Queuing Models for Dimensioning Interactive and Streaming Services in High-Speed Downlink Packet Access Networks

Sonia Aïssa, *Senior Member, IEEE*, and Ghassane Aniba, *Student Member, IEEE*

***Abstract*—We consider modeling the statistical behavior of interactive and streaming traffics in High-Speed Downlink Packet Access (HSDPA) networks. Two important applications in these traffic categories are web-browsing (interactive service) and video streaming (streaming service). Web-browsing is characterized by its important sensitivity to delay. Video streaming on the other hand is less sensitive to delay, however, due to its large frame sizes, video traffic is more affected by the packet loss resulting from a limited buffer size at the base station. Taking these characteristics into account, we consider modeling the queuing delay probability density function (PDF) of the web-browsing traffic, and modeling the queuing buffer size distribution of video streaming traffic. Specifically, we show that the queuing delay of the web-browsing traffic follows an exponential distribution and that the queuing buffer size of video streaming traffic follows a weighted Weibull distribution. Model fitting based on simulated data is used to provide simple mathematical formulations for the different parameters that characterize the PDFs under consideration. The provided equations could be used, directly, in HSDPA network dimensioning and, as a reference, to satisfy a certain quality of service (QoS).**

***Index Terms*—HSDPA, network dimensioning, queuing buffer size modeling, queuing delay modeling, scheduling, video streaming, web-browsing.**

I. Introduction

In 3G+ networks [1], packet-based data services have different requirements in terms of quality of service (QoS) and tolerance to delay. In order to satisfy the diverse requirements of future services, new transmission techniques are introduced in the new and promoting 3G+ technologies. One of the most supported technologies is high-speed downlink packet access (HSDPA), the new enhancement of the WCDMA downlink. HSDPA introduces new adaptive link techniques, namely, adaptive modulation and coding (AMC) and hybrid automatic retransmission query (HARQ), as means to increase a user equipment (UE) air throughput up to 10 Mbps. The air interface throughput in HSDPA networks is shared between different UEs. In order to increase the capacity of these networks, their capability to carry packet-switched traffic is used along with multiplexing the traffic of different UEs. The use of AMC and fast HARQ is intended to improve the HSDPA air throughput and reduce, at the same time, the delay introduced by the queuing and retransmission processes. However, overloading the network causes overflow of the queuing buffer at the base station and/or high queuing delay which, in turn, can cause degradation in the overall performance of the network. Hence, modeling the traffic of packet-based data services as well as their statistical behavior at the base station is of major importance. In fact, such models are needed in order to dimension practical networks and design efficient resource management strategies and communication protocols for these networks.

There exist four types of services: conversational, interactive, streaming and background. In this work, we focus on streaming and interactive services since they are those which differ the most from circuit-switched services, such as conversational services, while having specific QoS requirements unlike background services. In this context, this paper considers the two main HSDPA applications of interactive and streaming traffics, and is accordingly organized into two parts. The first is dedicated to modeling the queuing delay probability density function (PDF) of interactive services, and precisely the web-browsing traffic, and the second is devoted to modeling the distribution of the queuing buffer size of streaming services, namely, MPEG-4 encoded video streaming traffic.

For this purpose, the web-browsing traffic model provided in [2] is used to derive the parameters that specify the queuing delay PDF corresponding to the service under consideration. The web-browsing traffic can be modeled over three layers: session, packet call, and packet data, with each layer defined by its own PDF and parameters. In [3], it is shown that the packet-data delay of web-browsing traffic follows a weighted exponential distribution. Herein, we consider the delay at the packet call level, which represents the delay for browsing a web page, and address the queuing delay modeling taking into consideration key parameters, such as the *reading time*, i.e., the time needed for a user to read a web page, and packet retransmissions defined in terms of the packet loss probability.

The video traffic can be partitioned over scene layer, group of pictures (GOP) layer and frame layer, and modeled taking into consideration the intra and inter-GOP correlations. In general, queues corresponding to this kind of traffic are characterized by *weak stability* due to the wireless channel variability [4]. Indeed, the base station can not guarantee an instantaneous air data rate for the transmission of video frames at each transmission time interval (TTI), but rather an average transmission rate.

Manuscript received April 14, 2006; revised April 19, 2006. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grants (DG) program of Canada.

The authors are with INRS-EMT, University of Quebec, Montreal (QC), H5A 1K6, Canada (e-mail: aissa@emt.inrs.ca; ghassane@emt.inrs.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2007.903611

On the other hand, it was shown in [4] that the intra-GOP correlation (fast time-scale) has little impact on the queuing modeling for a weak stability scenario. Thus, we neglect the effect of the intra-GOP correlation, and focus for our modeling on the inter-GOP correlation. In particular, we show that the latter follows an $M/G/\infty$ model [5] with autocorrelation function (ACF) of the form $\rho(k) \sim e^{-\beta\sqrt{k}}$, and explain the contradictions noted in previous works [5]–[7] though the modeling therein considers the same movie trace.

Because of the complexity of the packet traffic involved, modeling the web-browsing queuing delay and the video-streaming queuing buffer size is analytically untractable. In [8], simulation-based model fitting is used to model the packet-data queuing delay PDF of video and voice traffic services. In this work, with the use of traffic models proposed herein and based on the approach followed therein, we propose a general queuing delay model for web-browsing traffic and a queuing buffer size model for video streaming traffic, in the context of transmission in HSDPA. For the web-browsing traffic, taking into consideration the use of discrete data rate values and ARQ retransmissions, we provide the different parameters that characterize the queuing delay PDF, as a function of the air throughput T, network loading defined by the number of UEs N, erroneous packet retransmission probability P_e, and the *reading time* parameter t_r. As for the video streaming traffic, we provide a general queuing buffer size distribution model, and define the relationship between its parameters as a function of the inter-GOP ACF coefficient β, and the aforementioned parameters N and T.

The remainder of this paper is organized as follows: Section II presents a brief description of the HSDPA transmission environment used in our simulations. Section III is devoted to the web-browsing traffic and presents its associated queuing delay modeling analysis and results. The video streaming traffic model and its queuing buffer size modeling are provided in Section IV, followed by concluding remarks drawn in Section V.

II. HSDPA Transmission Environment

A. System Model

In HSDPA, AMC is applied through the use of a mapping between the UE channel quality, defined by the signal-to-interference-and-noise ratio (SINR), and the modulation-level and coding-rate values that could be used in such channel condition [9]. Thirty couple values are defined, allowing thirty discrete rate values to be used, based on the channel quality indicator (CQI) that each HSDPA UE transmits on the uplink using the High-Speed Dedicated Physical Control Channel (HS-DPCCH). The CQI belongs to a set of thirty values used to map the UE's channel state, given by a SINR value which ranges over 30 dB in the interval [−5 dB, 25 dB], into a CQI value according to the following rule:

$$CQI \simeq \min\left(\max\left(0, \lfloor SINR_{\text{hsdsch}} + 6 \rfloor\right), 30\right), \qquad (1)$$

where $SINR_{\text{hsdsch}}$ is expressed in dB and $\lfloor \cdot \rfloor$ denotes the integer floor operator.

TABLE I
Mapping of Channel Quality Indicator (CQI) Into Transport Block (TB) Size

CQI	0	1	2	3	4	5	6	7
TB size (bits)	0	137	173	233	317	377	461	650
CQI	8	9	10	11	12	13	14	15
TB size (bits)	792	931	1262	1483	1742	2279	2583	3319
CQI	16	17	18	19	20	21	22	23
TB size (bits)	3565	4189	4664	5287	5887	6554	7168	9719
CQI	24	25	26	27	28	29	30	
TB size (bits)	11418	14411	17237	21754	23370	24222	25558	

Based on the CQI value, the Node-B assigns to UE i a maximum transmission rate r_i, or equivalently a transport block size TB_i (Table I), that could be used for transmission to the UE in the next TTI. The latter is of duration $T_{\text{TTI}} = 2$ ms [9].

As previously mentioned, HARQ is implemented in HSDPA in order to improve the performance of the network and reduce retransmission delays, using advanced combining techniques at the UE side to take advantage of each packet retransmission. When a packet is received in error, a Negative Acknowledgment (NACK) is transmitted to the base station. The scheduler decides then if the corresponding user will receive a retransmission for this packet in the next TTI. In this paper, the HARQ protocol is taken into consideration by means of an error transmission probability P_e, defining the probability that a UE receives an erroneous packet.

B. Transmission Mode

The choice of a scheduling algorithm for our queuing modeling is of a major importance. The main task of the scheduler is to select a UE for which data will be transmitted in a TTI. Given a specified objective in terms of throughput and/or fairness maximization, in order to make a decision, some scheduling algorithms such as the max-CIR [10] use the SINR of each UE, while some others use the instantaneous rate values (r_i) and their average values, for instance the Proportional Fairness method [11]. In this work, our focus being on dimensioning the network, the scheduling algorithm needs to be capable of providing a fair channel access to the UEs. A fair scheduler is defined as one that can provide, on the average, the same throughput for all users independently of their channels' quality, and that on a longer time-scale compared to the scheduling time scale. Such fairness can be achieved through Round Robin (RR) scheduling when the users exhibit the same channel quality on the average. The RR algorithm allocates the channel to UEs in a circular way, based on the CQI that is fed back from each UE as a measure of its corresponding SINR (1).

In a multi-cell network, a user's SINR accounts for the inter-cell interference and the user's channel gain including fast fading, shadowing and path loss effects. The channel quality can be different from a user to another, depending on the user's position within the cell and the fading characteristics of the link between the base station and the UE. The effects of large-scale variations and inter-cell interference can be assumed to be canceled out in average by using a fair scheduler and, accordingly, the analysis can be based on exploiting the multi-user diversity gain as a result of the small-scale variations of the channels. Hence, in our simulations, we consider that

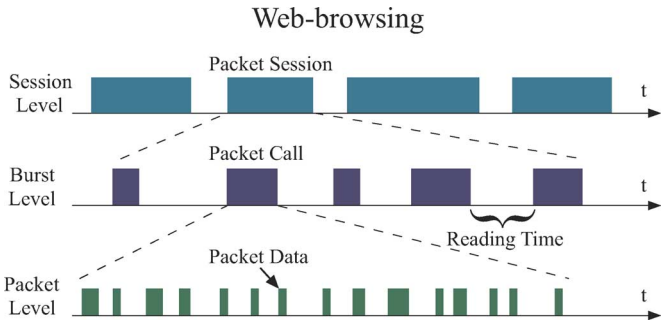

Fig. 1. Web-browsing layered traffic model.

all UEs exhibit the same channel variations. In particular, we consider that the instantaneous SINR of each user follows an exponential distribution (as a result of fast fading considered herein to follow Rayleigh distribution) with the average SINR value computed in order to achieve an average total throughput for all users equal to $R = 2$ Mbps. The latter corresponds to an average CQI value of 17 [9], but the results can be generalized for other values of R as will be seen later. Under the assumption of equal average channel quality for all UEs, the RR scheduler provides equal average data rates for all users [12], thus allowing to model the statistical behavior of the services under consideration in a fair transmission environment.

A queuing model for dimensioning HSDPA networks is a useful and practical tool, especially because of the existence of a standardized mapping table that limits the values of allowable discrete data rates [9]. By taking into consideration such limitation, the queuing models proposed herein will be general enough to be applied in realistic transmission scenarios to satisfy requirements in terms of QoS. On the other hand, in the case that users have different average channel qualities, our results can be used with any strict fair scheduling algorithm, such as the Adaptive Proportional Fairness (APF) [13] shown to provide any required level of fairness even under heterogeneous propagating conditions.

III. Web-Browsing Traffic

A. Web-Browsing Traffic Model

A general traffic model, based on the definition provided in [2] for the web-browsing service, is used. The model is divided into three levels (Fig. 1):

- *Session level*: user sessions (web-browsing) are modeled at this level. Session arrivals follow a Poisson distribution.
- *Burst level*: each packet session is formed by one or many *packet call(s)*, with each packet call representing a web page. The inter-arrival time between packet calls, named *reading time*, follows a Geometric distribution.
- *Packet level:* each packet call is composed of a number of data packets. The distributions of the packet size and of the inter-arrival time between packets are specified at this level.

In Table II, we provide a summary of the parameters considered for the web-browsing traffic. For each traffic parameter, the value provided corresponds to the mean of the random variable according to which the parameter is distributed. In order to provide a general packet-call delay model, we consider different arrival rates, namely, 144 kbps, 384 kbps and 2048 kbps, generated by choosing the value of the packet data inter-arrival time within a packet call equal to 0.0277 s, 0.0104 s, and 0.00195 s, respectively. Moreover, different values are considered for the reading time t_r (2 s, 4 s, 8 s, 16 s and 32 s). Higher values were also studied (64 s, 128 s and 256 s), nevertheless, only results corresponding to the values shown in Table II are presented, given that they provide the required accuracy for the proposed queuing delay model.

TABLE II
Parameters of the Web-Browsing Traffic

Parameter	Distribution	Value
Session arrival process	Poisson	5
Reading time (t_r)	Geometric	$\{2^n s\}_{n=1}^{5}$
Number of packets per call	Geometric	25
Packet inter-arrival time	Geometric	$0.0277s \cdots 0.00195s$
Arrival data rate	-	$144kbps \cdots 2048kbps$
Packet size	Pareto $(1.2, 81.5)$	480 bytes

B. Queuing Delay Modeling

As previously mentioned, following the model-fitting approach used in [3], [8], we provide results corresponding to the web-browsing service in HSDPA. Specifically, considering the simulation setting described in Section II, we provide the different parameters that characterize the queuing delay PDF as a function of the air throughput T, the number of UEs N, and the reading time t_r. In order to provide precision to our delay modeling, a large number of simulation runs, ranging between thirty and a hundred, was performed. On the other hand, unlike the packet-data queuing delay measurements performed in [8], measurement of the delay is performed herein for each packet call (web page) in each session.

Our simulations for the web-browsing traffic demonstrate that the packet-call queuing delay follows an exponential distribution (2) with μ denoting the mean packet-call queuing delay. This delay is defined by the time difference between the arrival of the first data packet in a packet call and the time its last data packet is correctly received. Fig. 2 shows the cumulative distribution function (CDF) of the packet-call delay using simulations along with its corresponding exponential CDF (3) for different values of N, given a reading time $t_r = 16$ s, an arrival data rate of 144 kbps, and $P_e = 0.5$.

$$f_\tau(\tau) = \frac{1}{\mu} \exp\left(-\frac{\tau}{\mu}\right). \tag{2}$$

$$F_\tau(\tau) = 1 - \exp\left(-\frac{\tau}{\mu}\right). \tag{3}$$

The mean packet-call queuing delay μ depends on the number of UEs N, the probability of correct packet transmission $P_c = (1 - P_e)$, and the reading time t_r. Fig. 3(a) shows the variation of μ as a function of N for different data arrival rates and reading time values, when $P_c = 0.5$. In order to alleviate the figure, plots corresponding to $t_r = 4$ s, and 16 s are shown for a data arrival rate equal to 144 kbps, while for $t_r = 2$ s, 8 s, and 32 s the

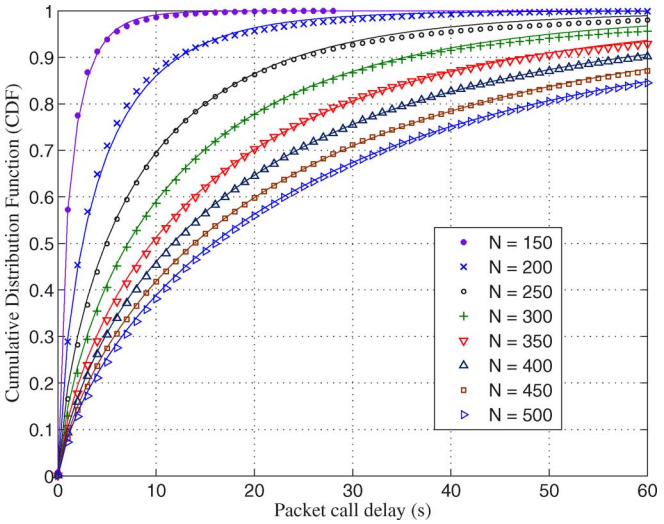

Fig. 2. Packet call delay CDFs based on simulation and their corresponding fitting curves, for different numbers of UEs, N.

curves illustrate the results for different arrival rates (144 kbps, 384 kbps and 2048 kbps). From these results, one important observation is that the mean delay μ is independent of the arrival rate. Indeed, for each reading time value, $t_r = 2$ s, 8 s, and 32 s, we get a superposition of the three mean delay curves corresponding to the considered arrival rates. This result can be explained by the inherent specificity of an interactive service. In fact, even if we change the arrival data rate, the volume of traffic corresponding to each packet remains the same; only the inter-arrival of data packets gets changed. Hence, for the considered high values of arrivals, we can suppose that the packet call arrives at the base station's buffer almost instantaneously and, as a consequence, the queuing delay will depend only on the volume of the packet call, which is the same in all cases. From Fig. 3(a), we also observe that when the number of UEs N exceeds a certain value, the mean delay μ becomes a linear function of N. This is one of the major differences between the mean packet-data delay [8] and the mean packet-call delay. Specifically, the former increases exponentially when the number of UEs gets higher whereas the latter increases linearly with N due to the interactive behavior of the web-browsing traffic. According to Table II, each packet call is formed by 25 data packets in average, with an average data packet size of 480 bytes. Hence, the average packet call size is equal to $25 \times 480 \times 8 = 96$ Kbits. Using Round Robin scheduling, each of the N users gets access to the channel $(N-1)$ TTIs after the last transmission he received from the base station. When the number of users is large, this waiting time becomes large enough for the packet call of a user to be received in whole in the user's corresponding buffer at the base station before the user is served. Thus, the average time that one packet call takes before being received at the UE, which is indeed the average queuing delay μ, can be approximated as

$$\mu \simeq \frac{\text{average packet call size}}{P_c * R} N + (N-1) T_{\text{TTI}} - t_r, \quad (4)$$

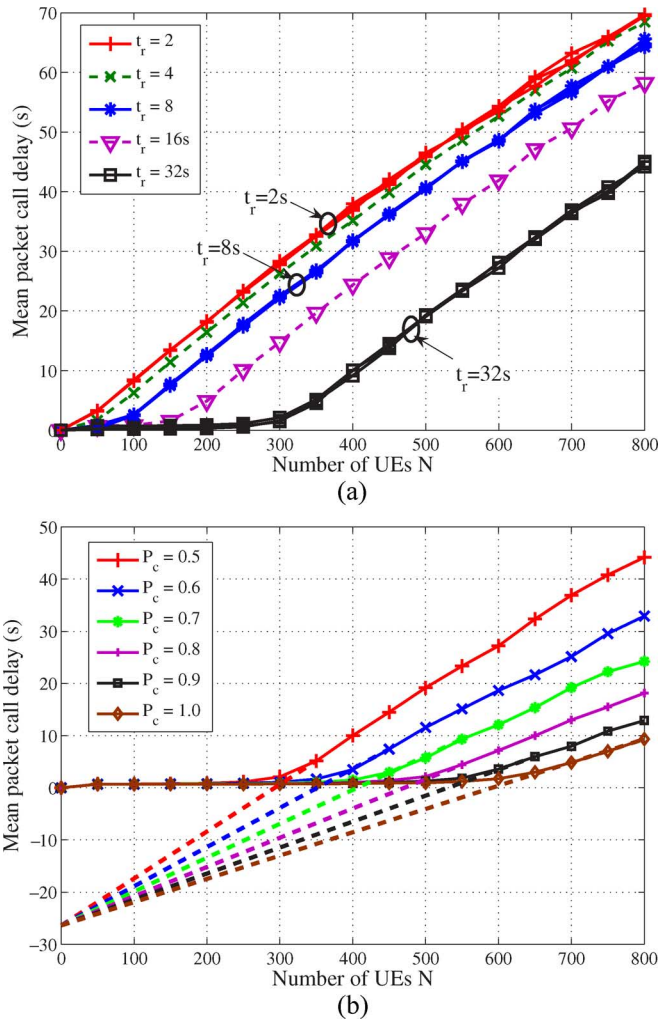

Fig. 3. Variation of the mean packet call delay μ as a function of the number of UEs: (a) for different reading time values t_r and data arrival rates; (b) for different values of the probability P_c, along with the slope of each linear approximation.

which illustrates a linear relationship between the average delay and the number of active users N as confirmed through simulations.

Considering an average packet-call delay value smaller than 1 s to be a negligible delay for the web-browsing service, we neglect values smaller than this threshold. Hence, as can be seen in Fig. 3(a), from a certain value N_0, the average μ becomes a linear function of N and, as such, can be expressed as

$$\begin{cases} \mu = \frac{1}{a} N - b & \text{if } N > N_0, \\ \mu = 0 & \text{if } N < N_0, \end{cases} \quad (5)$$

with $a[1/s]$ and $b[s]$ parameters used to define the linear model. The value N_0 defines the maximum number of allowed UEs when an average delay value smaller than 1 s is considered as one that meets the required QoS. This limiting value can easily be calculated using:

$$N_0 = a \times b. \quad (6)$$

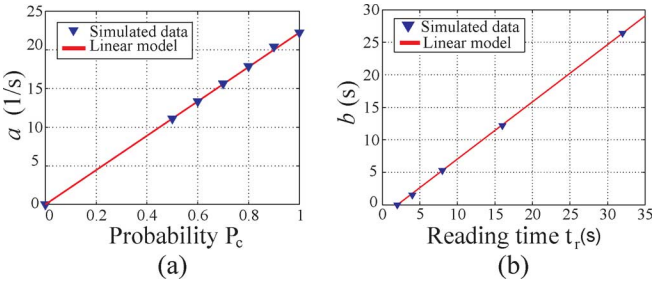

Fig. 4. Web-browsing model parameters: (a) parameter a as a function of the probability P_c; (b) parameter b as a function of the reading time t_r.

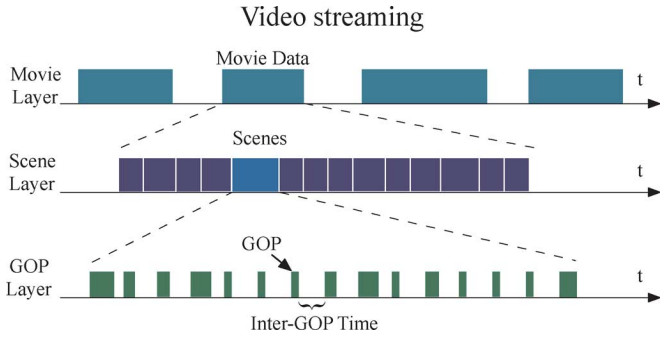

Fig. 5. Video-streaming layered traffic model.

We now present a formulation of the parameters a and b as a function of the air throughput T, the reading time t_r, and the probability P_c. Before proceeding, two observations must be made. Firstly, in Fig. 3(a), we observe that for all reading time values, the slope of the μ curve is the same, which means that the parameter a is independent of the reading time. Moreover, Fig. 3(b) shows that for different values of P_c, the slope of the μ curve is different, which means that the parameter a depends on the probability P_c. In addition, intersection of all μ curves in the ordinate axis means that b is independent of P_c. These important observations simplify the modeling of the mean delay μ. Indeed, Fig. 4 shows that there is a linear relationship between the parameter a and the probability P_c on one hand, and between the parameter b and the reading time t_r on the other hand, expressed as:

$$a = 22.357\, P_c, \tag{7}$$

$$b = 0.88\, t_r + 1.87. \tag{8}$$

Therefore, (5) and (6) can respectively be written as:

$$\begin{cases} \mu = \frac{1}{22.357\, P_c} N - (0.88\, t_r + 1.87) & \text{if } N > N_0, \\ \mu = 0 & \text{if } N < N_0. \end{cases} \tag{9}$$

$$N_0 = 22.357\, P_c (0.88\, t_r + 1.87). \tag{10}$$

The air throughput is given by $T = P_c \times R$. The equations given above are formulated when $R = 2$ Mbps, therefore, for a general formulation we can write the mean delay μ as function of the air throughput T according to

$$\begin{cases} \mu = \frac{1}{11.175\, T} N - (0.88\, t_r + 1.87) & \text{if } N > N_0, \\ \mu = 0 & \text{if } N < N_0, \end{cases} \tag{11}$$

where T is expressed in Mbps and N_0 is given by

$$N_0 = 11.175\, T (0.88\, t_r + 1.87). \tag{12}$$

Equation (11) is the general analytic formulation of the mean delay μ. Indeed, for a certain number N of UEs, an air throughput $T = (1 - P_e) \times R$ and a reading time t_r, we can directly evaluate the PDF of the packet-call queuing delay τ using (2). By means of the latter, different statistics can be generated, for instance, the 95-percentile value:

$$\nu = -\mu \ln 0.05 = 3\, \mu. \tag{13}$$

Thus, in order to satisfy a certain QoS for the UEs, defined in terms of average delay (μ_0) or 95-percentile value (ν_0), the maximum number of active UEs $N_{\max}$ that the base station can serve while meeting the required QoS, can simply be determined by means of (11), (12) and (13). Based on the value of $N_{\max}$, the call admission control (CAC) module can determine its admit/reject decisions so as to keep the users' QoS at the required level.

IV. Video Streaming Traffic

A. Video-Streaming Traffic Model

The video-streaming traffic model, used in this paper and represented in Fig. 5, is divided into three layers:

- *Movie layer*: user sessions (video streaming) are modeled at this level. Arrivals are assumed to follow a Poisson distribution.
- *Scene layer*: each movie is formed by a number of scenes. A scene is the period of time over which the GOP statistics are approximatively constant. The duration of each scene follows a distribution in direct relation with the autocorrelation function (ACF) of the GOP level.
- *GOP layer*: each scene is composed of *Group Of Pictures (GOPs)*. The distribution of the GOPs are specified at this level. Each GOP is composed of three types of frames: Intra-coded (I), Predictive (P) and Bidirectional (B). The I frames are those for which intra-frame coding is used (without motion estimation), the P frames are those for which inter-frame coding (with motion estimation) with forward prediction only is used, and the B frames specify the frames that can be predicted using forward and backward prediction [14]. In general, I frames are larger in size than the P frames which, in turn, are larger than the B frames. The GOP length is fixed to 12 frames. With this and since we consider a frame rate of 25 f/s, the inter-GOP time follows to 480 ms.

In the following, we present a detailed description of the statistical behavior of the video streaming traffic at the scene and GOP levels.

In order to generate the video streaming traffic, we use the spatial renewal process (SRP) model [4], [15]. In the latter, a first background process, which is responsible for the time-dependence structure, characterizes the scenes' durations modeled

TABLE III
ESTIMATED PARAMETERS OF THE SCENE LAYER, OBTAINED BY SELF SIMILAR (SS) AND $M/G/\infty$ PROCESSES, AND THE GOP LAYER PARAMETERS USING LEAST SQUARE ERROR (LSE) FITTING

Trace	Scene level				GOP level	
	Self Similar		$M/G/\infty$		M	V
	β	LSE	β	LSE		
Alladin	0.5615	6.2743	0.2447	3.4998	10.0390	0.5897
ARDTalk	0.7001	5.0513	0.3196	3.1962	10.3231	0.2417
DieHardIII	0.6398	1.8365	0.3056	0.7850	10.5137	0.5023
Dusk	0.7197	4.3613	0.3485	3.1767	10.5283	0.4109
BoulevardBio	0.5935	12.4894	0.2293	8.22	10.5037	0.3365
FirstContact	0.8317	2.2806	0.4431	1.6992	9.7626	0.5143
Formula	0.9375	1.6027	0.5378	1.2976	10.7843	0.2843
Futurama (GOP: 500-2500)	0.6641	1.4484	0.3428	1.153	10.9855	0.2285
Futurama (GOP: 1000-2500)	0.8142	0.9371	0.5173	0.9570	10.9855	0.2285
Jurassic	0.6260	14.3892	0.256	9.5006	10.6389	0.4768
N3Talk	0.7422	1.1588	0.4648	1.116	10.3445	0.3191
RobinHood	0.6871	2.4252	0.3341	1.7314	10.8743	0.3278
SilenceOfLambs	0.3739	8.9788	0.1459	2.9857	10.2314	0.6018
Ski	0.7496	3.8907	0.4150	3.3629	10.7195	0.5387
Soccer	1.0867	1.7483	0.6633	1.6409	11.0610	0.3480
StarWarsIV	0.4949	2.0616	0.2171	0.7822	9.6646	0.3652
SusiUndStrolch	0.8363	1.4209	0.4289	0.8363	9.8921	0.4640
TheFirm	0.6385	21.1092	0.267	16.72	9.6371	0.5088
Troopers	0.5624	5.743	0.2692	4.2868	10.4019	0.4552

as independent and identically distributed (i.i.d.) sequence $\{d_n\}$ (the scene length d_n denotes the number of GOPs in the nth scene) described by the CDF $F_d(k)$ given by

$$F_d(k) = 1 - \frac{\rho(k) - \rho(k+1)}{1 - \rho(1)}, \tag{14}$$

where $\rho(k)$ is the GOP ACF. In the SRP model, a second background process, which is independent of the first, characterizes the GOP sizes through an i.i.d. sequence $\{S_n\}$ distributed to the desired steady-state marginal distribution of the video traffic [15]. The SRP is composed of a sequence of scenes where the nth scene is d_n in length, and the sample path during this period takes on values from the random variable S_n. Generation of the second process requires a huge amount of computer resources and complicates the model. Hence, we simplify the latter by using the *fluid version* in which the $\{S_n\}$'s are replaced by their means $\{\overline{S}_n\}$. This simplification does not affect the accuracy of the video model [4].

1) Scene Layer—Correlation Model: In general, three groups of GOP correlation models can be distinguished [5]: (i) Short-range dependence (SRD) models, such as the Markovian ones, defined by an ACF $\rho(k)$ that drops off exponentially, $\rho(k) \sim e^{-\beta k}$, (ii) Long-range dependence (LRD) models, also referred to as self-similar models and represented by an ACF that drops off slowly, $\rho(k) \sim k^{-\beta}$, and (iii) The $M/G/\infty$ model, with an ACF of the form $\rho(k) \sim e^{-\beta\sqrt{k}}$. The last model fits between the two others given that it drops slower than the SRD model and faster than the LRD one. Based on empirical video traces, some papers [6], [7] show that the GOP ACF of MPEG-4 encoded video traffic follows a self-similar model, while others, for instance [5], show that it follows the $M/G/\infty$ model.

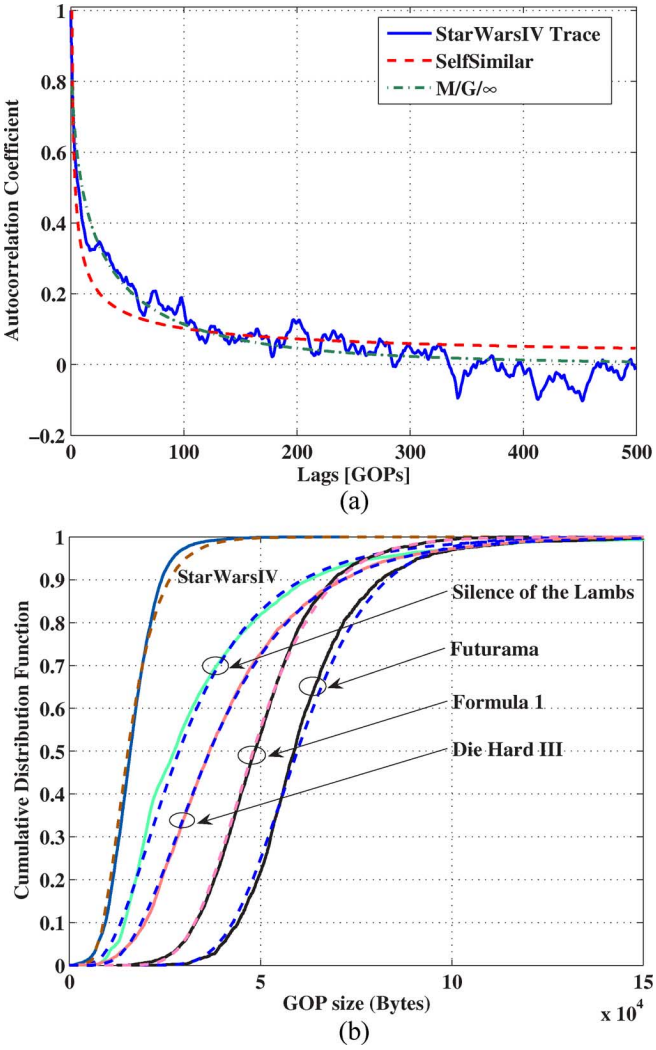

Fig. 6. Video trace modeling for the Scene and GOP levels: (a) Approximation for ACF of *StarWarsIV* GOPs by Self similar and $M/G/\infty$ processes; (b) GOP CDFs of empirical movie traces and their approximations through Lognormal distributions.

Before proceeding with our queuing buffer size modeling, an investigation of the appropriate ACF model to be used is needed. In this vein, we compare the fitting of the GOP ACF of 18 empirical video traces [14] to both models ($M/G/\infty$ and self-similar). Table III shows, at the scene level, the least square error (LSE) between the GOP ACF of the empirical data and that of the approximated GOP ACF, for different video sequences. In addition, taking the *StarWarsIV* movie as an example, we present in Fig. 6(a) the corresponding GOP ACF and the fitting to the considered models. These results show that the $M/G/\infty$ process is obviously more accurate than the self-similar model. Hence, the scene length duration will have a CDF of the form:

$$F_d(k) = 1 - \frac{e^{-\beta\sqrt{k}} - e^{-\beta\sqrt{k+1}}}{1 - e^{-\beta}}. \tag{15}$$

Furthermore, the differences in the ACF modeling noted in the above-cited papers can be explained by examining the *Futurama* GOP ACF statistics shown in Table III. In the latter, we provide statistics for two windows of observation (GOP

interval), the first between GOP 500 and GOP 2500, and the second between GOP 1000 and GOP 2500. The results show that, in the first case, the LSE pertaining to the $M/G/\infty$ model is smaller than that of the self-similar model. Hence, the $M/G/\infty$ model can be considered as a better fit in this case. The opposite is observed in the second case, which shows that even the interval size of empirical video data affects the ACF modeling. Moreover, we note that the larger the interval size is, the better is the modeling of the ACF through an $M/G/\infty$ process.

2) Gop Layer—Marginal Distribution: Pervious works studied the statistical behavior of frames I, B, and P, and suggested different distributions, such as the Gamma, lognormal, Beta and hybrid Gamma/Pareto [5]–[7]. As aforementioned, the intra-GOP correlation (fast time-scale) has little impact on the queuing modeling for a weak-stability scenario, i.e., the conditional mean of at least one scene is greater than the capacity of the server, which is generally the case in wireless systems [4]. Thus, the intra-GOP behavior (frame) is not modeled and the traffic unit is taken in terms of GOPs. Based on 18 empirical video traces, for some of which we provide in Fig. 6(b) fitting to the Lognormal CDF, a perfect match can be noted between the empirical curves and the fitting ones. Using the underlined distribution, the GOP size PDF with mean ψ and variance σ^2, can be expressed as

$$f(x) = \begin{cases} \frac{1}{V\sqrt{2\pi}x} \exp\left[\frac{-(\log x - M)^2}{2V^2}\right] & x > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{16}$$

with $V^2 = \log(1 + (\sigma^2/\psi^2))$ and $M = \log(\psi^2/\sqrt{\sigma^2 + \psi^2})$.

Designing and conducting such a set of experiments in justifying the utilization of the $M/G/\infty$ model, and the Lognormal distribution for the GOP size, is part of the contribution of this paper. Table III presents the parameters needed to model both layers (Scene and GOP) for different empirical video traces.

B. Queuing Buffer Size Modeling

Using the results of the previous section, we now present simulation results demonstrating the accuracy of our video streaming model, followed by results pertaining to the queuing buffer size distribution. Fig. 7(a) illustrates a comparison between the ACF of the real *StarWarsIV* trace and that of the simulated one using the proposed video model with the parameters of the real trace (Table III). As observed, both ACFs can be fitted by the same exponential function, $e^{-\beta\sqrt{k}}$ with the same parameter β, which characterizes the ACF of an $M/G/\infty$ process, hence demonstrating the accuracy of our model in reproducing the same correlation behavior as a real trace. The queuing buffer size for video traffic is more complicated to model than the queuing delay of the web-browsing traffic. Indeed, video traffic involves many independent parameters (e.g., ACF parameters, GOP size parameters, scene duration parameters, etc.) which makes establishing a general relationship between the parameters involved difficult to achieve. However, based on extensive simulations, several observations can be made with respect to the variation of the mean buffer

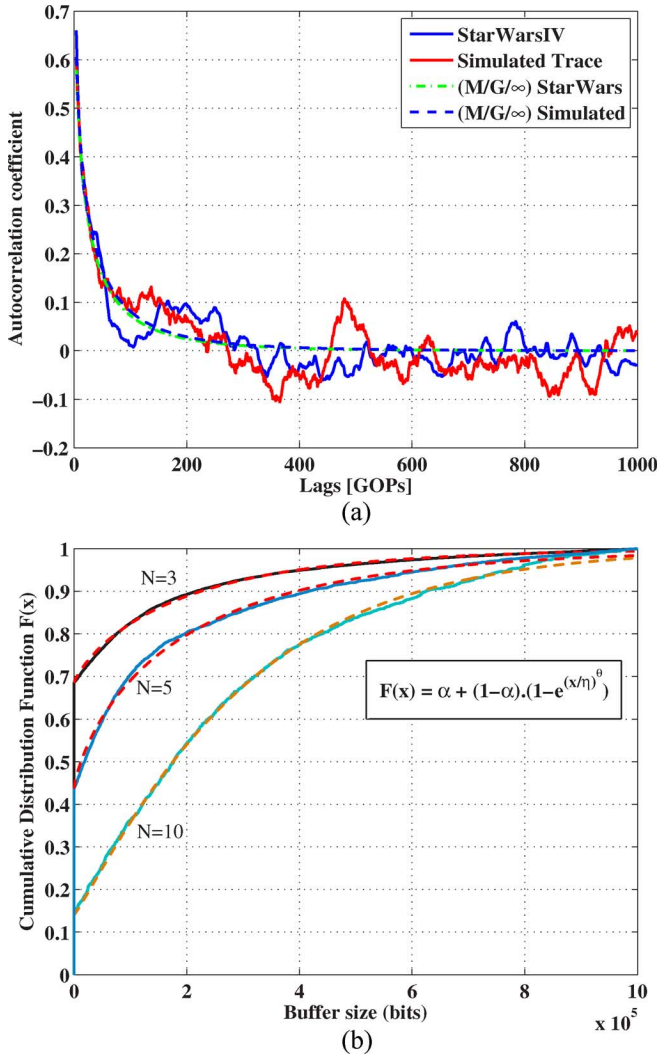

Fig. 7. Validation of the proposed video traffic model considering *StarWarsIV*: (a) ACF of the empirical GOPs and the ACF of the traffic generated by the proposed model, and their respective $M/G/\infty$ approximations; (b) CDF of the queuing buffer size at the Node-B, and its approximation through *weighted Weibull* function for different numbers of UEs, N.

size χ and that of the probability α of having an empty buffer, as a function of the ACF coefficient β. In particular, we note that the probability α is independent of β, while χ decreases linearly as the value of β increases. Thus, we can conclude that the probability to get an empty buffer is independent of the correlation of the traffic (scene layer), and is more affected by the traffic load (GOP layer). Furthermore, for a long-term correlated video traffic ($\beta \to 0$), the GOP ACF $\rho(k) \to 1$ and the mean buffer size χ becomes very large, which can lead to buffer overflow.

Following these observations, we now present the main important result, namely, that the queuing buffer size of video streaming traffic follows a weighted Weibull distribution. Considering the *StarWars IV* trace, and the channel variations of the different users to follow a Rayleigh distribution with an average SNR of 16 dB, thus yielding a total throughput equal to 3.58

Mbps [9], we show in Fig. 7(b) the CDF of the queuing buffer size for different values of N. The CDF $F(x)$ is given by

$$F(x) = \alpha + (1-\alpha)\left(1 - e^{\left(\frac{x}{\eta}\right)^{\theta}}\right), \tag{17}$$

where α is the probability of having an empty buffer, and parameters η and θ define the shape of the Weibull distribution with mean χ and variance ϕ^2 given by

$$\begin{aligned} \chi &= \eta\Gamma(1+\theta^{-1}), \\ \phi^2 &= \eta^2\left[\Gamma(1+2\theta^{-1}) - \Gamma^2(1+\theta^{-1})\right], \end{aligned} \tag{18}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

The proposed queuing buffer size distribution model, which is valid for any video traffic that can be well fitted by an $M/G/\infty$ process, servers as a practical tool for the development and evaluation of scheduling protocols that allow for an efficient sharing of the resources between streaming-service users. Indeed, without the need to resort to the complicated video traffic generating, the proposed queuing buffer size distribution model can be used to directly simulate the buffer size and hence consider a wide range of traffic scenarios, which does not only allow for a significant complexity reduction in the performance analysis of existing scheduling policies, but also for the development of new algorithms that dynamically adapt to the operating condition so as to achieve the goal of providing the required QoS for the users.

V. Conclusion

This paper proposed models for dimensioning the web-browsing and video streaming services in HSDPA networks. The web-browsing traffic was modeled through the layered structure: session, burst and packet levels. We showed that the packet-call queuing delay is exponentially distributed and presented a mathematical formulation for the parameters that characterize the corresponding PDF. By means of this formulation or directly from our curves, for a given reading time value and a required mean queuing delay, the maximum number of allowed UEs can directly be deduced. In addition, the proposed expressions can be used jointly with any fair scheduling algorithm which provides the same air throughput for all UEs. Besides, we provided a simple video-streaming traffic model over three layers: movie, scene and GOP layers, and showed that the GOP autocorrelation function of the video traffic follows an $M/G/\infty$ process, and that the GOP size follows a Lognormal distribution. Furthermore, we showed that the queuing buffer size follows a weighted Weibull distribution for all video traffics that can be modeled by $M/G/\infty$ processes. The proposed model can be directly used to simulate the video streaming buffer behavior, which significantly reduces the complexity of the development and evaluation of scheduling and call admission policies for the video streaming service.

Acknowledgment

The authors would like to thank the Associate Editor and the anonymous reviewers for helpful suggestions and comments.

References

[1] *3rd Generation Partnership Project; Technical Specification Group Services and System Aspects Service aspects; Services and Service Capabilities (Release 6)*, 3GPP Std. 22.105, Rev. 6.3.0, Mar. 2005.

[2] *Selection procedures for the choice of radio transmission technologies of the UMTS*, ETSI Std. 101.112, Rev. 3.2.0, Apr. 1998.

[3] K. Butterworth, M. Shafi, and P. J. Smith, "A flexible model for dimensioning mixed service 3{G} wireless networks," in *Proc. IEEE International Conference on Communications (ICC'05)*, Seoul, Korea, May 2005, vol. 4, pp. 2207–2212.

[4] P. Jelenkovic, A. Lazar, and N. Semret, "The effect of multiple time scales and subexponentiality in {MPEG} video streams on queueing behavior," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1052–1071, Aug. 1997.

[5] M. Krunz and A. Makowski, "Modeling video traffic using M/G/ ∞ input processes: a compromise between markovian and LRD models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 733–748, June 1998.

[6] N. Ansari, L. Hai, Y. Shi, and H. Zhao, "On modeling {MPEG} video traffics," *IEEE Trans. Broadcast.*, vol. 48, pp. 337–347, Dec. 2002.

[7] M. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar {VBR} video traffic," in *SIGCOMM '94: Proc. of the Conference on Communications Architectures, Protocols and Applications*, London, UK, 1994, pp. 269–280.

[8] G. Aniba and S. Aïssa, "A general traffic and queueing delay model for 3{G} wireless packet networks," in *Proc. 11th International Conference on Telecommunications (ICT'04)*, Fortaleza, Brazil, Aug. 2004, vol. 3, pp. 942–949.

[9] *3rd Generation Partnership Project; Technical Specification Group Radio Access Networks; Physical Layer Procedures (FDD) (Release 6)*, 3GPP Std. 25.214, Rev. 6.3.0, 2004.

[10] C. J. Ong, P. H. J. Chong, and R. Kwan, "Effects of various packet scheduling algorithms on the performance of high speed downlink shared channel in a WCDMA network," in *Proc. IEEE Pacific Rim Conference on Communications, Computers and signal Processing (PACRIM'03)*, Aug. 28–30, 2003, vol. 2, pp. 935–938.

[11] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE Vehicular Technology Conference (VTC-S'00)*, Tokyo, Japan, May 2000, vol. 3, pp. 1854–1858.

[12] O. Shin and K. Bok, "Antenna-assisted Round Robin scheduling for MIMO cellular systems," *IEEE Commun. Lett.*, vol. 7, pp. 109–111, May 2003.

[13] G. Aniba and S. Aïssa, "Adaptive proportional fairness for packet scheduling in HSDPA," in *Proc. IEEE Global Telecommunications Conference (Globecom'04)*, Dallas, TX, Nov. 2004, vol. 6, pp. 4033–4037.

[14] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H.263 Video Traces for Network Performance Evaluation," 2001 [Online]. Available: http://www-tkn.ee.tu-berlin.de/research/trace/trace.html

[15] T. Taralp, M. Devetsikiotis, and I. Lambadaris, "Efficient fractional Gaussian noise generation using the spatial renewal process," in *Proc. IEEE International Conference on Communications (ICC'98)*, Jun. 1998, vol. 3, pp. 1456–1460.

Sonia Aïssa (S'93–M'00–SM'03) received her Ph.D. degree in Electrical and Computer Engineering from McGill University, Canada, in 1998. She is now Associate Professor at INRS-EMT, University of Quebec, Montreal, Canada, and Adjunct Professor at Concordia University, Montreal, Canada.

From 1996 to 1997, she was a visiting researcher at the department of electronics and communications of Kyoto University, Kyoto, Japan, and at the wireless systems laboratories of NTT, Kanagawa, Japan. From 1998 to 2000, she was a research associate at INRS-Telecommunications, Montreal, Canada. From 2000 to 2002, she was a principal investigator in the major program of personal and mobile communications of the Canadian Institute for Telecommunications Research, conducting research in CDMA systems. In 2006, she was Invited Professor at the Graduate School of Informatics, Kyoto University, Kyoto, Japan. Her research interest is in wireless communications and networking, and includes radio resource management, cross-layer design and MIMO systems.

Dr. Aïssa is a recipient of the Quebec government FQRNT fellowship "Strategic Program for Professors-Researchers", received the Performance Award in 2004 from INRS-EMT for outstanding achievements in research, teaching and service, and the Community Service Award in 2007 from the FQRNT Center for Advanced Systems and Technologies in Communications (SYTACom). She serves as Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and Associate Editor for the IEEE Communications Magazine and the IEEE Wireless Communications Magazine. As Guest Editor, she served for the 2006 EURASIP journal on Wireless Communications and Networking: Special Issue on Radio Resource Management in 3G+ Systems. She is the founding chair of the Montreal Chapter IEEE Women In Engineering society, served as Technical Program Chair for the Wireless Communications Symposium of IEEE ICC'2006, and acted as PHY/MAC Program Chair for the IEEE WCNC'2007.

Ghassane Aniba (S'04) received the Dipl.-Ing. degree in telecommunication engineering from the Institut National des Postes et Telecommunications (INPT), Rabat, Morocco, in 2002. He is currently working toward the Ph.D. degree at the Institut National de la Recherche Scientifique, Energie, Matériaux et Télécommunications (INRS-EMT), University of Quebec, Montreal, Canada.

For his Ing. degree, he worked on the adaptation of multicast protocols over satellite links (IP over DVB), with the Planéte Research Team at the Institut National de Recherche en Informatique et Automatique (INRIA), Sophia Antipolis, France. His current research interests include traffic modeling in wireless networks, MIMO systems in HSDPA, and QoS in multiuser wireless networks.