A Flow-Level Performance Model for Mobile Networks Carrying Adaptive Streaming Traffic

Thomas Bonald, Salah-Eddine Elayoubi, Yu-Ting Lin

A Flow-Level Performance Model for Mobile Networks Carrying Adaptive Streaming Traffic

Thomas Bonald
Telecom ParisTech
Paris, France
thomas.bonald@telecom-paristech.fr

Salah Eddine Elayoubi
Orange Labs
Issy les Moulineaux, France
salaheddine.elayoubi@orange.com

Yu-Ting Lin
Orange Labs and Telecom ParisTech
France
yuting.lin@orange.com

***Abstract*—This paper proposes a performance model for mobile networks carrying adaptive streaming traffic. The proposed model takes into account the flow dynamics in addition to the main parameters influencing the performance of adaptive streaming, such as the playout buffer and the video bit rates. We show how to compute several performance metrics like the average video bit rate, the deficit rate, defined as the probability of having an instantaneous throughput lower than the chosen video bit rate, and the average buffer surplus, related to the amount of data accumulated in the buffer. Considering the coexistence of multiple services and heterogeneous radio conditions that make the exact solution intractable, we propose a simple yet accurate approximation that is easy to integrate in the operator's dimensioning tools. Our numerical results investigate the performance trade-offs between the different parameters of the adaptive streaming service.**

I. Introduction

According to Cisco forecast 2013 [1], video services will account for increasing proportion of data transmitted in future mobile networks and adaptive streaming will represent most of the video streaming services. Several implementations of adaptive streaming exist, such as Microsoft Smooth Streaming, Adobe HTTP Dynamic Streaming, Apple HTTP Live Streaming (HLS) and the MPEG Forum specified a standard Dynamic Adaptive Streaming over HTTP (MPEG-DASH). In this paper, we do not look on a specific protocol implementation but rather develop a general performance model that aims at helping operators assessing the quality of service perceived by their users and properly dimensioning their networks.

In [2], authors show typical implementations for adaptive streaming, where the video files are divided into several chunks (segments) [3][4] and the server has several encoded versions of each chunk. At the client side, a playout buffer stores the downloaded video segments that are not yet played. Even if this buffer is of finite size, we suppose in this paper that it is able to store the whole video files and that the video chunks are downloaded by the user at the maximum achievable throughput so that the amount of buffered chunks is maximized. We also consider small chunk sizes, ensuring an instantaneous adaptation of the video bit rate to the available throughput, when the number of active flows changes in the system. A traffic flow stands for a download task such as data transfer or HTTP video streaming.

Users are satisfied if they watch the video at a high coded rate and if the video playout is smooth, i.e. the playout buffer never gets empty. The Key Performance Indicators (KPIs) that directly influence the users' Quality of Experience (QoE) are the average video bit rate observed during the video session and the starvation probability, i.e. the probability that the buffer gets empty. While our flow-level model is able to compute the average video bit rate, the starvation probability computation needs a packet-level analysis as it depends on the behavior of the player in terms of prefetching policy and the detailed buffer state. As our objective is to provide simple models that can be used for mobile network dimensioning purposes, we provide two KPIs that are related to the starvation probability and can be computed using the flow-level model: the deficit rate defined as the probability that the instantaneous user throughput is lower than the chosen video bit rate, and the buffer surplus representing the average buffer variation during the video download.

Related literature

An important set of works considered flow-level modeling in mobile networks. Note that a flow represents for instance a file download or a video streaming session and that it may be subject to different radio conditions due to the position of the user in the cell and to throughput variations due to the dynamics of arrivals and departures of other users. In [5], a flow-level model has been proposed for elastic traffic in mobile networks assuming static users. This model is extended to mobile users in [6] and to advanced radio features in [7]. Another set of works considers the performance of real-time streaming services at the flow level [8] and combining flow and packet levels [9]. The integration of elastic and real-time streaming services is considered in [10] that provides performance bounds for both services. This integration has been also considered in [11] where the streaming performance is evaluated through the blocking rate.

The above mentioned flow-level models focus on classical elastic and real-time streaming services and do not consider HTTP video streaming with buffering. A flow-level model is proposed in [12] for HTTP video streaming with infinite buffers and KPIs like starvation probability are computed using a detailed buffer analysis. However, this work does not consider the video bit rate adaptability and the integration of elastic traffic.

Main contribution

We develop a flow-level model for the performance of adaptive streaming taking into account the main characteristics of this service like the presence of a buffer and the video bit rate limitations. We extend this model to consider heterogeneous radio conditions and a mix between streaming and elastic traffic. We then propose and validate an approximation model to facilitate the calculation of performance metrics. Finally, we demonstrate the performance tradeoffs of different maximum video bit rates.

Paper organization

The remainder of the paper is organized as follows. Section II describes the system and develops the flow level performance model for adaptive streaming traffic. In section III, we extends the model with the consideration of heterogeneous radio conditions and the integration of elastic and streaming services. The approximation model is introduced in section IV. In section V, numerical analysis shows the validation of approximation model and the impacts of different video bit rates. Section VI concludes the paper.

II. Model for adaptive streaming

A. System description

In this section, we consider a cellular network and focus on the performance of a typical cell as described in Fig. 1. We will begin, for the ease of understanding of the models, by a homogeneous scenario where all users have the same physical rate R corresponding to the average radio condition over the cell. Section III will show how to extend this model to multiple radio conditions and how to include elastic traffic.

Moreover, to facilitate the flow-level modeling, for the streaming system configuration, we assume that each flow has an infinite buffer size so that each leaves the system as soon as its corresponding video is fully downloaded. During its stay in the system, each flow continuously downloads the video chunks according to the allocated resource. The streaming chunk size is assumed to be significantly small, enabling users to adapt their video bit rates at any time.

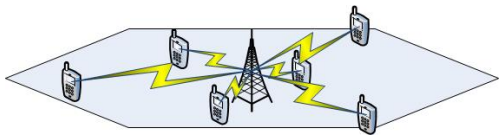


Fig. 1: System Model for Mobile Wireless Networks

B. Markov model

We make the classical assumption that traffic flows arrive according to a Poisson process with intensity λ. We only consider adaptive streaming services in this section, the mix with other services is considered later. The video durations are assumed to be independent and exponentially distributed with mean T.

Let $X(t)$ be the number of flows at time t. When $X(t) = x$, which means that there are x flows served in the system, the flow departure rate, $\mu(x)$, can be expressed as

$$\mu(x) = \frac{\phi(x)}{\sigma(x)}, \tag{1}$$

where $\phi(x)$ stands for the physical throughput allocated to all UEs of the cell at state x and $\sigma(x)$ stands for the remaining flow size at state x. As we consider an infinite buffer size at the user side, users can fully utilize the throughput allocated to them, occupying thus the whole throughput, $\phi(x) = R$.

As for the flow size, as we consider a small chunk size and an instantaneous adaptation to the radio conditions, the video bit rate depends on state x only and not on the history of the system. Furthermore, as the video duration is assumed exponential, the memoryless property applies and the flow size at state x is also exponentially distributed with mean

$$\sigma(x) = v(x)T, \tag{2}$$

where $v(x)$ is the video bit rate at state x. $X(t)$ is thus a Markov process whose transition rates depend on the selection of video bit rate. We will show in the following sections how the selected video bit rate is computed and how to derive the performance model in different scenarios of video bit rate limitations.

C. No video bit rate limitation

We first assume that number of available video bit rates is not limited so that the bit rate assigned to each flow is exactly equal to the instantaneous throughput, $\gamma(x)$, that the user gets. This corresponds to an extremely adaptive streaming service where very low and very high video bit rates are allowed. Then, we have

$$v(x) = \gamma(x) = \frac{R}{x}. \tag{3}$$

With equation (3), the flow departure rate then can be expressed as

$$\mu(x) = \frac{\phi(x)}{v(x)T} = \frac{R}{T \times R/x} = \frac{x}{T}. \tag{4}$$

The system can be easily shown to have a product form and the stationary distribution is

$$\pi(x) = \pi(0)\frac{\rho^x}{x!}, \tag{5}$$

where $\pi(0) = e^{-\rho}$ and $\rho = \lambda T$.

The number of users in the system evolves here like real-time communications as flows are always admitted in the system and the video duration is independent of the system state. Indeed, the buffer never fills as the user always watches the video with a rate equal to its available throughput, and starvation never happens as the video bit rate can go to zero without a playout interruption.

D. Video bit rate limitation

We now consider a more realistic case where minimal $v_{\min}$ and maximal $v_{\max}$ video bit rates are setup. However, we still assume a continuous set of video bit rates between these limits, although our model is sufficiently general to consider a discrete set. Therefore, the bit rate can be expressed as

$$v(x) = \max\Big(\min\Big(\gamma(x), v_{\max}\Big), v_{\min}\Big). \tag{6}$$

The flow departure rate becomes

$$\mu(x) = \frac{\phi(x)}{v(x)T} = \frac{R}{\max\Big(\min\Big(\gamma(x), v_{\max}\Big), v_{\min}\Big)T}. \tag{7}$$

When the available throughput $\gamma(x) = \frac{R}{x}$ is larger than $v_{\max}$, the buffer starts filling as the download capacity is larger than the playout rate. The flow behaves thus as an elastic one as there is no limitation on the buffer size. On the other hand, when this available throughput is smaller than $v_{\min}$, the behavior is also elastic but with possible starvation if the buffer empties due to the constant playout rate. Between $v_{\max}$ and $v_{\min}$, the flow behaves as a real time one and the buffer occupation remains constant. With the reversibility property [13], the stationary distribution $\pi(x)$ is

$$\pi(n) = \frac{\lambda^n}{\prod_{x=1}^{n}\mu(x)}\pi(0), \tag{8}$$

where $\pi(0)$ can be obtained by calculating $\sum_{x=0}^{\infty}\pi(x) = 1$.

E. Stability condition

If $v_{\min} = 0$, the system is always stable as the average flow duration is always equal to T. When $v_{\min} > 0$, the system may become unstable for a large number of arrivals. The flow arrival rate should be smaller than the max flow departure rate, leading to the following stability condition:

$$\lambda < \frac{R}{v_{\min}T}.$$

F. KPIs definition

To evaluate the performance of adaptive streaming service, we propose three key performance indicators, mean video bit rate, deficit rate and buffer surplus. All of them are defined based on the stationary distribution $\pi(x)$.

1) Mean video bit rate: The mean video bit rate stands for the average bit rate that a user experiences while watching the video. When the mean video bit rate is high, user has a better video experience. Due to the observers' paradox [13], users see a biased distribution of the state, proportional to $x\pi(x)$ in state x. Thus, we define the overall mean video bit rate as

$$\bar{v} = \sum_{x:x>0}\frac{x\pi(x)}{\bar{x}}v(x), \tag{9}$$

with $\bar{x} = \sum_{x>0} x\pi(x)$.

2) Deficit rate: A popular QoE indicator used to evaluate streaming performance is the starvation probability [12]. Even if starvation happens only when the video bit rate is larger than the instantaneous throughput, the latter condition is not a sufficient condition for starvation as the buffer may counteract the impact of short periods of low throughput. The computation of the starvation probability has to take into account the memory of the system by introducing the buffer size in the Markovian analysis, as in [12].

Here, we introduce and examine the performance of a metric called deficit rate. The deficit rate is equal to the probability an ongoing flow sees its instantaneous throughput lower than its chosen bit rate. As the rate adaptation is assumed to occur instantaneously in reaction to the variations of the observed throughput, the deficit rate is defined by the probability that the instantaneous throughput, $\gamma(x) = \frac{\phi(x)}{x}$, is smaller than $v_{\min}$. The overall deficit rate is defined by weighting the stationary distribution at different state x with the number of flows:

$$D = \mathrm{P}\{\gamma(x) < v_{\min}\} = \sum_{x:x>0}\frac{x\pi(x)}{\bar{x}}\mathbb{1}_{\{\gamma(x)<v_{\min}\}}, \tag{10}$$

where $\mathbb{1}$ stands for the indicator function.

3) Buffer surplus: We also introduce another performance metric called buffer surplus, which represents the average relative buffer variation of each flow. It is expressed as

$$B = \sum_{x:x>0}\frac{x\pi(x)}{\bar{x}}\Big(\frac{\gamma(x) - v(x)}{v(x)}\Big). \tag{11}$$

When $\gamma(x) > v(x)$, the video contents accumulates in the buffer. When $\gamma(x) < v(x)$, then user starts to consume the video packets stored in the buffer. When $B > 0$, the video buffer of each flow increases on average. We shall see that $B = 0$ is a good objective for dimensioning the maximum traffic load.

III. Model extension

Section II proposed performance metrics for adaptive streaming and showed how to compute them when only adaptive streaming flows share the capacity. Here, we generalize the model to heterogeneous radio conditions and discuss the performance of these flows in the presence of elastic traffic. We also show how to integrate opportunistic scheduling in the model.

A. Heterogeneneous radio conditions and mixed service

Based on the 3GPP LTE-A standards [14], users with various positions have different discrete CQI (Channel Quality Indicator), for example from CQI-1 to CQI-15. Therefore, traffic flows can be separated into several classes, $i \in I = \{1, \cdots\}$. Each class has a radio condition R_i and accounts for a proportion of flow arrivals, p_i, with $\sum_{i\in I} p_i = 1$.

We also consider an extended setting where streaming flows share the cell capacity with elastic flows. Therefore, the system can be represented as a network of two groups of coupled processor-sharing queues. Flows in queue e, i correspond to the elastic traffic with radio condition, R_i, and flows in queue

s, i correspond to the streaming ones with R_i. The flow arrival rates at each queue are calculated as $\lambda_{e,i} = q_e p_i \lambda$, $\lambda_{s,i} = q_s p_i \lambda$, where q_e and q_s stand for the proportions of elastic and streaming arrival rates and $q_e + q_s = 1$.

Applying the same concept to formulate the departure rate as equation (4), the flow departure rate of elastic and adaptive streaming services can be expressed as

$$\mu_{e,i}(\boldsymbol{x}) = \frac{\phi_{e,i}(\boldsymbol{x})}{\sigma}, \quad \mu_{s,i}(\boldsymbol{x}) = \frac{\phi_{s,i}(\boldsymbol{x})}{v_i(\boldsymbol{x})T}, \tag{12}$$

where $\boldsymbol{x} = (x_{e,1}, \cdots, x_{e,i}, x_{s,1}, \cdots, x_{s,i})$ represents the number of flows for each service and σ is the mean flow size of elastic data. $\phi_{e,i}(\boldsymbol{x})$ and $\phi_{s,i}(\boldsymbol{x})$ stand for the allocated wireless resources for each class. Assuming the Round-Robin scheduling, these capacity shares are

$$\phi_{e,i}(\boldsymbol{x}) = \frac{x_{e,i} R_i}{|\boldsymbol{x}|}, \ \phi_{s,i}(\boldsymbol{x}) = \frac{x_{s,i} R_i}{|\boldsymbol{x}|}, \tag{13}$$

where $|\boldsymbol{x}| = \sum_i (x_{e,i} + x_{s,i})$. Here, the video bit rate, $v_i(\boldsymbol{x})$, is chosen depending on the instantaneous throughput as follows,

$$v_i(\boldsymbol{x}) = \max\Big(\min\Big(\gamma_{s,i}(\boldsymbol{x}), v_{\max}\Big), v_{\min}\Big). \tag{14}$$

where $\gamma_{s,i}(\boldsymbol{x}) = \frac{\phi_{s,i}(\boldsymbol{x})}{x_{s,i}} = \frac{R_i}{|\boldsymbol{x}|}$. With the formulation above, note that the balanced property in [10] is not valid as $\mu_{s,i}(\boldsymbol{x} - \mathbf{e}_{e,j})\mu_{e,j}(\boldsymbol{x}) \neq \mu_{e,j}(\boldsymbol{x} - \mathbf{e}_{s,i})\mu_i(\boldsymbol{x})$, where $i, j \in I$ and $\mathbf{e}_i$ is a vector with only 1 user belonging to class i. The Markov chain is not reversible and we have to solve the balance equations (by truncating the state space and inversing the corresponding matrix) for the stationary distribution $\pi(\boldsymbol{x})$:

$$\begin{aligned} &\forall \boldsymbol{x}, \Big(\sum_{j=e,s}\sum_i \lambda_{j,i} + \mu_{j,i}(\boldsymbol{x})\Big)\pi(\boldsymbol{x}) \qquad (15)\\ &= \sum_{j=e,s}\sum_i \Big(\lambda_{j,i}\pi(\boldsymbol{x} - \boldsymbol{e}_{j,i}) + \mu_{j,i}(\boldsymbol{x} + \boldsymbol{e}_{j,i})\pi(\boldsymbol{x} + \boldsymbol{e}_{j,i})\Big), \end{aligned}$$

B. Stability condition

We begin by assessing the stability region of the system. When the system approaches the stability limit, the video bit rate of streaming users decreases to $v_{\min}$. The system behaves like a system of elastic traffic. Stability holds only when the sum of offered loads for both services is less than 1, that is:

$$\sum_i \Big(\frac{\lambda q_e p_i \sigma}{R_i} + \frac{\lambda q_s p_i v_{\min} T}{R_i}\Big) \leq 1.$$

We deduce the maximum flow arrival rate:

$$\lambda_{\max} = \Big(\sum_i \frac{R_i}{q_e p_i \sigma + q_s p_i v_{\min} T}\Big)^{-1}. \tag{16}$$

C. KPIs definition

We extend the adaptive streaming related KPIs to this general case and define a KPI for the performance of elastic traffic.

1) Mean video bit rate: We define the overall mean video bit rate of general model, $\bar{v}$, by summing up all the classes and weighting by the flow number of each class at state $\boldsymbol{x}$ and the mean video bit rate of class i, $\bar{v}_i$ as

$$\bar{v} = \sum_{\boldsymbol{x}:|\boldsymbol{x}_s|>0} \frac{\pi(\boldsymbol{x})}{\bar{\boldsymbol{x}}_s} \sum_i x_{s,i} v_i(\boldsymbol{x}), \tag{17}$$

$$\bar{v}_i = \sum_{\boldsymbol{x}:x_{s,i}>0} \frac{\pi(\boldsymbol{x})}{\bar{x}_{s,i}} x_{s,i} v_i(\boldsymbol{x}), \tag{18}$$

where $\bar{\boldsymbol{x}}_s = \sum_{\boldsymbol{x}} \pi(\boldsymbol{x})|\boldsymbol{x}_s|$, $\bar{x}_{s,i} = \sum_{\boldsymbol{x}} x_{s,i}\pi(\boldsymbol{x})$ and $|\boldsymbol{x}_s| = \sum_i x_{s,i}$.

2) Deficit rate: By using the same concept, we define the overall deficit rate of multiple class model, D and the deficit rate of class i, D_i as

$$D = \sum_{\boldsymbol{x}:|\boldsymbol{x}_s|>0} \frac{\pi(\boldsymbol{x})}{\bar{\boldsymbol{x}}_s} \sum_i x_{s,i} \mathbb{1}_{\{\gamma_{s,i}(\boldsymbol{x}) < v_{\min}\}}, \tag{19}$$

$$D_i = \sum_{\boldsymbol{x}:x_{s,i}>0} \frac{\pi(\boldsymbol{x})}{\bar{x}_{s,i}} x_{s,i} \mathbb{1}_{\{\gamma_{s,i}(\boldsymbol{x}) < v_{\min}\}}, \tag{20}$$

where $\mathbb{1}$ is the indicator function equal to 1 when the condition is satisfied, otherwise the indicator function will become 0.

3) Buffer surplus: Same as the previous concept, we define the overall buffer surplus of multiple class model, B and the buffer surplus of class i, B_i as

$$B = \sum_{\boldsymbol{x}:|\boldsymbol{x}_s|>0} \frac{\pi(\boldsymbol{x})}{\bar{\boldsymbol{x}}_s} \sum_i x_{s,i} \Big(\frac{\gamma_{s,i}(\boldsymbol{x}) - v_i(\boldsymbol{x})}{v_i(\boldsymbol{x})}\Big), \tag{21}$$

$$B_i = \sum_{\boldsymbol{x}:x_{s,i}>0} \frac{\pi(\boldsymbol{x})}{\bar{x}_{s,i}} x_{s,i} \Big(\frac{\gamma_{s,i}(\boldsymbol{x}) - v_i(\boldsymbol{x})}{v_i(\boldsymbol{x})}\Big). \tag{22}$$

4) Average elastic throughput: The average elastic throughput is chosen as the performance metric for elastic flows, with $\bar{\gamma}_e$ and $\bar{\gamma}_{e,i}$ standing for the overall metric and mean throughput for each class i,

$$\bar{\gamma}_e = \sum_{\boldsymbol{x}:|\boldsymbol{x}_e|>0} \frac{\pi(\boldsymbol{x})}{\bar{\boldsymbol{x}}_e} \sum_i x_{e,i} \frac{\phi_{e,i}(\boldsymbol{x})}{x_{e,i}}, \tag{23}$$

$$\bar{\gamma}_{e,i} = \sum_{\boldsymbol{x}:x_{e,i}>0} \frac{\pi(\boldsymbol{x})}{\bar{\boldsymbol{x}}_e} \frac{x_{e,i}\phi_{e,i}(\boldsymbol{x})}{x_{e,i}}, \tag{24}$$

where $\bar{\boldsymbol{x}}_e = \sum_{\boldsymbol{x}} \pi(\boldsymbol{x})|\boldsymbol{x}_e|$, $\bar{x}_{e,i} = \sum_{\boldsymbol{x}} x_{e,i}\pi(\boldsymbol{x})$ and $|\boldsymbol{x}_e| = \sum_i x_{e,i}$.

D. Impact of opportunistic scheduling

The capacity shares in equation (13) are computed assuming a round robin scheduling. Now consider a channel aware scheduling, like the proportional fair scheduler, that operates at the fast fading time scale [15]. In this case, when there are $|\boldsymbol{x}|$ flows that are active in the LTE cell, the throughput of a user of radio condition i that gets a proportion Φ of the cell resources is equal to $\Phi R_i G(|\boldsymbol{x}|)$, where $G(|\boldsymbol{x}|)$ is the opportunistic scheduling gain that depends on many parameters such as the channel model, the receiver and the Multiple Input Multiple Output (MIMO) scheme [16]. Note that, contrary to real time

streaming that does not benefit from the opportunistic scheduling gain due to its stringent delay constraints, http streaming with buffering takes advantage of this type of scheduling like elastic traffic. The performance model proposed above can thus be extended to the opportunistic scheduling case by introducing the fast fading gain, $G(|\boldsymbol{x}|)$ into the formulation of allocated resource in Eq. (13) and video bit rate in Eq.(14). Note that, $G(|\boldsymbol{x}|) = 1$ when round-robin scheduling is applied.

IV. Approximation

Section III introduces the general model with heterogeneous radio conditions and elastic flows. However, the complexity of solving numerically the balance equations mentioned in Eq. (15) increases exponentially with the number of classes, making the resolution of the stationary distribution very difficult. In this section, we propose an approximation model simplifying multiple queues into two coupled processor-sharing queues for elastic and streaming flows respectively.

A. Markov model

In our approximation, we make use of the intuition indicating that the contribution of a radio class to the load of the system is proportional to the ratio of the data volume it generates to the throughput it sees:

- For elastic traffic, class-i users generate a traffic volume of $\lambda_e p_i$ and observe a throughput of R_i, their contribution to the load of the elastic traffic queue, $\frac{\lambda_e p_i}{R_i}$, then can be used to calculate the equivalent throughput R_e.
- For the adaptive streaming traffic, class-i users generate a volume proportional to the video bit rate, i.e. that depends on the state of the system, $|\boldsymbol{x}| = x_s + x_e$. The load it generates in the streaming queue is proportional to $\frac{\lambda_s p_i v_i(\boldsymbol{x})}{R_i}$, where we use to calculate the equivalent throughput, $R_s(\boldsymbol{x})$.

As we mentioned, system is approximated as the one carrying only two equivalent classes of adaptive streaming and elastic traffic and their flow departures are calculated as

$$\hat{\mu}_e(\boldsymbol{x}) = \frac{\hat{\phi}_e(\boldsymbol{x})}{\sigma}, \quad \hat{\mu}_s(\boldsymbol{x}) = \frac{\hat{\phi}_s(\boldsymbol{x})}{v_{\min} T}, \tag{25}$$

where $\boldsymbol{x} = (x_e, x_s)$ denotes the number of elastic and streaming flows and the allocated resources for elastic and streaming are expressed as $\hat{\phi}_e(\boldsymbol{x}) = \frac{R_e}{x_e + x_s}$, $\hat{\phi}_s(\boldsymbol{x}) = \frac{R_s(\boldsymbol{x})}{x_e + x_s}$ with equivalent physical throughput

$$R_e = \Big(\sum_i \frac{p_i}{R_i} \Big)^{-1}, R_s(\boldsymbol{x}) = \Big(\sum_i \frac{p_i \alpha_i(\boldsymbol{x})}{R_i} \Big)^{-1}. \tag{26}$$

Moreover, $\alpha_i(\boldsymbol{x})$ is calculated by Eq. (14) as

$$\alpha_i(\boldsymbol{x}) = \frac{v_i(\boldsymbol{x})}{v_{\min}} = \frac{\max\Big(\min\Big(\gamma_{s,i}(\boldsymbol{x}), v_{\max} \Big), v_{\min} \Big)}{v_{\min}} \tag{27}$$

With the flow departure rate mentioned above and flow arrival rate shown as $\lambda_e = p_e \lambda$ and $\lambda_s = p_s \lambda$, the approximated stationary distribution of $\hat{\pi}(\boldsymbol{x})$ can be obtained using the same concept of balance equations shown in Eq. (15). However, the complexity is much lower. Note that the stability condition of this approximation is the same as that gotten in Eq.(16). Approximation also holds when applying opportunistic scheduling; it is thus sufficient to multiply the throughput by the state dependent scheduling gain $G(|\boldsymbol{x}|)$.

B. KPIs definition

As for the approximated performance metrics, the mean video bit rate, $\hat{v}_s$, the deficit rate, $\hat{D}_s$ and the buffer surplus, $\hat{B}_s$ for adaptive streaming traffic and the mean throughput for elastic traffic, $\hat{\gamma}_e$, can be computed with the newly calculated stationary distribution $\hat{\pi}(\boldsymbol{x})$ as

$$\hat{v}_s = \sum_{\boldsymbol{x}: x_s > 0} \frac{x_s \hat{\pi}(\boldsymbol{x})}{\bar{x}_s} \sum_i \frac{\beta_i(\boldsymbol{x})}{\bar{\beta}(\boldsymbol{x})} v_i(\boldsymbol{x}), \tag{28}$$

$$\hat{D}_s = \sum_{\boldsymbol{x}: x_s > 0} \frac{x_s \hat{\pi}(\boldsymbol{x})}{\bar{x}_s} \sum_i \frac{\beta_i(\boldsymbol{x})}{\bar{\beta}(\boldsymbol{x})} \mathbb{1}_{\{\gamma_{s,i} < v_{\min}\}}, \tag{29}$$

$$\hat{B}_s = \sum_{\boldsymbol{x}: x_s > 0} \frac{x_s \hat{\pi}(\boldsymbol{x})}{\bar{x}_s} \sum_i \frac{\beta_i(\boldsymbol{x})}{\bar{\beta}(\boldsymbol{x})} \Big(\frac{\gamma_{s,i}(\boldsymbol{x}) - v_i(\boldsymbol{x})}{v_i(\boldsymbol{x})} \Big), \tag{30}$$

$$\hat{\gamma}_e = \sum_{\boldsymbol{x}: x_e > 0} \frac{x_e \hat{\pi}(\boldsymbol{x})}{\bar{x}_e} \frac{\phi_e(\boldsymbol{x})}{x_e} = \sum_{\boldsymbol{x}} \frac{\hat{\pi}(\boldsymbol{x}) \phi_e(\boldsymbol{x})}{\bar{x}_e}, \tag{31}$$

where

$$\beta_i(\boldsymbol{x}) = \frac{p_i \alpha_i(\boldsymbol{x})}{R_i}, \ \bar{\beta}(\boldsymbol{x}) = \sum_i \beta_i(\boldsymbol{x}), \ \gamma_{s,i}(\boldsymbol{x}) = \frac{R_i}{|\boldsymbol{x}|}, \tag{32}$$

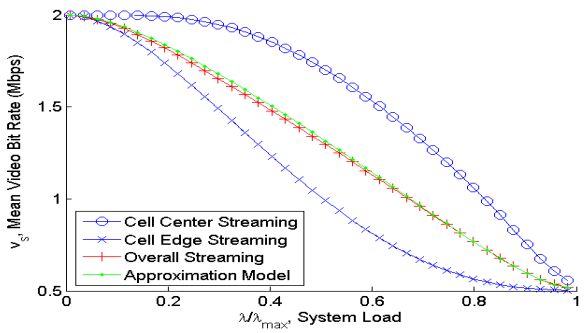
and $\frac{\beta_i(\boldsymbol{x})}{\bar{\beta}(\boldsymbol{x})}$ represents the fraction of load volume of class-i streaming users when there are $\boldsymbol{x}$ users in the system. In addition, $\bar{x}_s = \sum_{\boldsymbol{x}: x_s > 0} x_s \hat{\pi}(\boldsymbol{x})$ and $\bar{x}_e = \sum_{\boldsymbol{x}: x_e > 0} x_e \hat{\pi}(\boldsymbol{x})$ stand for the average number of streaming and elastic calls in the cell. It is also worth mentioning that the approximation model can also predict all the metrics for each class i.

V. Numerical Analysis

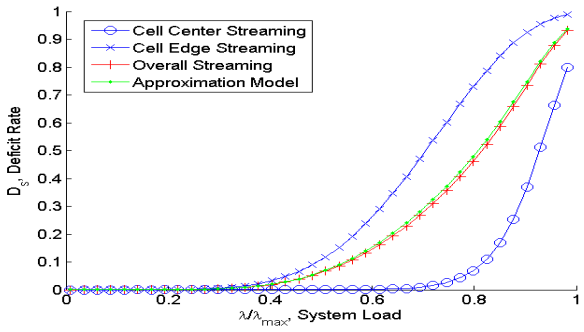
In our numerical analysis, we first validate the proposed approximation model, showing that the streaming and elastic performance with heterogeneous radio conditions can be accurately modeled by two classes. We then illustrate how to use our model for network dimensioning and study the impact of the video bit rates on the performance.

A. Validation of the approximation

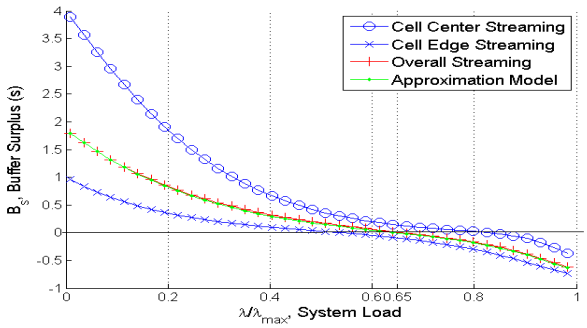
To make the exact model tractable, we consider two radio conditions only representing the cell center and the cell edge, respectively. Therefore, there are four classes for the general model and two for the approximate model. Then we examine the performance of the approximate model by setting the parameters as $(R_C, R_E) = (10, 4)$Mbps, $(p_C, p_E) = (\frac{1}{2}, \frac{1}{2})$ for cell center and cell edge, $(v_{\min}, v_{\max}) = (0.5, 2)$, $T = 10$s, $(p_e, p_s) = (\frac{1}{2}, \frac{1}{2})$, $\sigma = 5$Mbits. The simulation results in Fig.2 show the four defined metrics for both models when the offered traffic increases. These figures demonstrate that the predictions of the approximate model are very close to the ones obtained by the exact model for four different metrics, also for the metrics of each class.



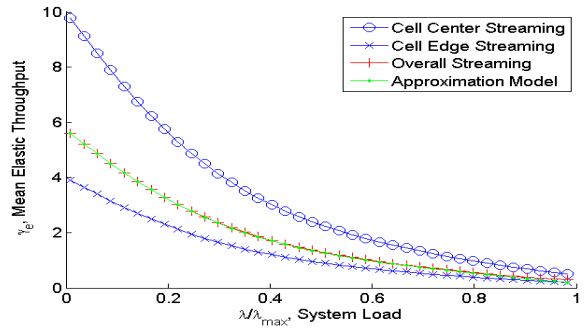
(a) Exact versus approximate model: mean video bit rate.



(b) Exact versus approximate model: deficit rate.



(c) Exact versus approximate model: buffer surplus.



(d) Exact versus approximate model: mean elastic throughput.

Fig. 2: Validation of approximation model

B. System dimensioning

Based on the results, operators can obtain the proper traffic intensity by setting some QoS constraints, which might be any combination of the performance metrics we defined. Take the buffer surplus, $B > 0$ as an example of QoS constraint in Fig.2c, the traffic load should be lower than $0.655\lambda_{\max}$. Otherwise, the average buffer surplus is smaller than zero, meaning the starvation events happen very often. Likewise, more QoS constraints can be considered together.

C. Performance trade-offs of $v_{\max}$ configuration

After having validated the approximation using a simple radio model, we use it in the following scenario with realistic radio conditions based on measurement data from a 4G network in a large European city, with an average cell radius of 350 meters. The concerned frequency band is LTE 1800 MHZ. Figure 3 shows the measured probability distribution function of the Channel Quality Indicator (CQI) obtained from base station measurements collected using an O&M tool. Each CQI is associate to an MCS, determining its spectral efficiency. Using the CQI-MCS association figures of [17] and considering a bandwidth of 10 MHZ, the corresponding harmonic capacity of an LTE cell is computed as equal to $R_e = 16.82$ Mbps.

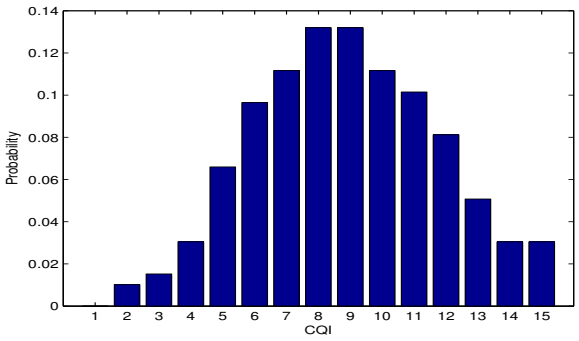


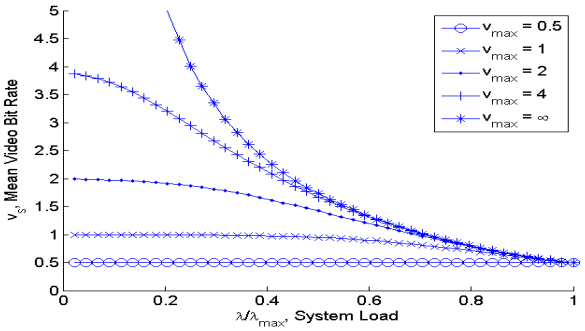
Fig. 3: Measured CQI probability distribution function on a live LTE network.

Other traffic-related parameters are configured as $T = 20$s, $\sigma = 5$Mbits and $v_{\min} = 0.5$Mbps. Based on the stability condition, we have $\lambda_{\max} = 2.189$ flows/s. As of the opportunistic scheduling gain, we make use of the scheduling gain calculated in [18] for a MIMO 2×2 LTE system and an Additive White Gaussian Noise (AWGN) channel and that converges to $G(\infty) = 1.7$ starting from a number of active uses in the LTE cell equal to 15.

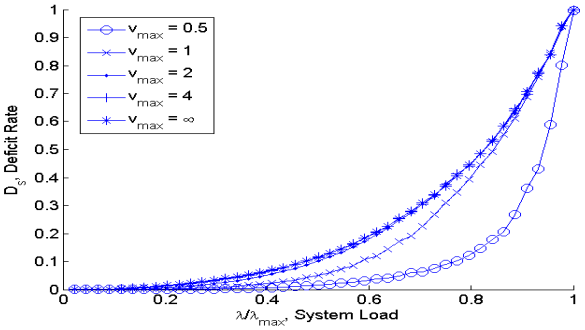
In Fig. 4, four performance metrics are shown with different $v_{\max}$ configurations. It can be observed that when $v_{\max}$ increases, $\bar{v}_s$ and D_s increases, but B_s and $\bar{\gamma}_e$ decreases, meaning that in the case of $v_{\min} = 0.5$Mbps, decreasing $v_{\max}$ can benefit $\bar{v}_s$, D_s and B_s regardless of the trade-offs of $\bar{\gamma}_e$ reduction. We can also observe that the deficit rate is not highly influenced when $v_{\max}$ is large. Therefore, we believe that buffer surplus is a better metric than deficit rate for dimensioning purposes.

VI. Conclusions

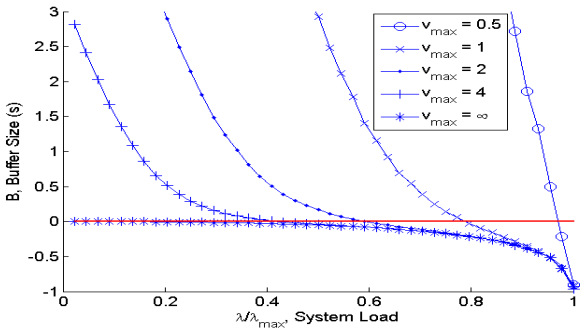
In this paper, we develop a flow-level model to describe the performance of adaptive streaming in wireless mobile



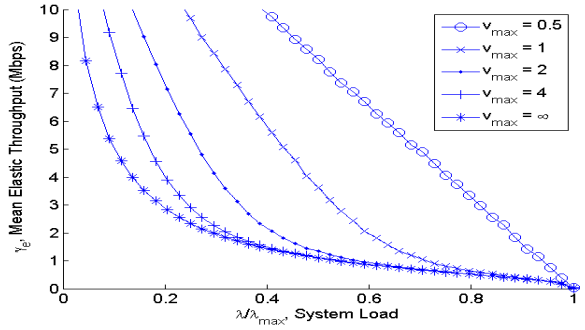
(a) Mean streaming video bit rate



(b) Streaming deficit rate



(c) Streaming buffer surplus



(d) Mean elastic throughput.

Fig. 4: System performance with different $v_{\max}$ settings.

networks carrying elastic traffic. We propose an approximate model which reduces the complexity of solving system metrics with heterogeneous conditions. Several performance metrics like mean video bit rate, deficit rate, buffer surplus and elastic mean throughput are taken as the main KPIs. It is shown that good buffer surplus and mean elastic throughput can be obtained by properly controlling the maximum video bit rate. In the future, the impacts of other system parameters such as the chunk duration and the buffer size need to be examined.

ACKNOWLEDGEMENT

This work has been carried out in the framework of IDEFIX project, funded by the ANR under the contract number ANR-13-INFR-0006.

REFERENCES

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," Tech. Rep.

[2] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *Communications Magazine, IEEE*, vol. 50, no. 4, pp. 20–27, April 2012.

[3] I. D. 23009-1, "Information Technology - Dynamic Adaptive Streaming Over HTTP (DASH) - part 1: Media Presentation Description and Segment Formats," Tech. Rep., 2011.

[4] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the internet," *MultiMedia, IEEE*, vol. 18, no. 4, pp. 62–67, April 2011.

[5] T. Bonald and A. Proutière, "Wireless downlink data channels: user performance and cell dimensioning," in *Proceedings of the 9th annual international conference on Mobile computing and networking*. ACM, 2003, pp. 339–352.

[6] T. Bonald, S. Borst, N. Hegde, M. Jonckheere, and A. Proutiere, "Flow-level performance and capacity of wireless networks with user mobility," *Queueing Systems*, vol. 63, no. 1-4, pp. 131–164, 2009.

[7] A. Khlass, T. Bonald, and S. Elayoubi, "Flow-level performance of intra-site coordination in cellular networks," in *Modeling Optimization in Mobile, Ad Hoc Wireless Networks (WiOpt), 2013 11th International Symposium on*, May 2013, pp. 216–223.

[8] B. Blaszczyszyn, M. Jovanovic, and M. Kadhem Karray, "Quality of Real-Time Streaming in Wireless Cellular Networks - Stochastic Modeling and Analysis," *ArXiv e-prints*, Apr. 2013.

[9] Y.-T. Lin, S. Elayoubi, and R. Nasri, "Capacity dimensioning for real-time video services in wireless mobile networks," *IEEE VTC Workshop*, 2015.

[10] T. Bonald and A. Proutière, "On performance bounds for the integration of elastic and adaptive streaming flows," in *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '04/Performance '04. New York, NY, USA: ACM, 2004, pp. 235–245. [Online]. Available: http://doi.acm.org/10.1145/1005686.1005716

[11] S. Borst and N. Hegde, "Integration of streaming and elastic traffic in wireless networks," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, May 2007, pp. 1884–1892.

[12] Y. Xu, S. Elayoubi, E. Altman, and R. El-Azouzi, "Impact of Flow-level Dynamics on QoE of Video Streaming in Wireless Networks," in *INFOCOM, 2013 Proceedings IEEE*, April 2013, pp. 2715–2723.

[13] T. Bonald and M. Feuillet, *Network Performance Analysis*. Wiley, 2011.

[14] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 1st ed. Academic Press, 2011.

[15] D. Tse, "Multiuser diversity in wireless networks," in *Wireless Communications Seminar, Standford University*, 2001.

[16] R. Combes, S.-E. Elayoubi, and Z. Altman, "Cross-layer analysis of scheduling gains: Application to lmmse receivers in frequency-selective rayleigh-fading channels," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2011 International Symposium on*. IEEE, 2011, pp. 133–139.

[17] Y. Bouguen, E. Hardouin, and F.-X. Wolff, *LTE et les réseaux 4G*. Editions Eyrolles, 2012.

[18] Y. Wang, "System level analysis of lte-advanced: with emphasis on multi-component carrier management," Ph.D. dissertation, Department of Electronic Systems, Aalborg University, 2010.