mbps, 4 Mbps	1 h

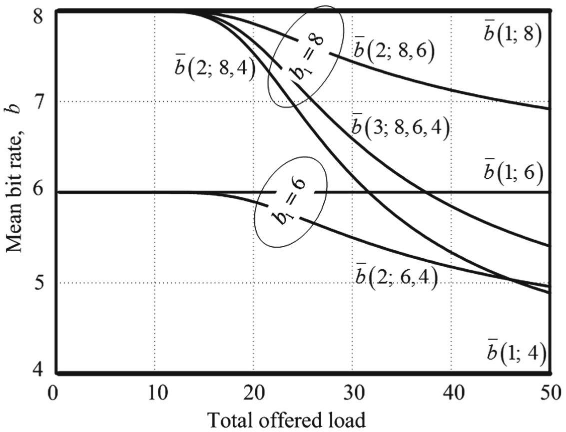

Fig. 5. Mean bit rates of video conference service.

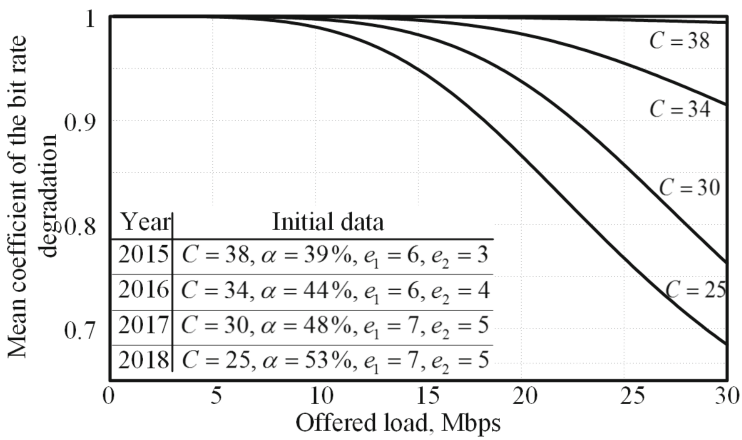

Fig. 6. Mean coefficient of the bit rate degradation for elastic traffic.

To analyze the RAC scheme for the model with elastic traffic, let us define maximum bit rate values e_1 and e_2 as the minimum rate values based on "group of devices" and "technology" criterion (e.g. 4G and 3G, respectively). Note that the ITU-T G.1010 contains information on the preferred (15 s) and acceptable (60 s) data transfer time values for a 10 MB file, which allows us to define a limit on the minimum elastic traffic bit rate and, consequently – on the maximum number of the transferred data blocks U. Suppose, in this case the ratio α is between the volume of traffic generated within a 4G network and the total volume of traffic generated within the 2G and 3G networks. Then, for the annual range of 2015-2018 [13], according to the Cisco Systems data, the values of α can be obtained. Using the obtained initial data, let us conduct the analysis of the mean coefficient of the bit rate degradation. The Fig. 6 shows that the growth of the total offered load results in decreasing the mean degradation factor.

5 Conclusion

The paper discusses the principle of RAC schemes construction within multiservice networks with the pre-emption based admission control. A set of three Markov models of RAC schemes with unicast streaming, multicast streaming, and elastic traffics was proposed. Different scenarios of priority admission control were analyzed, that are based, first and foremost, on the mechanism of bit rate degradation and service interruption of users with lower priority levels. In addition to the results described in the paper, a unique input data for numerical experiments is developed for further performance analysis.

Currently, the understanding and specifications of the fifth-generation networks are formed. In these specifications there are new types of services, generating traffic, which does not always fit into the existing classification of LTE network services, for example, machine-to-machine traffic. Thereby, it is assumed that in future releases of 3GPP specifications the number of services will be expanded again, and the formal description of RAC schemes suggested in the paper could be also applied.

References

1. Stasiak, M., Glabowski, M., Wisniewski, A., Zwierzykowski, P.: Modelling and Dimensioning of Mobile Wireless Networks: From GSM to LTE. Willey, Chichester (2010)
2. Samouylov, K., Gudkova, I.: Recursive computation for a multi-rate model with elastic traffic and minimum rate guarantees. In: IEEE International Congress on Ultra Modern Telecommunications and Control Systems ICUMT-2010, pp. 1065–1072. IEEE Press, Moscow (2010)
3. Basharin, G., Gaidamaka, Y., Samouylov, K.: Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks. Autom. Control Comput. Sci. **47**(2), 62–69 (2013)

4. Borodakiy, V., Gudkova, I., Samouylov, K., Markova, E.: Modelling and performance analysis of pre-emption based radio admission control scheme for video conferencing over LTE. In: 6th International Conference ITU Kaleidoscope: Living in a Converged World - Impossible Without Standards?. K-LCW-2014, pp. 53–59. ITU Press, St. Petersburg (2014)
5. Ghaderi, M., Boutaba, R.: Call admission control in mobile cellular networks: a comprehensive survey. Wirel. Commun. Mob. Comput. **6**(1), 69–93 (2006)
6. Gudkova, I.A., Samouylov, K.E.: Modelling a radio admission control scheme for video telephony service in wireless networks. In: Andreev, S., Balandin, S., Koucheryavy, Y. (eds.) NEW2AN/ruSMART 2012. LNCS, vol. 7469, pp. 208–215. Springer, Heidelberg (2012)
7. Khabazian, M., Kubbar, O., Hassanein, H.: A fairness-based pre-emption algorithm for LTE-advanced: In: 10th IEEE Global Telecommunications Conference GLOBECOM-2012, pp. 5320–5325. IEEE Press, Anaheim (2012)
8. Kwan, R., Arnott, R., Trivisonno, R., Kubota, M.: On pre-emption and congestion control for LTE systems. In: 72nd Vehicular Technology Conference VTC2010-Fall, pp. 6–8. IEEE Press, Ottawa (2010)
9. Liao, H., Wang, X., Chen, H.H.: Adaptive call admission control for multi-class services in wireless networks. In: IEEE International Conference on Communications ICC-2008, pp. 2840–2844. IEEE Press, Beijing (2008)
10. Shorgin, S.Y., Samouylov, K.E., Gudkova, I.A., Markova, E.V., Sopin, E.S.: Approximating performance measures of radio admission control model for non real-time services with maximum bit rates in LTE. In: AIP Conference Proceedings, vol. 1648, pp. 1–4. AIP Publishing, USA (2015). doi:10.1063/1.4912508
11. Gudkova, I., Plaksina, O.: Performance measures computation for a single link loss network with unicast and multicast traffics. In: Balandin, S., Dunaytsev, R., Koucheryavy, Y. (eds.) ruSMART 2010. LNCS, vol. 6294, pp. 256–265. Springer, Heidelberg (2010)
12. Iversen, V.: Teletraffic Engineering and Network Planning. Technical University of Denmark, Lyngby (2011)
13. Cisco Systems: Cisco visual networking index: Global Mobile Data Traffic Forecast Update, 20132019: usage: White paper. Cisco Systems, 40 (2015)
14. Skype 2015: How much bandwidth does Skype need?. https://support.skype.com/en/faq/FA1417/how-much-bandwidth-does-skype-need