Evaluating user-perceived performance in high-speed backhaul networks

Raluca-Maria Indre, Philippe Olivier, Bruno Kauffmann, Nabil Benameur
Orange Labs
38-40 Rue du Général Leclerc
92794 Issy-les-Moulineaux, France
E-mail: {ralucamaria.indre, phil.olivier, bruno.kauffmann, nabil.benameur}@orange.com

***Abstract*—In this paper, we evaluate the performance perceived by end-users with very high access rates, connected to a common backhaul link that aggregates the traffic of multiple access areas. We model, at flow level, the way a finite population of users with heterogeneous access rates and traffic demands shares the capacity of this common backhaul link. We then evaluate several practically interesting use cases, focusing particularly on the performance of users subscribing to recent FTTH offers in which the user access rates may be of the same order of magnitude as the backhaul link capacity. We show that, despite such high access rates, reasonable performance can be achieved as long as the total offered traffic is well below the backhaul link capacity. The obtained performance results are used to derive simple dimensioning guidelines for backhaul networks.**

I. Introduction

Ever-growing Internet traffic raises the problem of efficient link dimensioning. The goal is to achieve a trade-off between capacity investments and the guarantee of a certain Quality of Service (QoS) level. The links of backhaul networks are particularly critical in this respect since they aggregate heterogeneous traffic from different kinds of end-users and are placed close to the edge of the network, which makes them particularly cost-sensitive. The goal of this paper is to analyze the performance perceived by users connected to high speed access networks, e.g., Fiber To The Home (FTTH) or Very high bit-rate Digital Subscriber Line (VDSL), and to derive simple dimensioning rules accordingly. This work is more specifically motivated by the introduction of very high speed (500 Mbit/s, 1 Gbit/s) offers for FTTH residential users. The actual use of this high bandwidth by some of these users clearly poses some acute QoS and/or dimensioning problems regarding the Passive Optical Network (PON) access and associated backhaul links with capacities of the order of a few Gbit/s.

In this paper, we propose a flow-level performance model which allows to evaluate the QoS perceived by the end-users. Given that users typically perceive performance at flow level (e.g., the duration of a file download), flow-level modeling has proven to be a convenient approach for both performance evaluation and link dimensioning purposes [1].

Most of the flow-level models proposed in the literature assume a Poisson flow arrival process, thus neglecting the problem of finite source population typical of an access/backhaul area. Flow-level models for a finite set of sources were first proposed by [2] and [3]. These papers, however, only considered uniform access rates, whereas we propose a model that handles both heterogeneous access rates and traffic demands. The present paper elaborates on [4] where simple models were developed to dimension IP access links carrying data traffic. The latter paper focused mainly on some useful approximations and only tackled the multirate aspect in a preliminary manner. Moreover, the derived approximations do not readily apply to the case of FTTH access where the access rates are in the same order of magnitude as the backhaul links.

In Section II of the present paper, we introduce the network context, the traffic model and related assumptions. Section III develops the multirate performance model which analyzes the way a finite number of active users share the downlink capacity of a common backhaul link. The considered users may have different access rates (i.e., users that subscribe to different offers) and different traffic demands. The bandwidth of the common link is shared according to the *Balanced Fairness* policy [5] which allows to allocate bandwidth among different classes of flows in a tractable way. In this model, like in the Erlang model for circuit-switched telephone networks, the main performance outputs depend only on one key parameter, namely, the average offered traffic.

We then define and compute several user-centric performance indicators such as the average user throughput and the congestion probability. We also compute the user *insatisfaction probability*, a novel performance indicator representing the probability that a user obtains less than a minimum satisfying rate. Reciprocally, the proposed model can be used to dimension the considered link, or to determine how many users can be accepted in an access area such that a target QoS is attained.

A significant contribution of the paper is to provide, in Section IV, some simple illustrative use cases with user populations having different access rates and different offered traffics. We assess the performance obtained by end-users and show the potential impact of greedy users on the QoS perceived by ”standard” users. Based on these results, we derive some guidelines for backhaul link dimensioning which are summarized in the Conclusion.

II. Network and traffic model

This section presents the considered network architecture, the traffic model and the related assumptions.

A. Network architecture

We consider a backhaul link which aggregates traffic from several access areas. These links are among the most limiting in today's networks. Indeed, backhaul links have capacities of a few Gbit/s and may aggregate the traffic of a few thousands of users. As FTTH offers with access rate of 500 Mbit/s or 1 Gbit/s are being launched, the access rates of the end users become of the same order of magnitude as the backhaul link, which raises some legitimate questions regarding the QoS perceived by the users. In the following, we analyze the end-user performance assuming that the considered backhaul link is the bottleneck, i.e., the most limiting link of the network. We focus on the downlink of this backhaul link since traffic is generally higher in this direction.

B. *Traffic context*

Todays Internet traffic mainly consists of video streaming, TCP-controlled data transfers (web browsing, downloads, P2P, etc.) and real-time flows such as audio/video conversational flows. TCP-controlled traffic and HTTP adaptive streaming represent about 90% of the total traffic [6]. This traffic is said to be *elastic* in the sense that the duration of the transmission dynamically adjusts to the available bandwidth. The performance of an elastic flow is thus perceived by the user through the time it takes to complete a transfer, or equivalently through the realized throughput.

In the following, we assume all traffic to be elastic which is a conservative assumption. Indeed, for elastic traffic, all packets that are lost are retransmitted until they are correctly received. On the contrary, packets from real-time flows (e.g., voice calls) that are lost are typically not retransmitted due to strict delay requirements thus decreasing the total network load in case of congestion.

C. *User session model*

Traffic generated by an active user is typically composed of a random succession of flow transfers and periods of inactivity. Each user flow corresponds to a sequence of contiguous transfers originated by the same user: a digital document (e.g., web page, e-mail, video) or several documents transferred successively or in parallel (e.g., elements of a web page or successive e-mails). The inactivity period typically corresponds to the time during which the user consults the transferred document and is referred to as the "think time"or "silence time", see, e.g., [4] for details. We assume that users can fully use their access rate during flow transfers (apart from congestion situations on the considered link). This assumption may be proven to be a worst case scenario, which makes it robust for dimensioning, see the discussion below in IV-A1. OFF periods of silence correspond to zero bit rate.

We consider a total population of N *active* users, i.e., users that generate active sessions as described above during the considered busy period. Note that this corresponds only to a fraction of total connected users. The total population consists of K classes with N_k users each, with $\sum_{1\leq k\leq K} N_k = N$. Each user class may have specific traffic requirements and different access rates. For each set of class-k users:

- $\mathrm{E}[V_k]$ is the mean flow volume generated;
- $\mathrm{E}[S_k]$ is the mean silence time duration;
- $\mathrm{E}[T_k]$ is the mean flow transfer duration;
- c_k is the common access rate of class-k users, it constitutes a kind of "peak rate", the maximum rate at which a user is able to transmit at any given time.

The above parameters describe the ON/OFF profiles of active user sessions. They are all input parameters to the model except for the mean flow transfer time $\mathrm{E}[T_k]$ which is an output performance parameter. Indeed, as explained above, the transfer time of elastic flows depends on the possible congestion situations in the network.

We are now able to characterize the average traffic generated in each class. The per-user *carried traffic* is the average bit rate generated by a user that can be measured on the considered link during the busy period. It is defined as follows:

$$b_k \equiv \frac{\mathrm{E}[V_k]}{\mathrm{E}[T_k] + \mathrm{E}[S_k]}. \quad (1)$$

Note that the carried traffic is actually a performance parameter as it depends on the time necessary to transfer documents.

The user demand is characterized by one key parameter, namely, per-user *offered traffic* a_k. The offered traffic represents the average bit rate a user would generate if it would not be constrained by the network. The flow transfer time would thus be $\mathrm{E}[V_k]/c_k$ and the offered traffic writes:

$$a_k = \frac{\mathrm{E}[V_k]}{\mathrm{E}[V_k]/c_k + \mathrm{E}[S_k]}. \quad (2)$$

III. Performance model

This section details the multirate performance model and introduces the proposed user-centric performance metrics.

A. *System state equations*

We model the way in which N active users share the capacity C of a backhaul link. The system state is a random process where user flows arrive and disappear dynamically. We assume a stationary regime where the state vector $x = (x_1, ..., x_k, ..., x_K)$, $0 \leq x_k \leq N_k$, gives the number of flows in progress in each class at equilibrium. We denote by $\mathcal{S} = \{x = (x_k)_{k=1,K} \in \mathbb{N}^K;\, \forall k,\, x_k \leq N_k\}$ the set of admissible states and by $x \cdot c = \sum_{1\leq k\leq K} x_k\, c_k$ the scalar product of state and access rate vectors, which represents the overall required bandwidth at a given state.

Since we consider elastic traffic only, we model this traffic at flow level assuming the packet level mechanisms of TCP realize fair bandwidth sharing among ongoing flows of the considered class [1]. Thus, the system can be modeled as a network of Processor Sharing queues working in parallel. The service speed of each queue depend on the network state: we assume a total amount of bandwidth $\phi_k(x)$ is allocated to each class k at state x (with $x_k > 0$). This amount is equally shared so that each class k flow receives a portion $\phi_k(x)/x_k$. This bandwidth allocation scheme must satisfy the constraints set by the link capacity and user access rates:

$$\sum_{1\leq k\leq K} \phi_k(x) \leq C \;\text{ and }\; \forall k = 1, K,\; \phi_k(x) \leq x_k\, c_k. \quad (3)$$

Assume for now that the flow volumes $(V_k)_{k=1,K}$ and silence times $(S_k)_{k=1,K}$ random variables have exponential distributions and are independent to each other. The state vector $x(t)$ then forms an homogeneous Markov process, and more precisely a multi-class Birth-Death process, whose per-class arrival and departure rates at state x are the following:

$$\lambda_k(x) = (N_k - x_k)/\mathrm{E}[S_k], \qquad x \in \mathcal{S}, \qquad (4)$$
$$\mu_k(x) = \phi_k(x)/\mathrm{E}[V_k], \qquad x \in \mathcal{S}. \qquad (5)$$

Such a Markov process is known to have an equilibrium regime whose stationary distribution $\pi(x)$ is the unique solution to the system of Kolmogorov balance equations [7]

$$\sum_{k=1}^{K} [\lambda_k(x - e_k)\pi(x - e_k) + \mu_k(x + e_k)\pi(x + e_k)] = \sum_{k=1}^{K} [\lambda_k(x) + \mu_k(x)]\,\pi(x), \quad \forall x \in \mathcal{S}, \qquad (6)$$

together with the normalizing condition $\sum_{x\in\mathcal{S}} \pi(x) = 1$. In these equations, e_k denotes the unit vector in the k^{th} direction.

B. Bandwidth allocation

When the total required bandwidth $x \cdot c$ at a given flow state x is no greater than link capacity C, each flow of class k may be allocated its peak rate c_k. When the link is congested, i.e., $x \cdot c > C$, it is no longer possible to ensure full access rate use for all flows in progress, and thus bandwidth must be divided among flow classes in a non trivial manner.

Numerous bandwidth allocation strategies have been proposed in the literature, see [8], [9], for instance. Most of these algorithms try to achieve fairness by maximizing some overall utility function. Such utility-based allocations have two major drawbacks. First, they generally do not lead to a reversible Markov process, and thus do not allow tractable analysis using closed-form solution to the equilibrium equations [7]. This is true in particular in the multirate scenario we are interested in. The equilibrium equations (6) must then be solved numerically, leading to high computation times as the system size grows. Second, these allocations are *sensitive* in the sense that the performance results depend on detailed traffic characteristics such as the flow arrival process or flow size distribution [5]. This is an important limitation since performance results are in that case valid for exponential flow sizes and silence times only, which do not represent realistic assumptions [10].

The Balanced Fairness allocation proposed in [5] overcomes the above limitations of utility-based allocations. An allocation is said to be balanced if, for any pair of classes (i, j), the product of bandwidth consumptions from state x to state $x - e_i - e_j$ does not depend on the path followed between these two states:

$$\phi_i(x)\phi_j(x - e_i) = \phi_j(x)\phi_i(x - e_j), \forall i, j, \forall x \in \mathcal{S}, x_i, x_j > 0.$$

This condition is equivalent to the existence of a positive balance function $\Phi(x)$ such that

$$\phi_k(x) = \frac{\Phi(x - e_k)}{\Phi(x)}, \ \forall k = 1, K, \ \forall x \in \mathcal{S}, \ x_k > 0. \qquad (7)$$

The *Balanced Fairness* is the unique balanced allocation that maximizes resources utilization, i.e., it saturates at least one of the constraints expressed in (3) at each state. The corresponding balance function can be defined as follows:

$$x \cdot c \le C : \phi_k(x) = x_k\, c_k, \forall k, \text{and } \Phi(x) = \prod_{k=1}^{K} \frac{1}{c_k^{x_k} x_k!},$$
$$x \cdot c > C : \sum_{k=1}^{K} \phi_k(x) = C \text{ and } \Phi(x) = \frac{1}{C} \sum_{k=1}^{K} \Phi(x - e_k).$$

In the following, we consider that bandwidth allocation is performed according to Balanced Fairness.

In a Markovian setting, the balance property is equivalent to the reversibility of the Markov process (as long as the arrival rates are also balanced). Thus, for exponentially distributed flow volumes and silence times, the Markov process $x(t)$ converges to an equilibrium regime, whatever its initial state. The stationary distribution $\pi(x)$ is the unique solution to local balance equations. The state probabilities have the following product form solution:

$$\pi(x) = \pi(0)\frac{\Phi(x)}{\Phi(0)} \prod_{k=1}^{K} r_k^{x_k} \frac{N_k!}{(N_k - x_k)!}, \ x \in \mathcal{S}, \qquad (8)$$

where $r_k = \mathrm{E}[V_k]/\mathrm{E}[S_k] = a_k c_k/(c_k - a_k)$ is a quantity related to the offered traffic a_k. This result is shown to have the insensitivity property [5] which ensures that performance results *only* depend on the mean traffic demand of each class.

C. Performance metrics

The probability distribution of the number of ongoing flows (8) allows us to derive various performance indicators. All of them clearly inherit the insensitivity property quoted above.

In the case of elastic traffic, the user perceives QoS in terms of the average time needed to transfer a document. The *useful rate* d_k defined as the mean flow volume to the mean flow transfer time ratio for class-k flows is an equivalent performance metric [1]. It can be expressed as

$$d_k \equiv \frac{\mathrm{E}[V_k]}{\mathrm{E}[T_k]}. \qquad (9)$$

From definition (9) and Little's formula applied to class-k flows in progress, $\mathrm{E}[x_k] = \mathrm{E}[\lambda_k]\,\mathrm{E}[T_k]$, we deduce $d_k = \mathrm{E}[V_k]\,\mathrm{E}[\lambda_k]/\mathrm{E}[x_k]$. The useful rate finally writes

$$d_k = r_k \frac{N_k - \mathrm{E}[x_k]}{\mathrm{E}[x_k]}. \qquad (10)$$

Note that in view of (1) and (10), we obtain a conservation law between the mean rate of all users of class k and the useful rate of their active flows, $N_k b_k = \mathrm{E}[x_k] d_k$.

Although the useful rate is an adequate performance measure for elastic traffic, it remains only an average metric. We now define performance metrics which reflect the proportion of time the end users perceive a degradation of their QoS. The *link congestion probability* is the stationary probability that flows in progress are allocated a bandwidth less than the user access rate. It is denoted as P_C and is independent of the class of flows:

$$P_C = \sum_{x\in\mathcal{S},\, x\cdot c > C} \pi(x). \qquad (11)$$

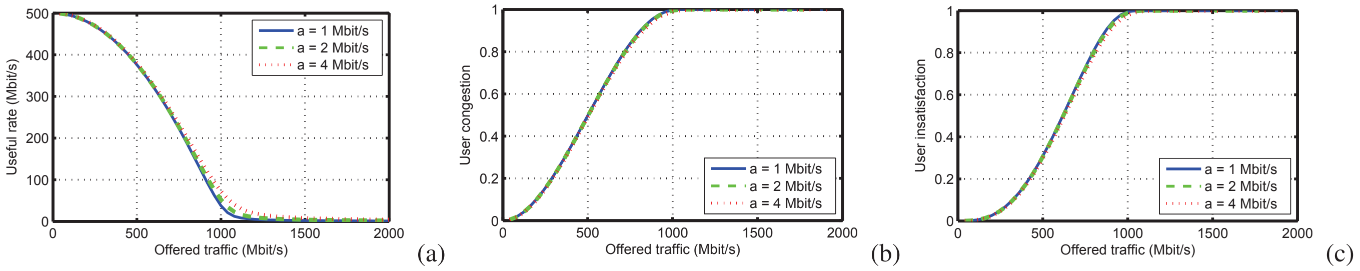

Fig. 1. Useful rate (a), user congestion probability (b), and user insatisfaction probability with $\beta = 60\%$ (c) vs total offered traffic for different per-user offered traffics - Link capacity = 1 Gbit/s and access rate = 500 Mbit/s

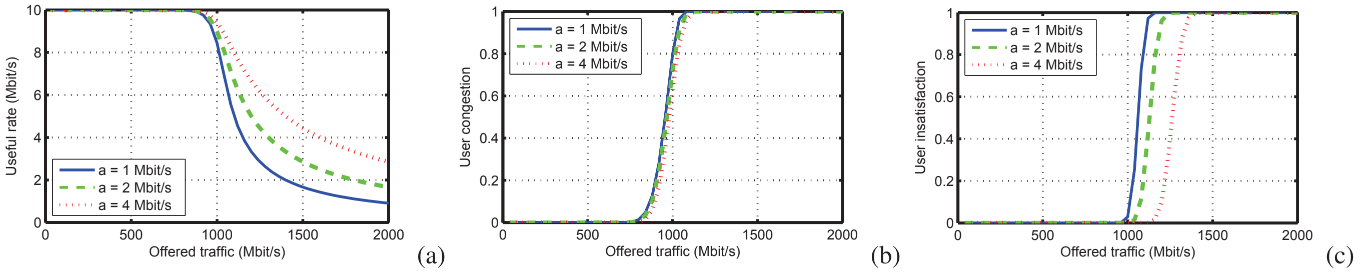

Fig. 2. Useful rate (a), user congestion probability (b), and user insatisfaction probability with $\beta = 60\%$ (c) vs total offered traffic for different per-user offered traffics - Link capacity = 1 Gbit/s and access rate = 10 Mbit/s

The link congestion probability is a performance metric that reflects the network point of view. For present purposes in this paper, it is desirable to consider the user point of view. We thus define the *user congestion probability* $P_U(k)$ which measures the probability that a flow is allocated less than its access rate given that the flow is in transfer. Unlike the link congestion, the user congestion probability depends on the flow class; it is obtained by considering the weighted distribution $x_k\pi(x)/\mathrm{E}[x_k]$:

$$P_U(k) = \frac{1}{\mathrm{E}[x_k]} \sum_{x \in \mathcal{S},\, \phi_k(x) < x_k c_k} x_k \pi(x), \quad k = 1, K. \qquad (12)$$

The congestion probability is a rather strict performance criteria since it only relates to whether or not the flows in transfer attain their access rates. In practice, users may be satisfied with a rate which is only a fraction of the access rate $\beta_k\, c_k$, $\beta_k \in [0,1]$, that we refer to as the *satisfying rate*. The latter may represent a bit rate that the operator aims at guaranteeing. We define the *user insatisfaction probability* as the probability that a flow of class k is allocated a bandwidth less than the satisfying rate, given that the flow is in transfer:

$$P_I(k) = \frac{1}{\mathrm{E}[x_k]} \sum_{x \in \mathcal{S},\, \phi_k(x) < \beta_k x_k c_k} x_k \pi(x), \quad k = 1, K. \qquad (13)$$

IV. Use cases

We now apply the proposed model to a few practically interesting use cases. We consider both the case in which all users have the same access rate and the case of different access rates and different traffic demands.

A. *Uniform access rate*

1) Uniform demand: We first consider the case of an FTTH access network in which a 1 Gbit/s backhaul link aggregates traffic from FTTH clients that have subscribed to high access rate offers. We assume all users have a 500 Mbit/s access rate and a uniform demand. Figure 1 gives the realized useful rate (a), the user congestion probability (b), and the user insatisfaction probability with $\beta = 60\%$ (c) as functions of the overall offered traffic Na for various values of the per-user offered traffic. An important observation is that performance mainly depends on the total offered traffic. Indeed, as long as Na is well below the backhaul capacity, reasonable performance can be achieved despite the high access rates of end users. On the contrary, when the total offered traffic exceeds the link capacity, the link becomes saturated and the useful rate drops to 0, while the user congestion and insatisfaction probabilities attain 100%. Note also that since the per user offered traffic is very small compared to the backhaul link capacity, performance results are practically insensitive to this parameter.

From plots (a) and (b) we note that the user-perceived congestion level may appear rather large, say, about 20-40%, while the useful rate remains at an acceptable level, more than 80% of the access rate. The congestion probability is thus a very strict performance indicator. The insatisfaction probability shown in plot (c) may be a more practical performance metric although the value of the satisfying rate must be set carefully.

Consider now that the same 1 Gbit/s backhaul link is shared by a population of users having an access rate of c = 10 Mbit/s, typical of a classical Asynchronous Digital Subscriber Line (ADSL) access network. Figure 2 gives the same performance metrics as above; all traffic parameters are the same, except for the access rate. Observe that, as long as the total offered traffic is lower than the backhaul link capacity, the useful rate is approximately equal to the access rate. We say that the link is *transparent* in this region since it does not impact the user perceived performance. This result is in accordance with the ones obtained in [4].

In the case of FTTH networks, the transparent regime is almost inexistent and the approximation $d \approx c$ whenever $Na < C$ no longer holds. Indeed, in Figure 1, the useful rate d is almost always below the access rate c and thus the link is far from being transparent.

Comparing Figures 1 and 2, we observe that all performance parameters behave better (although in a relative way since the access rates are different) when all users have an access rate of 10 Mbit/s. Performance is thus better when the peak rate attained by the users is lower. Assuming that all users can attain their full access rate is therefore a conservative assumption since performance results would be better if flows would be rate-limited elsewhere in the network, e.g., by a codec or a different bottleneck link.

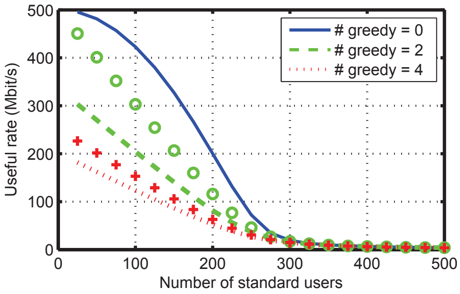

Fig. 3. Useful rate vs number of standard users in the presence of greedy users - Link capacity = 1 Gbit/s and access rate = 500 Mbit/s; lines refer to standard users useful rate, circles and crosses refer to that of greedy users

2) Impact of greedy users: We now evaluate the impact of having some users which are greedy, i.e., they generate flow transfers almost permanently such that their individual traffic demand is close to their access rate. Figure 3 provides the useful rate obtained by each class as a function of the number of standard users. Standard users have a 4 Mbit/s per-user traffic demand, while greedy users have an offered traffic equal to 99% of their access rate. In this case with very contrasted demand, the performance perceived by greedy users is slightly better than that of standard users, but both of them quickly deteriorate as the population size increases. Actually, even for a small population size, the presence of only two greedy users is not acceptable if we require that standard users get more than half of their access rate, on average.

Admittedly, the considered scenario is an extreme case. Although it may seem natural that the presence of one or more greedy users with an average demand of 500 Mbit/s on a 1 Gbit/s link will considerably degrade performance, the proposed model helps quantify the precise impact of such greedy users. Contrary to naive intuition, even in the presence of two greedy users that require nearly 500 Mbit/s each, some bandwidth is still available for standard users. This is due to the principle of fair bandwidth sharing among ongoing flows with same access rate.

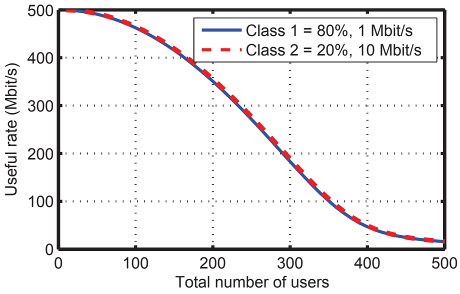

Fig. 4. Useful rate vs total number of users for two classes of users - 1 Gbit/s link capacity and 500 Mbit/s access rate

We now consider the case of heterogeneous demand with non-greedy users. Figure 4 shows the mean useful rate for two classes of end users, each with 500 Mbit/s access and with average offered traffic of 1 and 10 Mbit/s, respectively. Given that both classes have traffic demands which are very small compared to the capacity of the backhaul link, their useful rates are nearly the same, which is consistent with results of Figure 1. Performance in terms of user congestion or user insatisfaction probability is also very similar (not shown here for the sake of brevity).

B. Multiple access rates

1) Standard users only: We now analyze the practically interesting use case in which end-users subscribe to different offers, thus benefiting from different access rates. Specifically, we assume that the user population is made of two classes of 'standard' customers with reasonable traffic demand: class 1 users subscribe to a 100 Mbit/s offer and express a 2 Mbit/s traffic demand, while class 2 users subscribe to a high access rate offer at 500 Mbit/s and have a 4 Mbit/s per-user traffic demand. Class 1 users represent 80% of the total population.

Figure 5 provides performance results in terms of the useful rate (a) and the user insatisfaction probability corresponding to a fraction $\beta_1 = \beta_2 = 80\%$ of the respective access rates (b), as functions of the total number of active users N. Also reported are the results of a fluid approximation according to which the low access rate users are assumed to occupy a constant bandwidth $N_1 a_1$ (and thus to perceive perfect performance) while the high access rate flows share the remaining capacity $C - N_1 a_1$.

First, remark that the simple fluid approximation is quite accurate. Consequently, if we are interested in determining only the performance of high rate users, it is sufficient to consider a single user class sharing a link capacity of $C - N_1 a_1$. Results shown in Figure 5 also indicate that in case of high link congestion, when the number of users is very large, there is no performance differentiation between the two classes of customers, both classes obtaining equally small useful rates.

We then observe that the performance of low rate users degrades at a much slower pace than that of high rate users, which at the same time explains the accuracy of the fluid approximation. It is noteworthy that low rate users maintain a high fraction of their access rate until, say, a total population of 300 users, while high rate users have already perceived

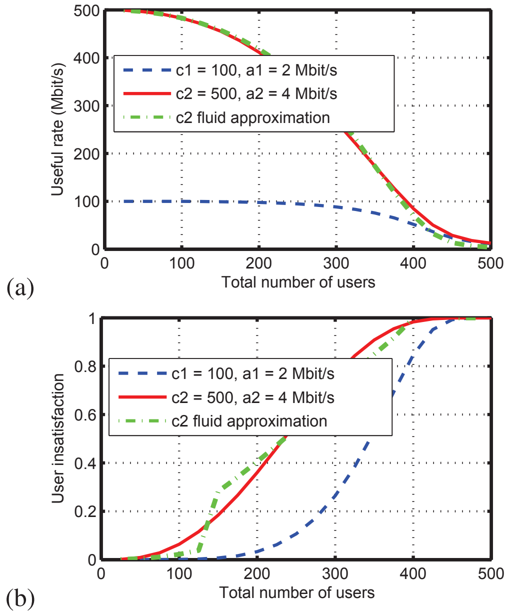

Fig. 5. Useful rate (a) and user insatisfaction probability with $\beta_1 = \beta_2 = 80\%$ (b) vs total number of users for two classes of users with distinct access rates - Comparison with the fluid approximation for high rate users

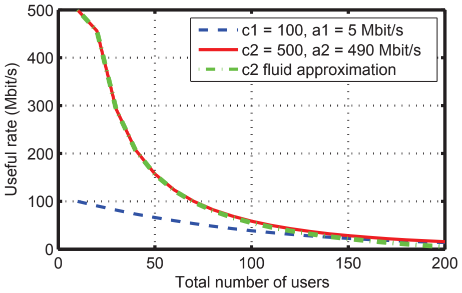

Fig. 6. Useful rate vs total number of users in the presence of high access rate greedy users - Comparison with a fluid approximation for high rate users

a considerable degradation of their performance. Thus, a useful lesson we can learn from this is that when commercial offers with high access rates are launched, network operators should primarily ensure that capacity is sufficient to provide a reasonable performance for high access rate users; low rate users are less likely to experience severe QoS degradation.

2) Impact of greedy users: Finally, we consider a pessimistic scenario in which all high rate customers are greedy users. Figure 6 provides the useful rate obtained by each class as a function of the total number of users, when 10% of users have an access rate of 500 Mbit/s and a 490 Mbit/s per-user traffic demand; the remaining 90% have an access rate of 100 Mbit/s and a 5 Mbit/s demand. The overall behaviour of the system is rather similar to the one presented in Figure 5, except that performance degradation occurs very quickly, as soon as there are a few tens of active customers. Such a scenario is clearly unacceptable for an operational network, and an upgrade of the considered link capacity would be required in this case. Once more, the fluid approximation is perfectly able to accurately predict performance of high rate users.

V. Conclusion

We proposed a general flow-level performance model which we applied to evaluate the performance of end-users with very high access rates. The capacity of the common backhaul link is shared according to the Balanced Fairness policy. To the best of our knowledge, this is the first time that Balanced Fairness is applied to a finite source population model. In order to assess user-perceived performance, we defined and evaluated user-centric performance metrics such as the useful rate, the user congestion probability and the user insatisfaction probability.

Another significant contribution is the application of the model to a few illustrative use cases where some users have very high access rates, possibly comparable to the backhaul link capacity. The impact of having multiple access rates and users with various traffic demands (such as heavy users) has been analyzed. Some approximations were discussed and benchmarked against previous studies on ADSL networks. The main guidelines that could be derived from the considered use cases are:

- For dimensioning purposes, QoS analysis may be simplified by assuming all users attain their full access rate which constitutes a conservative approach;
- When launching high rate commercial offers, a network provider should primarily ensure the QoS of the high rate users; the QoS of low rate users is only marginally impacted;
- When considering users with different access rates, the performance of high access rate users can be determined using a single rate model applied to the residual link capacity left by low rate users.
- The impact of greedy users on the overall performance can be severe; the proposed model allows to quantify this impact and to dimension the network accordingly.

References

[1] S. B. Fredj, T. Bonald, A. Proutière, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," in *Proceedings of ACM SIGCOMM*, 2001.

[2] D. P. Heyman, T. V. Lakshman, and A. L. Neidhardt, "A new method for analysing feedback-based protocols with applications to engineering web traffic over the internet," *SIGMETRICS Perform. Eval. Rev.*, vol. 25, no. 1, Jun. 1997.

[3] A. W. Berger and Y. Kogan, "Dimensioning bandwidth for elastic traffic in high-speed data networks," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, 2000.

[4] T. Bonald, P. Olivier, and J. Roberts, "Dimensioning ip access links carrying data traffic," *Annals of Telecomm.*, vol. 59, no. 11-12, 2004.

[5] T. Bonald and A. Proutière, "Insensitive bandwidth sharing in data networks," *Queueing Syst. Theory Appl.*, vol. 44, no. 1, 2003.

[6] A. Finamore, M. Mellia, M. Meo, M. Munafo, and D. Rossi, "Experiences of internet traffic monitoring with tstat," *Network, IEEE*, vol. 25, no. 3, 2011.

[7] R. Serfozo, *Introduction to Stochastic Networks*, ser. Applications of mathematics. Springer New York, 1999.

[8] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journ. of the Op. Research Society*, vol. 49, no. 3, 1998.

[9] L. Massoulié and J. Roberts, "Bandwidth sharing: Objectives and algorithms," *IEEE/ACM Trans. Netw.*, vol. 10, no. 3, 2002.

[10] P. Olivier and N. Benameur, "Flow level ip traffic characterization," in *Proceedings of ITC '17*, 2001.