A Flow-Level Performance Evaluation of Elastic Traffic Under Low Latency Queuing System

Mohamed El Hedi Boussada, Jean-Marie Garcia, Mounir Frikha

A Flow-Level Performance Evaluation of Elastic Traffic under Low Latency Queuing System

Mohamed El Hedi Boussada Mounir Frikha
Mobile Network and Multimedia
SUPCOM, University of Carthage
Ariana-Tunisia
{med.elhadi.boussada, m.frikha}@supcom.tn

Jean Marie Garcia
Services and Architectures for Advanced Networks
LAAS-CNRS
Toulouse, France
jmg@laas.fr

***Abstract*—Internet tends progressively to be more interactive by supporting real-time communications into this packet-based environment. It is anticipated that online applications will rapidly develop and contribute by a significant amount of traffic in the near future. These applications have strict requirements in terms of delay and throughput which can't be met only if they are prior. However, at high demand of real-time traffic, elastic traffic (which is transported generally by TCP) may not have sufficient resources to be transported with reasonable quality of service (QoS). Therefore, it is primordial to manage properly the resource-network in order to provide the QoS required by both each type of services. In this paper, we present a new fluid model to evaluate the performance of elastic traffic under Low Latency Queuing (LLQ) system combining a priority queue for delay-sensitive applications with a number of Class Based Weighed Fair Queues (CBWFQ) for elastic traffic. The originality of our contribution consists on focusing on the average total number of flows passing through the whole system by approximating it as a lossless best effort system with total load equal to the total carried load of the original system (no blocked load). The core of our analysis is based on some approximations proven for balanced fairness allocation, which provides a reasonable framework for estimating bandwidth sharing among elastic traffic for best effort allocations. Results issued from this allocation are then exploited to deduce the performance of CBWFQ system. Detailed packet level simulations are used to verify the effectiveness and the accuracy of our analysis. The approach presented in this paper allows a rapid performance evaluation of elastic or rate-adaptive traffic circulating in the actual networks.**

Keywords*—Elastic traffic, Real time traffic, Low Latency Queuing (LLQ), Class Based Weighted Fair Queieng (CBWFQ), Quality of Service (QoS), Balanced Fairness.

I. INTRODUCTION

Upgrades in Internet connection speeds, with the emergence of 4G LTE cellular and fiber optic communications, has led to a multiplication of services offered by modern telecommunication networks and to unprecedented growth in the number of users and traffic volumes that they generate [1], [2], [3], [4]. Internet traffic is forecasted to increase nearly threefold over the next 5 years. Major portion of this traffic will be dedicated mainly to various forms of video. Globally, IP video traffic will be 82 percent of all consumer Internet traffic by 2021. Live Internet video will contribute by 13 percent in this video traffic and online gaming will grow by 62 percent from 2016 to 2021 [5].

Internet tends, then, to be more interactive by supporting real-time communications in this packet-based environment. Therefore, we shall distinguish two broad categories of Internet traffic: real-time traffic and elastic traffic. Real-time traffic is generated by applications such as video-conference applications, Voice over Internet protocol applications (VoIP applications), online videos, online gaming, etc. These applications have strict bandwidth, end-to-end packet delay and jitter requirements for reliable operations [6]. Elastic traffic is generally transported by TCP (Transport Control Protocol) and it can adapt its rate to the available capacity [7]. Although that TCP is typically designed to carry data (transfer of documents such as files or web pages) , it is also suitable, today, for transporting modern video streaming services such YouTube and Netflix traffic for example [8], [9], [10]. It is reported that most commercial streaming traffic is carried over TCP [9], [11], [12].

Facing this diversity of applications supported by the actual IP networks, transporting all this traffic over one converged packet-oriented network is not built in the same manner. Operators and service providers generally offer all services at different service levels according to the Quality of Service (QoS) level required by both the service and the client [1]. The Diffserv architecture presents an architecture where packets are marked and divided into service classes as they enter the network [13]. Network nodes are then able to provide different services using some scheduling algorithms. In this context, different scheduling policy were proposed. In priority queueing (PQ), higher priority traffic is transmitted before lower priority traffic, with separate buffers for each class of traffic. Weighted Fair Queueing (WFQ) allocates an equal share of the bandwidth for each flow according to the weight attributed to it [14]. The Class Based Weighted Fair Queueing (CBWFQ) is an advanced form of WFQ that supports user defined traffic classes. A queue is allocated for each class, and the traffic belonging to that class is directed to the queue assigned to it [15]. The CBWFQ dynamically allocates the available bandwidth to each traffic class based on the queue's weight. When one class does not use all its bandwidth, the other classes will share this residual capacity, which creates a coupling aspect between queues in order to optimize the resource sharing and ensure a more efficient traffic transport. Although that this policy may be able to meet the QoS measures needed by the different elastic traffic types, it seems to be not adopted to guarantee the strict QoS requirements of delay-sensitive applications [16]. The composition between these two policies of traffic management techniques (fixed priority policy and bandwidth sharing-based policy) is considered by many telecommunication equipment constructors like Cisco and Huawei [17], [18]. The low-latency queuing (LLQ), which is being used frequently on the Internet, is a feature developed by Cisco to bring strict priority queuing to class-based weighted fair queuing [18]. Priority queuing is used to satisfy the strict delay constraints for real-time traffic, whereas CBWFQ is used to assure acceptable throughput for elastic traffic classes.

During high demand of real-time traffic, the blocking probability of coming streaming calls becomes important and the number of ongoing elastic flows indefinitely increases. Continuous monitoring of the network is then essential to detect as quickly as possible any degradation of performance and be able to reconfigure the network to cope with it [19]. At this level, end-to-end measurements are extremely complex to be implemented [19], [20]. For operators, the solution is to use traffic-engineering techniques to anticipate the degradation of quality of service resulting from the phenomena of congestion. However, the use of these techniques assumes to have models, theoretical methods, analytical approximations and appropriate software tools to properly manage the network [1], [20]. Until now, operators do not have a real reliable model to modelize and predict the performance of the traffic circulating in their infrastructures. It is still difficult to predict the performance of large-scale operator networks with millions of users, millions of flows, various access terminal, various types of multimedia applications, complex routing protocols and unexpected user behaviours.

Flow level models represent very well the great dynamicity of the traffic as it makes it possible to represent the connections and disconnections of users on an application (or service). However, studying the flow level performance of elastic traffic in a LLQ system (or an equivalent system) has not received much attention in literature. Most works evaluating the performance of elastic traffic in the presence of real-time traffic assumed that TCP traffic should not be transported in a differentiated manner [6], [21], [22], [23], [24], [25], [26]. In general, this viewpoint remains insufficiently appreciated in actual networks where TCP is used to carry several types of traffic with different QoS requirements. In this paper, we will overcome this limitation by proposing a general flow-level model of a LLQ system where elastic flows are distributed into a number of CBWFQ queues to be differently managed. The main idea behind our analysis is to focus on the average total number of flows crossing the system by approximating the original system as a lossless Best Effort system with a total load equal to the total carried load (no blocked load) of the original system. The conservation of the total average number of flows is not largely considered at the previous works in this field which generally tends to analyse the problem assuming that the dynamics of streaming requests are much slower or much faster than those of elastic requests. Our approach is based on adopting some approximations for balanced fairness allocation, which provides a reasonable framework for estimating bandwidth sharing among elastic traffic for best effort allocations, to predict the performance of elastic traffic under a LLQ system.
The rest of this paper is organised as follows: An overview of related work is shown in the next section. In the third section, we describe our model of the system to be studied. The fourth section is devoted to give useful approximations proven for balanced fairness allocation in best effort system. These approximations will be exploited in the fifth section to evaluate the performance of elastic traffic under a LLQ system. The accuracy of the proposed analytical results is discussed in the last section basing on some detailed packet level simulations with Network Simulator (NS 2).

II. Related work

The coexistence between elastic and real time traffic is a vexed problem [27]. Some authors have proposed that real time traffic should be TCP-friendly, so that it can fairly share network resources with TCP traffic [25], [27]. In practice, real time applications often need some form of priority to function adequately, which make this approach not applicable in a real context. In [6], [26] and [21], the authors was interesting in studying the performance of elastic flows in a network where real time traffic is prior and non-adaptive. In addition, they justified the need for an appropriate admission control mechanism for real-time traffic to not jeopardize the performance of these sensitive flows. In [22] and [23], elastic flows were supposed to have a guarantee bit rate (nominal bit rate) and each arriving flow (elastic or non-elastic) will be blocked if it may negatively affect the performances of ongoing flows.

In the presence of variable service rate for elastic traffic, it is so difficult to have a regular representation of the generating matrix of the Markov chain modelling arrivals and departures of flows in such system. Approximation techniques are proposed then as an alternative to Markov chain analysis. The integration of streaming and elastic traffic was mainly treated under a quasi-stationary assumption. Under this assumption, we assume that the number of ongoing elastic flows rapidly evolves and attain the stationary regime while the number of active streaming calls remains unchanged [22], [23], [21], [26]. This assumption is reasonable when we consider the combination of voice calls and web browsing or email applications [22]. In each state of real-time flows, the remaining capacity (after satisfying the streaming flows requirements) will be reserved for elastic flows and then we construct a new elastic system with a reduced capacity. When the probability that this remaining capacity is inferior to the total elastic traffic intensity is not negligible, the elastic system becomes unstable and the performances of elastic flows can not be easily predicted [6]. Another drawback of this approach is that it is difficult to guarantee the quasi-stationary assumption in a system with more than one class of streaming flows characterising by different arrival rates, different mean service times and different bit rates.

Assuming now that the dynamics of elastic flows are much slower than those of streaming flows (considering the combination of voice calls and large file transfer applications), another classical approach is proposed in the literature and it is based on estimating of the mean available capacity for elastic traffic. In general, this approach does not yield meaningful results, especially when the average capacity used by real time flows is close to the system capacity [6].

Other approaches are also proposed. By adopting an admission control for both elastic and real-time traffic, the authors of [22] aim to find a weighted version of the two regime-approximations described above for neutral traffic regime (the relative dynamics of both request types are similar) in order to evaluate approximately the blocking probability of each type of flows. Malhotra et al. proposed in [6] a model where the streaming flows are prior and have the same rate as elastic one. The approximations proposed in [6] focus on the conservation of the total workload of the global system, which modeled as a General Processor Sharing (GPS) system without priorities having a total load equal to the elastic load plus the effective load of real time traffic (the carried real time load). The assumption of equal peak rate for both stream and elastic traffic is very restrictive and not always true in practice. Moreover, the Malhotra's approach is completely sensitive to the detailed characteristics of traffics.

All these papers treated a limited system with two priority queues admitting that there is no differentiation on transporting TCP traffic. In [22], [23] and [26], elastic flows have not a rate limit and they equally share the remaining capacity. In a real context, user flows always have a rate limit which depends on the used access technique. The real behaviour of a general LLQ system is hard to be exactly modelled, especially because of the coupling aspect existing between the priority queue and the CBWFQ system on the one hand and between the CBWFQ queues itself on the other hand.

Our paper aims to overcome the limitations detailed above by :
- Presenting a general model for LLQ system that support different elastic and real time flow-classes.
- Providing a large-scale modelization of the LLQ system taking into account the strong coupling between the queues.

III. MODEL

We consider a large number of ADSL/UMTS/LTE subscribers distributed over a single geographic area and connected to the backbone network through an access line of capacity C $(Mbits/Second)$. This access line is assumed to be the bottleneck link for all the traffic flows generated by users. This link is a full-duplex link, which means that the transmission capacity is C $(Mbits/Second)$ in both directions (See figure 1).
Users generate a random number of real-time/streaming and elastic flow-classes. Let E be the set of elastic flow classes and S be the set of real-time flow classes.
Real-time flows are mainly defined by their rate and their mean holding-time. For each class $j \in S$, we define:

•$\tau_j(Second)$:The mean holding-time of flows.
•$d_j^{(s)}(Mbits/Second)$:The rate of each flow.

For each elastic class-i flows $(i \in E)$, we define:

•$\sigma_i(Mbits/Flow)$:The mean volume transferred by flows.
•$d_i^{(e)}(Mbits/Second)$:The maximum bit rate of each flow.

Flows arrive as an independent Poisson process with rate $\lambda_j^{(s)}(Flows/Second)$ for streaming class j and $\lambda_i^{(e)}(Flows/Second)$ for elastic class i. We refer to the

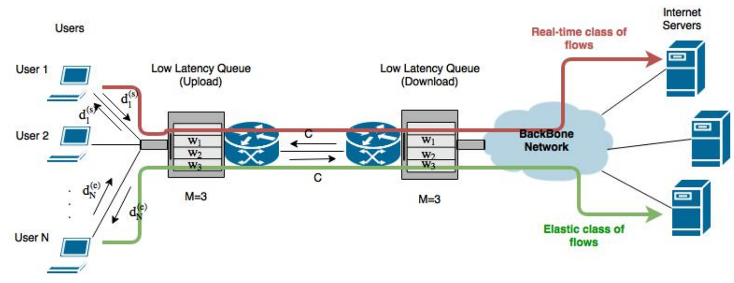

Figure 1. Studied system

product $\rho_i^{(e)} = \lambda_i^{(e)}\sigma_i(Mbits/Second)$ as the load of the elastic class i. In the same way, we give by $\rho_j^{(s)}$ the load of the streaming class j, where $\rho_j^{(s)} = \lambda_j^{(s)}\tau_j(Flows)$.
Let $x_j^{(s)}$,$j \in S$, (respectively $x_i^{(e)}$,$i \in E$) be the number of class-j flows in progress (respectively the number of class-i flows in progress) . We refer to the vector $x^{(s)} = (x_j^{(s)})_{j\in S}$ (respectively $x^{(e)} = (x_i^{(e)})_{i\in E}$) as the state of real-time flow-classes (respectively the state of elastic classes). Let $d^{(s)} = (d_j^{(s)})_{j\in S}$ and $d^{(e)} = (d_i^{(e)})_{i\in E}$) be the vectors of rate limits for streaming flow-classes and elastic flow-classes, respectively.
We note by $\theta^{(e)} = \sum_{i\in E}\rho_i^{(e)}$ (respectively $\theta^{(s)} = \sum_{j\in S}\rho_j^{(s)}d_j^{(s)}$) as the elastic load (respectively the real-time load) offered to the capacity C in $Mbits/Second$. In order to maintain the stability of the system, we assume that the total load is strictly inferior to the link capacity:

$$\theta = \theta^{(e)} + \theta^{(s)} < C \tag{1}$$

In a similar way to the configuration of Internet routers, we assume that, for both two directions, packets are affected to a LLQ queue before being sent into the link. This LLQ queue combines a priority queue with a number of M CBWFQ queues. Let w_m, $1 \le m \le M$, be the weight of the queue number m. We assume that:

$$\sum_{m=1}^{M} w_m = 1 \tag{2}$$

.
The priority queue is devoted to real-time flows, which have strict bandwidth and delay requirements that can be met only if the link capacity is completely allocated to them. Real-time flows whose requirements cannot be guaranteed will be blocked rather than allow them into the system and jeopardize the performance of ongoing real time traffic. This admission control coupled to the strict priority is generally considered sufficient to meet the quality of service requirements of the audio and video applications. Elastic traffic is distributed throughout the CBWFQ queues.
In a large scale modelization, each CBWFQ queue can be represented by a virtual link. We note by E_m the set of elastic flow classes passing through the virtual link number m and by $\theta_m^{(e)} = \sum_{i\in E_m}\rho_i^{(e)}$ the total elastic load offered to this virtual link. In the following, the virtual link m refer to the CBWFQ queue m and vice versa.
Since the above resource sharing is applied for both upload and download directions, our study will be limited to one

LLQ system. For elastic traffic, we assume that users perceive performance essentially through the average throughput per flow [7], [28], [29]. For each class $i \in E$, the average throughput, noted γ_i, is given by:

$$\gamma_i = \frac{\sigma_i}{\tau_i^{(e)}} \tag{3}$$

Where $\tau_i^{(e)}$ is the mean delay of the class-i flows. By Little's low, we have:

$$\tau_i^{(e)} = \frac{E[x_i^{(e)}]}{\lambda_i^{(e)}} \tag{4}$$

Where $E[x_i^{(e)}]$ is the average number of active flows for the elastic class i.
From 3 and 4 , we deduce:

$$\gamma_i = \frac{\rho_i^{(e)}}{E[x_i^{(e)}]} \tag{5}$$

The expression 5 is widely used in the literature [20], [28], [30], [31]. Evaluating the average throughput for each class of flows in a large scale modelization necessarily pass then through evaluating the average number of persistent flows for this class.
In the following, we use these notations:

$$|x^{(e)}| = \sum_{i \in E} x_i^{(e)}$$

$$x^{(e)} d^{(e)} = \sum_{i \in E} x_j^{(e)} d_j^{(e)}$$

The same previous notations are applicable for $x^{(s)}$ and $d^{(s)}$. We note also by $\pi(B)$ the stationary probability of a set B and by $\pi(x)$ the stationary probability of a vector x.

IV. Approximate results for balanced fairness allocation

In this section, we assume that the link capacity is only exploited by elastic traffic and there is no CBWFQ policy for this traffic.
The dynamic flow level modelling of Internet traffic in Best Effort architecture was practically introduced with Massouliè and Roberts with a model of fair bandwidth sharing (Processor-sharing Model) [32]. Theoretical allocations optimising an utility function of the instantaneous flow rates have also been proposed to estimate the TCP bandwidth sharing. Examples of such allocations are the classical max-min fairness and the Kelly's proportional fairness [29]. In general, the analysis of a network operating under these allocations scheme is quite difficult. One reason is that they do not lead to an explicit expression for the steady state distribution, which determines the typical number of competing flows of each class [30]. For this reason, balanced fairness was proposed by Bonald and Proutière to approximate the performance of TCP traffic under these equitable allocations [28], [31], [33]. A key property of balanced fairness is its insensitivity: the steady state distribution is independent of all traffic characteristics beyond the traffic intensity [30], [28], [31]. The only required assumption is that flows arrive as a Poisson process. The performance metrics of TCP traffic can be then evaluated for each class of flows. Nevertheless, the balanced fairness allocation remains complex to be used in a practical context as it requires the calculation of the probability of all possible states of the system, and thus it faces the combinatorial explosion of the space of states for large networks [21], [29], [34]. Recent works led by Jean Marie Garcia, Research Director at LAAS-CNRS and co-author of this paper, proposed some explicit approximations to effectively calculate the performance metrics of elastic traffic under this allocation without requiring the evaluation of individual probability of states [20], [34]. These approximations are based on numerical observations and are practically applicable for all network topologies. This section is devoted to present these approximations, which will be exploited in the next section to evaluate the performance of our studied LLQ system.

A. Identical rate limits

We suppose here that all elastic classes have the same maximum bit rate r. Let $N = \lfloor C/r \rfloor$ be the maximum number of flows that can have its maximum bit rate. In [20], Henda proved that, using balanced fairness allocation, the average number of flows for each class i is exactly given by:

$$E[x_i^{(e)}] = \frac{\rho_i^{(e)}}{r} + \pi(B)\frac{\rho_i^{(e)}}{C - \theta^{(e)}} \tag{6}$$

Where B is the set of congestion stats:

$$\pi(B) = \pi(|x^{(e)}| \geq N) = \frac{(\frac{\theta}{r})^N}{N!}\frac{C}{C - \theta^{(e)}}\pi(0) \tag{7}$$

$\pi(0)$ is the probability that the link is empty, and it is given by:

$$\pi(0) = \left(\sum_{k=0}^{N-1} \frac{(\frac{\theta^{(e)}}{r})^k}{k!} + \frac{(\frac{\theta^{(e)}}{r})^N}{N!}\frac{C}{C - \theta^{(e)}} \right)^{-1} \tag{8}$$

The average total number of flows can be then deduced from 6 by:

$$E[x^{(e)}] = \frac{\theta^{(e)}}{r} + \pi(B)\frac{\theta^{(e)}}{C - \theta^{(e)}} \tag{9}$$

And then, we have the following relation:

$$E[x_i^{(e)}] = \frac{\rho_i^{(e)}}{\theta^{(e)}} E[x^{(e)}] \tag{10}$$

B. General rate limits

For different transmission rates, an approximation was given in [20] and [21]:

$$E[x_i^{(e)}] \approx \frac{\rho_i^{(e)}}{d_i^{(e)}} + \pi(B_i)\frac{\rho_i^{(e)}}{C - \theta^{(e)}} \tag{11}$$

Where B_i is the set of congestion stats for the class i: $B_i = \{x^{(e)}, x^{(e)}d^{(e)} >= C - d_i^{(e)}\}$. The probability $\pi(B_i)$ is written as follows:

$$\pi(B_i) = \frac{1}{C - \theta^{(e)}} \sum_{k \in E} \rho_k^{(e)} \pi(W_k) + \pi(W_i) \tag{12}$$

W_k is defined as $W_k = \{x^{(e)}, C^{(e)} - d_k \le x^{(e)} d^{(e)} < C\}$ and its probability is given by:

$$\pi(W_k) = \pi(0) \sum_{x^{(e)} \in W_j} P(x^{(e)} d^{(e)}) \qquad (13)$$

Where $P(x^{(e)} d^{(e)})$ is given by the Kaufman-Roberts formula. This formula is written as follows [35]:

$$P(n) = \sum_{k \in E} \frac{\rho_k^{(e)}}{n} P(n - d_k^{(e)}) \qquad \forall 1 \le n \le C \qquad (14)$$

with $P(0) = 1$ and $P(n) = 0$ for all $n < 0$.

The probability that the link is empty is then given by:

$$\pi(0) = \Bigg(\sum_{0 \le x^{(e)} d^{(e)} < C} P(x^{(e)} d^{(e)}) + \frac{1}{C - \theta^{(e)}} \sum_{k \in E} \rho_k^{(e)} \sum_{x \in W_k} P(x^{(e)} d^{(e)}) \Bigg)^{-1} \qquad (15)$$

V. Analysis of the LLQ system

A. High Priority traffic

If the capacity C is already occupied by high priority traffic then further incoming real-time flows are lost. For each class $j \in S$, we note by $B_j = \{x^{(s)}, C - d_j^{(s)} < x^{(s)} d^{(s)} \le C\}$ the set of blocking states. The probability of this set is given by:

$$\pi(B_j) = \sum_{C - d_j^{(s)} < x^{(s)} d^{(s)} \le C} \pi(x^{(s)}) \qquad (16)$$

For each state $x^{(s)}$, the probability $\pi(x^{(s)})$ is given by:

$$\pi(x^{(s)}) = \prod_{j \in S} \frac{\rho_j^{(s)\, x_j^{(s)}}}{x_j^{(s)}!} \pi(0)^{(s)} \qquad (17)$$

Where:

$$\pi(0)^{(s)} = \Bigg(\sum_{0 \le x^{(s)} d^{(s)} \le C} \frac{\rho_j^{(s)\, x_j^{(s)}}}{x_j^{(s)}!} \Bigg)^{-1} \qquad (18)$$

Kaufman and Roberts proposed a more efficient solution than the direct calculation of the product form of the blocking probability [36], [37]. After splitting the link capacity into virtual circuits, their solution based on a recursive algorithm to determine the probability that m circuits are occupied.

Let δd the greatest common divisor between $d_j^{(s)}$ $\forall j \in S$. We have then:

- $K = C/\delta d$: The maximum number of circuits.
- $m_j = d_j^{(s)}/\delta d$: The number of circuits required for one flow of class j.

According to Kaufman and Roberts, the state probabilities are calculated by the following recursive algorithm:

$$\pi^*(m) = \begin{cases} 1 & \text{if} \quad m = 0 \\ 0 & \text{if} \quad m < 0 \\ \frac{1}{m} \sum_{j \in S} \pi^*(m - m_j) m_j \rho_j^{(s)} & \text{else} \end{cases} \qquad (19)$$

$$\pi(m) = \pi^*(m) \Bigg(\sum_{1 \le k \le K} \pi^*(k) \Bigg)^{-1}$$

The blocking probability for each class j is obtained by:

$$\pi(B_j) = \sum_{K - m_j + 1 \le m \le K} \pi(m) \qquad (20)$$

The average number of flows for each class j is written as follows:

$$E[x_j^{(s)}] = \rho_j^{(s)} (1 - \pi(B_j)) \qquad (21)$$

Thus, the average total number of flows for real-time traffic is given by:

$$E^{(s)} = \sum_{j \in S} E[x_j^{(s)}] \qquad (22)$$

B. Evaluation of the average total number of flows

Since each lost TCP packet is sent again until it reaches the destination, looking at flow-level, it is assumed that the total elastic traffic load is conserved in the system (there is no loss). Let $\theta^{(mix)}(Mbits/Second)$ the average total load carried by the system. $\theta^{(mix)}$ can be written as follows:

$$\theta^{(mix)} = \sum_{j \in S} E[x_j^{(s)}] d_j^{(s)} + \theta^{(e)} \qquad (23)$$

For a Best Effort system, the link capacity will be equally shared to carry the total traffic load $\theta^{(mix)}$. For a LLQ system, the same capacity will be shared in a differentiated manner to carry the same total traffic load. This leads to the fact that the mean total number of flows is approximately conserved between the two systems. Let $E_{\theta^{(mix)}}$ the average total number of flows carried by the system. We define a new set of elastic flow classes noted $E^{'}$ with the same cardinality as S such that for each $k \in E^{'}$, $\rho_k^{(e)} = E[x_k^{(s)}] d_k^{(s)}$ and $d_k^{(e)} = d_k^{(s)}$. $E_{\theta^{(mix)}}$ is given by:

$$E_{\theta^{(mix)}} = \sum_{i \in E \cup E'} E_{\theta^{(mix)}}[x_i^{(e)}] \qquad (24)$$

Where $E_{\theta^{(mix)}}[x_i^{(e)}]$ is deduced from 11 by replacing $\theta^{(e)}$ with $\theta^{(mix)}$.

The average total number of elastic flows can be then approximated by:

$$E_{\theta^{(mix)}}[x^{(e)}] \approx E_{\theta^{(mix)}} - \sum_{j \in S} E[x_j^{(s)}] \qquad (25)$$

C. Evaluating the average total number of flows for each CBWFQ queue: General case

In this section we assume that all elastic flow-classes in E have the same maximum bit rate r.

As we showed in [38] for a system of two CBWFQ queues carrying traffic with identical maximum bit rate, the evolution of the average total number of flows for each queue m in function of its weight can be estimated by the expression 26, where β_1, β_2 and β_3 are calculated according the system parameters and α is numerically adjusted to 3. Our numerical observations show that this assumption remains true even for a greater number of queues.

$$E_m^{CBWFQ} \approx \frac{\beta_1}{\beta_2 + w_m^{\alpha}} + \beta_3 \qquad (26)$$

The parameters β_1, β_2 and β_3 will be fixed basing on known results for three specific cases for w_m: w_m tends towards 1, w_m is equal to $1/M$ and w_m tends towards 0.

When the weight of a queue m, tends towards 1, it can be considered as a priority queue [39]. Let $E^{CBWFQ}_{m/w_m \to 1}$ the average number of flows passing through this queue in this case. $E^{CBWFQ}_{m/w_m \to 1}$ can be approximated by:

$$E^{CBWFQ}_{m/w_m \to 1} \approx E_{\theta^{(mix)} - \sum_{\substack{1 \le k \le M \\ k \ne m}} \theta^{(e)}_k} - \sum_{j \in S} E[x^{(s)}_j] \quad (27)$$

Where $E_{\theta^{(mix)} - \sum_{\substack{1 \le k \le M \\ k \ne m}} \theta^{(e)}_k}$ refers to the average total number of flows for a system with two priority queues. The first priority queue is dedicated for real-time traffic while the second is devoted to elastic flow-classes in E_m. $E_{\theta^{(mix)} - \sum_{\substack{1 \le k \le M \\ k \ne m}} \theta^{(e)}_k}$ is given by:

$$E_{\theta^{(mix)} - \sum_{\substack{1 \le k \le M \\ k \ne m}} \theta^{(e)}_k} = \sum_{i \in E_m \cup E'} E_{\theta^{(mix)} - \sum_{\substack{1 \le k \le M \\ k \ne m}} \theta^{(e)}_k} [x^{(e)}_i] \quad (28)$$

Where $E_{\theta^{(mix)} - \sum_{\substack{1 \le k \le M \\ k \ne m}} \theta^{(e)}_k} [x^{(e)}_i]$ is deduced from 11 by replacing $\theta^{(e)}$ with $\theta^{(mix)} - \sum_{\substack{1 \le k \le M \\ k \ne m}} \theta^{(e)}_k$.

Exploiting 26, $E^{CBWFQ}_{m/w_m \to 1}$ can be written as follows:

$$E^{CBWFQ}_{m/w_m \to 1} \approx \frac{\beta_1}{\beta_2 + 1} + \beta_3 \quad (29)$$

When the weight of each queue m is equal to $1/M$, the remaining capacity for elastic traffic is supposed to be equally shared among the queues, and we are practically in the best effort case. Given that all elastic flows have identical rate limits and using 10, we have:

$$E^{CBWFQ}_{m/w_m = 1/M} \approx \frac{\theta^{(e)}_m}{\theta^{(e)}} E_{\theta^{(mix)}}[x^{(e)}] \quad (30)$$

Where $E_{\theta^{(mix)}}[x^{(e)}]$ is given by 25. Using 26 we obtain:

$$E^{CBWFQ}_{m/w_m = 1/M} \approx \frac{\beta_1}{\beta_2 + (1/M)^{\alpha}} + \beta_3 \quad (31)$$

When the weight of a queue m tends towards zero, the average number of flows passing through this queue is obtained by:

$$E^{CBWFQ}_{m/w_m \to 0} \approx E_{\theta^{(mix)}}[x^{(e)}] - E_{\theta^{(mix)} - \theta^{(e)}_m}[x^{(e)}] \quad (32)$$

Where $E_{\theta^{(mix)} - \theta^{(e)}_m}[x^{(e)}]$ refers to the average total number of elastic flows passing through the system regardless of the traffic of the CBWFQ queue number m. It is deduced from 25 as follows:

$$E_{\theta^{(mix)} - \theta^{(e)}_m}[x^{(e)}] \approx E_{\theta^{(mix)} - \theta^{(e)}_m} - \sum_{j \in S} E[x^{(s)}_j] \quad (33)$$

Where $E_{\theta^{(mix)} - \theta^{(e)}_m}$ is given by:

$$E_{\theta^{(mix)} - \theta^{(e)}_m} = \sum_{i \in (E/E_m) \cup E'} E_{\theta^{(mix)} - \theta^{(e)}_m}[x^{(e)}_i] \quad (34)$$

$E_{\theta^{(mix)} - \theta^{(e)}_m}[x^{(e)}_i]$ is deduced from 11 by replacing $\theta^{(e)}$ with $\theta^{(mix)} - \theta^{(e)}_m$.

According to 26 ,$E^{CBWFQ}_{m/w_m \to 0}$ can be written as follows:

$$E^{CBWFQ}_{m/w_m \to 0} \approx \frac{\beta_1}{\beta_2} + \beta_3 \quad (35)$$

The average total number of active flows in each queue can be then approximated by 26 where β_1, β_2 and β_3 are given from 29, 31 and 35 as follows:

$$\beta_1 \approx \left(E^{CBWFQ}_{m/w_m = 1/M} - \beta_3\right)\left(\left(\frac{1}{M}\right)^{\alpha} + \beta_2\right) \quad (36)$$

$$\beta_2 \approx \frac{\delta_1}{(M^{\alpha} - 1)\delta_2 - \delta_1} \quad (37)$$

$$\beta_3 \approx E^{CBWFQ}_{m/w_m = 1/M} - \frac{1 + \beta_2}{1 - (1/M)^{\alpha}} \delta_1 \quad (38)$$

With:

$$\delta_1 = E^{CBWFQ}_{m/w_m = 1/M} - E^{CBWFQ}_{m/w_m \to 1} \quad (39)$$

And:

$$\delta_2 = E^{CBWFQ}_{m/w_m \to 0} - E^{CBWFQ}_{m/w_m = 1/M} \quad (40)$$

Since all flows passing through a queue m are supposed to have the same maximum bit rate, the average number of flows for each class $i \in E_m$ can be deduced from 10 as follows:

$$E[x^{(e)}_i] = \frac{\rho^{(e)}_i}{\theta^{(e)}_m} E^{CBWFQ}_m \quad (41)$$

VI. Simulations and validity of analytical results

In this section, we aim to compare our analytical analysis with the real behaviour of TCP traffic. We simulate then with NS 2 the case of a capacity link $C = 100(Mbits/Second)$ shared by two streaming flow classes (transporting by User Datagram Protocol (UDP)) and three TCP flow classes. We assume then that at the entry of the link there is a LLQ queue that combines a priority queue with three CBWFQ queues and each CBWFQ queue is dedicated to a single class of TCP flows. Let $w_1 = 0.7$, $w_2 = 0.25$ and $w_3 = 0.05$.

Let $d^{(s)}_1 = 10(Mbits/Second)$, $d^{(s)}_2 = 5(Mbits/Second)$ and $r = 3(Mbits/Second)$. We assume that UDP traffic constitutes 60% of the total system load which varies from 50% to 80% of the link capacity. The two streaming classes have the same load. In the same manner, we assume that all elastic flow classes have the same contribution in the total elastic load. For our simulation, we took $\lambda^{(s)}_j = \lambda^{(e)}_i = 1$ $(\forall i \in E, \forall j \in S)$.

Each TCP flow is used to transfer a series of packets of 1000 bytes representing a document of a certain size. We chose FTP (File Transfer Protocol) and CBR (Constant Bit Rate) to be implemented respectively for TCP and UDP traffic. To ensure compatibility with our theoretical analysis, we assumed that there is no additional time added by the link.

Figure 2 plots a comparison between analytical and simulations results in term of the average throughput per queue for different ratios of θ/C. At first, it is necessary to note that

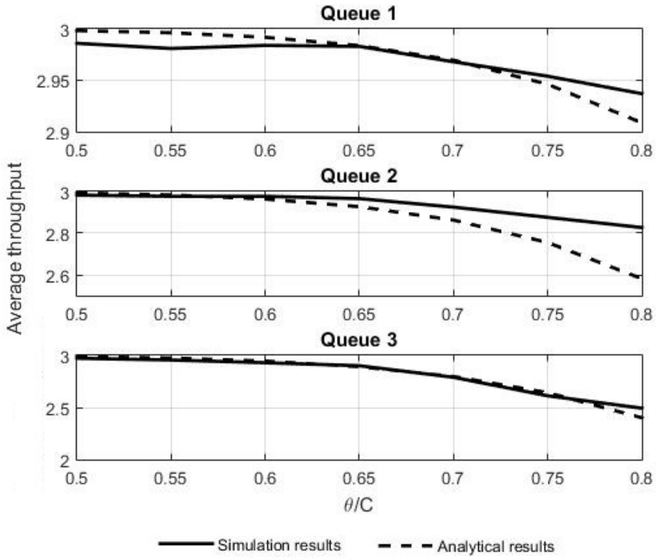

Figure 2. Comparison in term of the average throughput per queue between analytical and simulations results

the proposed analytical method is more accurate when the total load system is less than 75% with a negligible error rate that doesn't globally exceed the 4%. When the system tends to the instability zone, our analytical approach keep its accuracy for the first and the third queue, in contrast to the second queue where the error rate increases to 8%. In fact, for high and very low weights, our numerical approximation is more accurate since it is based on the explicit mathematical expressions 29 and 35 .

The effect of the coupling between CBWFQ queues is especially clear on the third queue. Although that the elastic load of this queue is always superior to its minimum capacity (w_3C), flows crossing this queue have a good average throughput when the system is far from the instability zone and reasonable throughput when the system tends to the instability. This is explained by the fact that the third queue exploited the remaining capacity not used by the other queues. It is clear that our approximation captures very well this coupling.

For a total load equal to 80% of the capacity, the probability that the elastic traffic becomes unstable(the probability that the remaining capacity for elastic traffic is inferior or equal to the total elastic load) is 0.15 which is not negligible. Therefore, the quasi-stationary analysis , largely adopted in the literature, can't give a reasonable results since it bases on calculating the average number of elastic flows for each state of streaming flows. Our approach overcomes this limitation by analysing the whole system as one entity without separating between the elastic and real-time traffic.

VII. CONCLUSION

In this paper, we propose a new fluid approach to evaluate the performance of elastic approach under a LLQ system giving the head of priority to delay-sensitive applications. Our analysis focus on the total average flows carried by the whole system by approximating the system as a lossless Best Effort system with total load equal to the total carried load of the original system (no blocked load). Approximations issued from balanced fairness, which is considered by many authors as an efficient tool for estimating bandwidth sharing among elastic traffic for best effort allocations, are used to approximate the average total number of flows in this equivalent system and deduce the performance of CBWFQ system. Detailed packet level simulations proved the accuracy and the good behaviour of our analysis.

Our approach is original in the sense that it overcomes the limitations of previous works studying the integration of elastic and real-time traffic under classical approaches (quasi-stationary approach, calculating the average remaining capacity for elastic traffic) and assuming that all elastic traffic have the same requirement in quality of service. The problem that we studied reflects the reality (and the complexity) of the Internet multimedia processes with heterogeneous flows, differentiated classes of services and different transport protocols. The expression given to evaluate the average end-to-end throughput of elastic traffic under a multi-queuing system is precise and can capture the strong coupling between the queues with a reasonable computation time.

Another key result is that the approximation proposed is insensitive to detailed traffic characteristics since it is based on approximations issued from balanced fairness allocations. This is particularly important for data network engineering since performance can be predicted from an estimate of overall traffic volume alone and is independent of changes in the mix of user applications. We expect results such as those presented in this paper to eventually lead to simple and robust traffic engineering rules and performance evaluation methods that are lacking for actual data networks.

REFERENCES

[1] V. Foteinos, K. Tsagkaris, P. Peloso, L. Ciavaglia, and P. Demestichas, "Energy-aware traffic allocation to optical lightpaths in multilayer core networks," *International Journal of Network Management*, vol. 25, no. 6, pp. 394–417, 2015.

[2] T. Jursonovics and S. Imre, "Quality-based charging solutions for wireless multimedia services," *International Journal of Network Management*, vol. 24, no. 5, pp. 357–401, 2014.

[3] K. Shim, J. Ham, S. BD, and M. S. Kim, "Application traffic classification using payload size sequence signature," *International Journal of Network Management*, vol. 27, no. 5, 2017.

[4] Y. Bian, B. Liu, Y. Li, and J. Gao, "Characterizing network traffic behaviour using granule based association rule mining," *International Journal of Network Management*, vol. 26, no. 4, pp. 308–329, 2016.

[5] Networking Index and Cisco Visual, "Forecast and methodology, 2016-2021, white paper," *San Jose, CA, USA*, vol. 1, 2016.

[6] R. Malhotra and J. Van Den Berg, "Flow level performance approximations for elastic traffic integrated with prioritized stream traffic," in *Proceedings of the 12 th International Telecommunications Network Strategy and Planning Symposium, NETWORKS 2006*. New Delhi, India: IEEE, November 6–9, 2006, pp. 1–9.

[7] T. Bonald, A. Proutiere, J. Roberts, and J. Virtamo, "Computational aspects of balanced fairness," *Teletraffic Science and Engineering*, vol. 5, pp. 801–810, 2003.

[8] A. Rao, A. Legout, Y. S. Lim, D. Towsley, C. Barakat, and W. Dabbous, "Network characteristics of video streaming traffic," in *The Seventh COnference on emerging Networking EXperiments and Technologies*, New York, USA, December 6–9, 2011.

[9] E. Brosh, S. A. Baset, V. Misra, D. Rubenstein, and H. Schulzrinne, “The delay-friendliness of tcp for realtime traffic,” *IEEE/ACM Transactions On Networking*, vol. 18, no. 5, pp. 1478–1491, 2010.

[10] J. Martin, Y. Fu, N. Wourms, and T. Shaw, “Characterizing netflix bandwidth consumption,” in *Proceeding of the 10th Consumer Communications and Networking Conference (CCNC)*. Las Vegas, AA, USA: IEEE, January 11-14,2013, pp. 230–235.

[11] B. Wang, J. Kurose, P. Shenoy, and D. Towsley, “Multimedia streaming via tcp: An analytic performance study,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 4, no. 2, p. 16, 2008.

[12] A. Biernacki and K. Tutschku, “Performance of http video streaming under different network conditions,” *Multimedia Tools and Applications*, vol. 72, no. 2, pp. 1143–1166, 2014.

[13] T. Balogh and M. Medvecky, “Performance evaluation of wfq, wf2q+ and wrr queue scheduling algorithms,” in *Proceeding of the 34th International Conference on Telecommunications and Signal Processing*, Budapest, Hungary, August 18–20, 2011, pp. 136–140.

[14] T. Balogh and M. Medveckỳ, “Average bandwidth allocation model of wfq,” *Modelling and Simulation in Engineering*, vol. 2012, p. 39, 2012.

[15] M. J. Fischer, D. M. B. Masi, and J. F. Shortle, “Simulating the performance of a class-based weighted fair queueing system,” in *Proceedings of the 40th Conference on Winter Simulation*, Miami, FL, USA, December 07–10, 2008, pp. 2901–2908.

[16] R. B. Bahaweres, A. Fauzi, and M. Alaydrus, “Comparative analysis of llq traffic scheduler to fifo and cbwfq on ip phone-based applications (voip) using opnet (riverbed),” in *Proceedings of the 1st International Conference on Wireless and Telematics (ICWT 2015)*. Manado,Indonesia: IEEE, November 17-18, 2015, pp. 1–5.

[17] “Congestion avoidance and congestion management configuration,” http://support.huawei.com/enterprise/en/doc/DOC1000088754?section=j009, accessed: 2018-01-29.

[18] S. Szilágyi and B. Almási, “A review of congestion management algorithms on cisco routers,” *Journal of Computer Science and Control Systems*, vol. 5, no. 1, p. 103, 2012.

[19] B. Dalmazo, J. Vilela, and M. Curado, “Online traffic prediction in the cloud,” *International Journal of Network Management*, vol. 26, no. 4, pp. 269–285, 2016.

[20] H. B. Cheikh, “Evaluation et optimisation de la performance des flots dans les réseaux stochastiques à partage de bande passante,” Ph.D. dissertation, INSA de Toulouse, 2015.

[21] H. Ben Cheikh, “Integration of streaming and elastic traffic: Modeling and performance analysis,” in *Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools*. Berlin, Germany: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), December 14–16, 2015, pp. 65–68.

[22] O. J. Boxma, A. F. Gabor, R. Núnez-Queija, and H. P. Tan, “Integration of streaming and elastic traffic in a single umts cell: modeling and performance analysis,” *EURANDOM Report*, 2006.

[23] N. Benameur, S. B. Fredj, F. Delcoigne, S. Oueslati-Boulahia, and J. W. Roberts, “Integrated admission control for streaming and elastic traffic,” in *Proceedings of the International Workshop on Quality of Future Internet Services*. Coimbra, Portugal: Springer, September 24-26, 2001, pp. 69–81.

[24] S. Borst and N. Hegde, “Integration of streaming and elastic traffic in wireless networks,” in *Proceedings of the 26th International Conference on Computer Communications INFOCOM 2007. IEEE*. Anchorage, Alaska, USA: IEEE, May 6–12, 2007, pp. 1884–1892.

[25] T. Bonald and A. Proutière, “On performance bounds for the integration of elastic and adaptive streaming flows,” in *ACM SIGMETRICS Performance Evaluation Review*, vol. 32, no. 1. ACM, 2004, pp. 235–245.

[26] F. Delcoigne, A. Proutiere, and G. Régnié, “Modeling integration of streaming and data traffic,” *Performance Evaluation*, vol. 55, no. 3, pp. 185–209, 2004.

[27] P. Key, L. Massoulié, A. Bain, and F. Kelly, “Fair internet traffic integration: network flow models and analysis,” *Annals of Telecommunications*, vol. 59, no. 11, pp. 1338–1352, 2004.

[28] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo, “A queueing analysis of max-min fairness, proportional fairness and balanced fairness,” *Queueing systems*, vol. 53, no. 1, pp. 65–84, 2006.

[29] T. Bonald and J. Virtamo, “Calculating the flow level performance of balanced fairness in tree networks,” *Performance Evaluation*, vol. 58, no. 1, pp. 1–14, 2004.

[30] T. Bonald, J. P. Haddad, and R. R. Mazumdar, “Congestion in large balanced multirate links,” in *Proceedings of the 23rd International Teletraffic Congress*, San Francisco, USA, September 6–8, 2011, pp. 182–189.

[31] T. Bonald and A. Proutiere, “Insensitive bandwidth sharing in data networks,” *Queueing systems*, vol. 44, no. 1, pp. 69–100, 2003.

[32] F. P. Kelly and R. J. Williams, “Fluid model for a network operating under a fair bandwidth-sharing policy,” *The Annals of Applied Probability*, vol. 14, no. 3, pp. 1055–1083, 2004.

[33] L. Massoulié, “Structural properties of proportional fairness: stability and insensitivity,” *The Annals of Applied Probability*, pp. 809–839, 2007.

[34] J. M. Garcia and M. E. H. Boussada, “Evaluation des performances bout en bout du trafic tcp sous le régime ≪équité équilibrée≫,” in *The 11th Performance Evaluation Workshop*, Toulouse,France, March 15–17, 2016.

[35] T. Bonald and J. Virtamo, “A recursive formula for multirate systems with elastic traffic,” *IEEE Communications Letters*, vol. 9, no. 8, pp. 753–755, 2005.

[36] J. Kaufman, “Blocking in a shared resource environment,” *IEEE Transactions on communications*, vol. 29, no. 10, pp. 1474–1481, 1981.

[37] J. Roberts, “A service system with heterogeneous user requirements: Application to multi-service telecommunications systems,” *Proc. Performance of Data Communications Systems and their Applications, North Holland, Amsterdam, 1981*, pp. 423–431, 1981.

[38] M. E. H. Boussada, J. M. Garcia, and M. Frikha, “Numerical key results for studying the performance of tcp traffic under class-based weighted fair queuing system,” *Journal of Communication and Computer*, vol. 13, pp. 195–201, 2016.

[39] C. Bockstal, J. M. Garcia, and O. Brun, “Approximation du régime stationnaire d’un système wfq,” in *Proceeding of the 6th French speaking meetings on the algorithmic aspects of telecommunication*, Batz-sur-Mer, France, May 26-28,2004.