Int. J. Communications, Network and System Sciences, 2016, 9, 219-233
Published Online June 2016 in SciRes. http://www.scirp.org/journal/ijcns
http://dx.doi.org/10.4236/ijcns.2016.96021

End-to-End Performance Evaluation of TCP Traffic under Multi-Queuing Networks

Jean Marie Garcia[1], Mohamed El Hedi Boussada[2]

[1]Services and Architectures for Advanced Networks, LAAS-CNRS, University of Toulouse, CNRS, Toulouse, France
[2]Mobile Network and Multimedia, SUP'COM, Ariana, Tunisia
Email: jmg@laas.fr, med.elhadi.boussada@supcom.tn

Received 13 March 2016; accepted 21 June 2016; published 24 June 2016

Abstract

While Internet traffic is currently dominated by elastic data transfers, it is anticipated that streaming applications will rapidly develop and contribute a significant amount of traffic in the near future. Therefore, it is essential to understand and capture the relation between streaming and elastic traffic behavior. In this paper, we focus on developing simple yet effective approximations to capture this relationship. We study, then, an analytical model to evaluate the end-to-end performance of elastic traffic under multi-queuing system. This model is based on the fluid flow approximation. We assume that network architecture gives the head of priority to real time traffic and shares the remaining capacity between the elastic ongoing flows according to a specific weight.

Keywords

Flow-Level Modelling, Multi-Queuing Network, Quality of Service, Streaming Traffic, Elastic Traffic

1. Introduction

The expansion of mobile communications, the increase of access rates and the convergence of access technologies (which allowed users to access at the same services regardless of the terminal used and where they are) lead to a multiplication of services offered by networks and to an unprecedented growth in the number of users and traffic volumes that they generate.

In addition to its traditional services, there has been interest in supporting real-time communication applications in the packet-based environments. Therefore, we shall distinguish two broad categories of Internet traffic: stream and elastic [1]. Streaming traffic is generated by applications such as Voice over Internet protocol appli-

How to cite this paper: Garcia, J.M. and Boussada, M.E.H. (2016) End-to-End Performance Evaluation of TCP Traffic under Multi-Queuing Networks. *Int. J. Communications, Network and System Sciences*, **9**, 219-233.
http://dx.doi.org/10.4236/ijcns.2016.96021

cations (VoIP applications), streaming video, interactive voice, online gaming, and videoconference applications. These applications have strict bandwidth, end-to-end packet delay and jitter requirements for reliable operation. Elastic traffic on the other hand is generated by applications such as file transfer, web-browsing, etc. Since these applications rely on the Transport Control Protocol (TCP) for packet transmission, the traffic generated is elastic in nature. This is because TCP's congestion control adapts to the available capacity in the network (congestion avoidance and slow start adaptive mechanism) and results in an elastic packet transmission rate [2].

To support both streaming and elastic traffic types, the network's architecture has been evolved beyond the best-effort model. The Diffserv architecture goes towards meeting the distinct quality of service requirements of these two types of traffic [3]. Many studies have been done to perform service's differentiation. Today, several scheduling algorithms are implemented to achieve this process, and are classified into two categories: fixed priority policies and bandwidth sharing-based policies like Weighted Round Robin (WRR) and Weighted Fair Queuing (WFQ) [1]. The composition between the two policies is considered by many telecommunications equipment constructors like Cisco [4] and Huawei [5]. The low-latency queuing (LLQ), for example, is a feature developed by Cisco to bring strict priority queuing (PQ) to class-based weighted fair queuing (CBWFQ) [4].

Today, with more than two billion Internet users worldwide [6]-[8], the information and communication technologies are increasingly present in our daily activities. In this context, the interruption of services provided by networks, or even a significant degradation of quality of service, is becoming less and less tolerable. Ensuring the continuity and quality of services is thus a major challenge for network operators.

For operators, the solution is to have a more regular monitoring of their infrastructures and to use traffic engineering techniques to anticipate the degradation of quality of service resulting from the phenomena of congestion. The use of these techniques, however, assumes to have models, theoretical methods and appropriate software tools to predict and control the quality of service of traffic flows.

In the literature, we can basically distinguish two types of models: the packet level models and flow level models. The packet level defines the way in which packets are generated and transported during the communication [9]. The packet level models incorporate many details about the system (Round Trip Times, buffer size, etc.) but generally consider a fixed number of persistent flows [10]. Although these models may be relevant to calculate packet level performance metrics (loss rate or transmission delay for example), they don't consider the dynamic flow-level (the arrival of flows at random times and random amounts of data to be transmitted).

Flow-level models, are an idealized models that include random flow-level dynamics (arrivals and departures of flows) and use highly simplified models of the bandwidth sharing [10]. The complex underlying packet-level mechanisms (congestion control algorithms, packet scheduling, buffer management…), at short-time scales, are then simply represented by a long-term bandwidth sharing policy between ongoing flows [1].

In general, a flow is defined as a series of packets between a source and a destination having the same transport protocol number and port number [11]. In flow level modelling, a flow is seen like an end-to-end connection between two entities whose rate varies dynamically in each arrival or departure of another flow. We refer to class of flows as all flows of the same service between a source and a destination, having a common rate limitation and the same resources requirements.

This paper presents a fluid model to evaluate and qualify performance characteristics of elastic traffic under multi-queuing architecture. In the next section, we present useful results applying to a network whose resources are dedicated for elastic traffic only. Section 3 is devoted to present our analytical model able to evaluate the performance of elastic traffic merging with streaming flows. The results presented in this manuscript are validated by simulations with NS2 in Section 4.

2. Bandwidth Sharing with Elastic Traffic

The network consists of a set of links $\mathcal{L}$ where each link l has a capacity C_l. A random number of elastic flows compete for the bandwidth of these links. Let E be the set of elastic flow classes. Elastic flows are generally characterized by their maximum bit rate and the mean size of the file that is transferred. For each elastic class-*i* flows ($i \in E$), we define:

- σ_i: The mean volume transferred by flows.
- $d_i^{(e)}$: The maximum bit rate of each flow.

Flows arrive as an independent Poisson process with rate $\lambda_i^{(e)}$ for class-*i* flows. We refer to the product $\rho_i^{(e)} = \lambda_i^{(e)}\sigma_i$ as the load of elastic class *i*. Let $\mathcal{A}$ be the incidence matrix defined as follows: $a_{i,l} = 1$ if class-*i* flows go through link l and it equals to zero otherwise.

Let $\theta_l^{(e)} = \sum_{i\in E} a_{i,l}\rho_i^{(e)}$ be the elastic load offered to a link l. To maintain the stability of the system, we assume that the total load of each link $l \in \mathcal{L}$ is strictly inferior to its capacity: $\theta_l^{(e)} < C_l$.

$d_i^{(e)}$ is the actual rate of each flow in the absence of congestion. Congestion forces elastic flows to reduce their rate and thus to increase their duration. We note $x_i^{(e)}$ the number of class-*i* flows and we refer to the vector $x^e = \left(x_i^{(e)}\right)_{i\in E}$ as the network state. Let $d^e = \left(d_i^{(e)}\right)_{i\in E}$. We note $n_l^{(e)} = \sum_{i\in E} a_{i,l} x_i^{(e)} d_i^{(e)}$.

Users essentially perceive performance through the mean time necessary to transfer a document [12]. In the following, we evaluate performance in terms of throughput, defined as the ratio of the mean flow size to the mean flow duration in steady state. Assuming network stability and applying Little's formula, the throughput of a flow of any class *i* is related to the mean number of class-*i* flows $\left(E\left[x_i^{(e)}\right]\right)$ through the relationship:

$$\gamma_i = \frac{\rho_i^{(e)}}{E\left[x_i^{(e)}\right]} \tag{1}$$

2.1. A Single Link Case

In this part, the system has a single link with capacity C. Let $n^{(e)} = \sum_{i\in E} x_i^{(e)} d_i^{(e)}$.

2.1.1. Single Rate Limits

In this section, we will suppose that all classes have the same maximum bit rate d. Let $N = \lfloor C/d \rfloor$ be the maximum number of flows that can be allocated exactly d units on the link. Above this limit, congestion occurs and flows equally share the link capacity C. Our system will be identical to a "Processor sharing" queue. We note by x the total number of flows presented in the link.

The average number of flows for each class $i \in E$ is given as follows [1]:

$$E\left[x_i^{(e)}\right] = \frac{\rho_i^{(e)}}{d_i^{(e)}} + \frac{\rho_i^{(e)}}{C - \theta^{(e)}}\pi(B) \tag{2}$$

where $\pi(B)$ is the probability of the set $B = \{x : x \geq N\}$ presenting all the congestion states. $\pi(B)$ is written as follows:

$$\pi(B) = \frac{\left(\frac{\theta^{(e)}}{d}\right)^N}{N!}\frac{C}{C-\theta^{(e)}}\pi(0) \tag{3}$$

$\pi(0)$ is the probability of $x = 0$. It is given by:

$$\pi(0) = \left(\sum_{x=0}^{N-1}\frac{\left(\frac{\theta^{(e)}}{d}\right)^x}{x!} + \frac{\left(\frac{\theta^{(e)}}{d}\right)^N}{N!}\frac{C}{C-\theta^{(e)}}\right)^{-1} \tag{4}$$

2.1.2. Multi Rate Limits

As there are many class of flows with different transmission rate, the evolution of the number of flows depends on how link capacity is allocated. Most work has focused on so-called utility based allocations, where bandwidth is shared so as to maximize some utility function of the instantaneous flow rates [12]. Examples of such allocations are classical max-min fairness [13] and Kelly's proportional fairness [14]. 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 [15]. It turns out that, for the flow-level dynamics, that we are interested in, proportional fairness can be well approximated by the slightly different notion of balanced fairness [16]-[18]. The notion of balanced fairness was introduced by Bonald and Proutière as a means to approximately evaluate the performance of fair allocations like max-min fairness and proportional fairness in wired networks. A key property of balanced fairness is its insensitivity: the steady state distribution is independent of all traffic characteristics beyond the traffic intensity [15]. The only required assumption is that flows arrive as a Poisson process, which is indeed satisfied in practice.

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 [2]. In [12], Bonald *et al*. propose a recursive algorithm to evaluate performance metrics, in which it is possible to identify congested network links for each system status. Although this algorithm makes it possible to calculate an accurate performance metrics, it is only applicable on some special cases. For complex networks, identification of saturated links is not always feasible. Another approach has been proposed in [19] by Bonald *et al*. to resolve this problem. Under the assumption that the flows do not have a peak rate, the authors propose explicit approximations of key performance metrics in any network topology. In practice, flows generally have a peak rate that is typically a function of the user access line.

In [2] and [9], we proposed some approximations to effectively calculate performance metrics under balanced fairness without requiring the evaluation of individual probabilities of states. These approximations are based on numerical observations and are practically applicable for all network topologies. Then, the average number of flows for each class $i \in E$ can be approximated as follows [2]:

$$E\left[x_i^{(e)}\right] \approx \frac{\rho_i^{(e)}}{d_i^{(e)}} + \frac{\rho_i^{(e)}}{C-\theta^{(e)}}\pi\left(B_i\right) \tag{5}$$

where $\pi\left(B_i\right)$ is the probability of the set $B_i = \left\{x^e : C - d_i^{(e)} < n^{(e)}\right\}$ representing all the congestion states.

$\pi\left(B_i\right)$ is written as follows:

$$\pi\left(B_i\right) = \frac{1}{C-\theta^{(e)}}\sum\nolimits_{k\in E}\rho_k^{(e)}\pi\left(W_k\right) + \pi\left(W_i\right) \tag{6}$$

where $W_i = \left\{x^e : C - d_i^{(e)} < n^{(e)} \le C\right\}$ for all $i \in E$ and $\pi\left(W_i\right)$ is given by:

$$\pi\left(W_i\right) = \pi\left(0\right)\sum\nolimits_{C-d_i^{(e)}<n^{(e)}\le C}\prod\nolimits_{k\in E}\frac{\left(\frac{\rho_k^{(e)}}{d_k^{(e)}}\right)^{x_k^{(e)}}}{x_k^{(e)}!} \tag{7}$$

And:

$$\pi\left(0\right) = \left(\sum\nolimits_{0\le n^{(e)}\le C}\prod\nolimits_{k\in E}\frac{\left(\frac{\rho_k^{(e)}}{d_k^{(e)}}\right)^{x_k^{(e)}}}{x_k^{(e)}!} + \frac{1}{C-\theta^{(e)}}\sum\nolimits_{i\in E}\rho_i^{(e)}\sum\nolimits_{C-d_i^{(e)}<n^{(e)}\le C}\prod\nolimits_{k\in E}\frac{\left(\frac{\rho_k^{(e)}}{d_k^{(e)}}\right)^{x_i^{(e)}}}{x_i^{(e)}!}\right)^{-1} \tag{8}$$

2.2. General Network Expansion

Let us now consider general networks where several flow classes cross various links.

2.2.1. Identical Rate Limits

The average number of class-i flows can be approximated as follows [3]:

$$E\left[x_i^{(e)}\right] \approx \frac{\rho_i^{(e)}}{d} + \sum\nolimits_{l\in\mathcal{L}} a_{i,l}\pi\left(B^{(l)}\right)\frac{\rho_i^{(e)}}{C_l-\theta_l^{(e)}} \tag{9}$$

where $\pi\left(B^{(l)}\right)$ is the probability of the set $B^{(l)} = \left\{x^e : C_l - d < n_l^{(e)}\right\}$ representing all the congestion states on the link *l*:

$$\pi\left(B^{(l)}\right)=\frac{\left(\frac{\theta_l^{(e)}}{d}\right)^{N_l}}{N_l\,!}\frac{C_l}{C_l-\theta_l^{(e)}}\pi^{(l)}\left(0\right) \tag{10}$$

With:

$$N_l=\frac{C_l}{d} \tag{11}$$

And

$$\pi^{(l)}\left(0\right)=\left(\sum\nolimits_{x=0}^{N_l-1}\frac{\left(\frac{\theta_l^{(e)}}{d}\right)^{x}}{x!}+\frac{\left(\frac{\theta_l^{(e)}}{d}\right)^{N_l}}{N_l\,!}\frac{C_l}{C_l-\theta_l^{(e)}}\right)^{-1} \tag{12}$$

2.2.2. Multi Rate Limits

The average number of class-*i* flows can be approximated as follows [2]:

$$E\left[x_i^{(e)}\right]\approx\frac{\rho_i^{(e)}}{d_i^{(e)}}+\sum\nolimits_{l\in\mathcal{L}}a_{i,l}\,\pi\left(B_i^{(l)}\right)\frac{\rho_i^{(e)}}{C_l-\theta_l^{(e)}} \tag{13}$$

where $\pi\left(B_i^{(l)}\right)$ is the probability of the set $B_i^{(l)}=\left\{x^e:C_l-d_i^{(e)}<n_l^{(e)}\right\}$ representing all the congestion states for the class *i* on the link *l*:

$$\pi\left(B_i^{(l)}\right)=\frac{1}{C_l-\theta_l^{(e)}}\sum\nolimits_{k\in E}a_{k,l}\,\rho_k^{(e)}\pi\left(W_k^{(l)}\right)+\pi\left(W_i^{(l)}\right) \tag{14}$$

with $W_i^{(l)}=\left\{x:C_l-d_i^{(e)}<n_l^{(e)}\le C_l\right\}$ for all $i\in E$.

$$\pi\left(W_i^{(l)}\right)=\pi^{(l)}\left(0\right)\sum\nolimits_{C_l-d_i^{(e)}<n_l^{(e)}\le C_l}\prod\nolimits_{k\in E}\frac{\left(\frac{a_{k,l}\,\rho_k^{(e)}}{d_k^{(e)}}\right)^{x_k^{(e)}}}{x_k^{(e)}\,!} \tag{15}$$

And:

$$\pi^{(l)}\left(0\right)=\left(\sum\nolimits_{0\le n_l^{(e)}\le C_l}\prod\nolimits_{k\in E}\frac{\left(\frac{a_{k,l}\,\rho_k^{(e)}}{d_k^{(e)}}\right)^{x_k^{(e)}}}{x_k^{(e)}\,!}+\frac{1}{C_l-\theta_l^{(e)}}\sum\nolimits_{i\in E}a_{i,l}\,\rho_i^{(e)}\sum\nolimits_{C_l-d_i^{(e)}<n_l^{(e)}\le C_l}\prod\nolimits_{k\in E}\frac{\left(\frac{a_{k,l}\,\rho_k^{(e)}}{d_k^{(e)}}\right)^{x_k^{(e)}}}{x_k^{(e)}\,!}\right)^{-1} \tag{16}$$

3. Integration of Streaming and Data Traffic under Multi-Queuing Networks

Little work has been devoted to evaluate the performances of elastic traffic in the existence of streaming flows. In [20], Bonald and Proutière offer an insensitive upper bound for the performances of TCP flows in a network where streaming flows are TCP-friendly and fairly share the bandwidth with elastic flows. In practice, as there are different requirements in term of quality of service, the two types of traffic cannot have the same amount of resources.

The authors of [21]-[23] are interested in the performance evaluation of elastic flows in a network where streaming traffic are adaptive and non-priority. In [21] and [24], the authors justified the need for an appropriate admission control mechanism for streaming flows to guarantee a minimum rate for elastic flows.

In [25], Malhotra proposed a model with priority queues giving the high priority to streaming traffic. He assumed that streaming and elastic traffic have the same peak rate and the capacity left over from serving streaming flows is equally divided among the elastic traffic flows. The approximation given by Malhotra to evaluate the average number of low priority traffic focus basically on the total workload and it is sensitive to the detailed characteristics of traffic. In practice, the network traffic has not the same peak rate, which makes this approximation inapplicable in a real context.

Although that many operators use nowadays the composition between priority queues and bandwidth sharing-based queues to handle the requirements of all traffic in term of quality of service, the existing work on flow modelling of such integration (integration between streaming and data traffic) did not treat this case. In this context, we propose a flow-level model to evaluate the performance of elastic traffic under such multi-queuing system. We assume that network architecture gives the head of priority to real time traffic and shares the remaining capacity between the elastic ongoing flows according to a specific weight.

3.1. The Model

The network consists of a set of links *L* where each link l has a capacity C_l. A random number of streaming and elastic flows compete for the bandwidth of these links. Let E be the set of elastic flow classes and S the set of streaming flow classes. Each class k is characterized by a route R_k consisting of a set of links. When link l is on route R_k we use the natural notation $l \in R_k$. Conversely, defining $E_l \subset E$ (respectively $S_l \subset S$) to be the set of elastic flow classes (respectively streaming flow classes) going through link l. we can equivalently write $k \in E_l$ (or $k \in S_l$ respectively).

Let $\mathcal{A}$ be the incidence matrix for elastic flow classes defined as follows: $a_{i,l} = 1$ if class-*i* flows ($i \in E$) go through link l and it equals to zero otherwise. In the same way we define $\mathcal{C}$ the incidence matrix for streaming flow classes: $c_{j,l} = 1$ if class-*j* flows ($j \in S$) go through link l and it equals to zero otherwise.

Streaming flows are mainly defined by their rate and their mean holding-time. For each streaming class-*j* flows ($j \in S$), we define:

- τ_j: The mean holding-time of flows.
- $d_j^{(s)}$: The rate of each flow.

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

- σ_i: The mean volume transferred by flows.
- $d_i^{(e)}$: The maximum bit rate of each flow.

Flows arrive as an independent Poisson process with rate $\lambda_j^{(s)}$ for streaming class-*j* flows and $\lambda_i^{(e)}$ for elastic class-*i* flows. We refer to the product $\rho_i^{(e)} = \lambda_i^{(e)}\sigma_i$ as the load of elastic class *i*. In the same way, we denote by $\rho_j^{(s)}$ the load of a streaming class $j \in S$, where $\rho_j^{(s)} = \lambda_j^{(s)} d_j^{(s)} \tau_j$.

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). Let us denote by the vector $x^s = \left(x_j^{(s)}\right)_{j\in S}$ (respectively $x^e = \left(x_i^{(e)}\right)_{i\in E}$) the state of streaming classes (respectively the state of elastic classes).

Let $\theta_l^{(e)} = \sum_{i\in E} a_{i,l}\rho_i^{(e)}$ (respectively $\theta_l^{(s)} = \sum_{j\in S} c_{j,l}\rho_j^{(s)}$) be the elastic load (respectively the streaming load) offered to a link l.

To maintain the stability of the system, we assume that the total load of each link $l \in \mathcal{L}$ is strictly inferior to its capacity:

$$\theta_l = \theta_l^{(e)} + \theta_l^{(s)} < C_l \tag{17}$$

In a similar way to the configuration of Internet routers, at the entrance of every link l, there is a LLQ queue combining a priority queue with a number of M_l WFQ queues. Let $\vartheta_{l,m}$, $1 \le m \le M_l$, the weight of the WFQ queue number m of the link l. We assume that $\sum_{m=1}^{M_l} \vartheta_{l,m} = 1$.

The priority queue is devoted to streaming flows, which have strict bandwidth and delay requirements that can be met if the requested capacity is allocated to them completely. Streaming flows whose requirements cannot be met will be blocked rather than allow them into the system and jeopardize the performance of real time traffic. The strict priority, coupled with an admission control (to limit the overall volume of streaming traffic) is generally considered sufficient to meet the quality of service requirements of the underlying audio and video applications [24].

Elastic traffic is distributed throughout the WFQ queues. We assume that each WFQ queue is characterized by a Code Point. A Code Point is an integer that distinguishes WFQ queues from each other. Along its path, each elastic flow pass on queues having the same Code Point.

3.2. Analysis

Initially, the capacity of each linkl is fully allocated to the streaming flows. Once this capacity is totally occupied, the real-time traffic will be blocked. The steady probability of x^s is given then by:

$$\pi^{(s)}\left(x^s\right)=\pi^{(s)}(0)\prod\nolimits_{j\in S}\frac{\left(\frac{\rho_j^{(s)}}{d_j^{(s)}}\right)^{x_j^{(s)}}}{x_j^{(s)}!} \tag{18}$$

where:

$$\pi^{(s)}(0)=\left(\sum\nolimits_{x^s}\prod\nolimits_{j\in S}\frac{\left(\frac{\rho_j^{(s)}}{d_j^{(s)}}\right)^{x_j^{(s)}}}{x_j^{(s)}!}\right)^{-1} \tag{19}$$

Let n_l the quantity of the capacity C_l used by streaming flows:

$$n_l=\sum\nolimits_{j\in S}c_{j,l}x_j^{(s)}d_j^{(s)} \tag{20}$$

For $n_l=0,\cdots,C_l$, we define the two following notations:

- The remaining capacity for elastic traffic on the link l:

$$C^e\left(n_l\right)=C_l-n_l \tag{21}$$

- The steady state probability of having n_l quantity of capacity link C_l used by streaming flows on the link *l*:

$$A\left(n_l\right)=\sum\nolimits_{x^s}\pi^{(s)}\left(\left(c_{j,l}x_j^{(s)}\right)_{j\in S}\right) \tag{22}$$

$C^e\left(n_l\right)$ can be viewed as a concatenation between M_l virtual links of capacity $\vartheta_{l,1}C^e\left(n_l\right),\vartheta_{l,2}C^e\left(n_l\right),\cdots,\vartheta_{l,M_l}C^e\left(n_l\right)$. We note by $E_{l,m}$ the set of elastic flow classes crossing the virtuallink m of the link l. Each virtual link is characterized by a Code Point $\left(CP_{l,m}\right)$.

Let $\theta_{l,m}^{(e)}=\sum\nolimits_{i\in E_{l,m}}\rho_i^{(e)}$ be the load offered to the virtual link m of the link l and $\psi_l=\max_{1\le m\le M_l}\frac{\theta_{l,m}^{(e)}}{\vartheta_{l,m}}$. ψ_l Represents the stability threshold: If $C^e\left(n_l\right)>\psi_l$ then the stability condition is satisfied for all virtual links on link *l*.

For each $l\in\mathcal{L}$, it is important to note that for $C_l-\psi_l\le n_l\le C_l$, there is at least one virtual link whose capacity is not enough to handle its load. Thus, if the probability $P_{C_l}=A\left(C_l-\psi_l\le n_l\le C_l\right)$ is not negligible, it will make our model "unstable" and the performance of elastic traffic unpredictable. Therefore, to maintain a maximum stability, which is the main objective of the network administrators in the IP network design phase, we assume that every capacity C_l is fixed in such way that P_{C_l} is negligible (In Section 4 we will suppose that P_{C_l} doesn't exceed 0.12).

The virtual link of capacity $C_{l,m}^*\left(n_l\right)=\vartheta_{l,m}C^e\left(n_l\right)$ is mainly dedicated to specific elastic flows, but it can be shared among the other elastic flow classes if it remains empty. Let $\mathcal{I}_l(n_l)=\{1\le m\le M_l : C^*_{l,m}(n_l)\le\theta_{l,m}\}$ and $\mathcal{S}_l(n_l)=\{1\le m\le M_l : C^*_{l,m}(n_l)>\theta_{l,m}\}$. We assume that if $m\in\mathcal{I}_l(n_l)$, this virtual link seems to be always occupied. If $\mathcal{I}_l(n_l)\ne\varnothing$ we say that there is a "local instability" on the link l. Therefore P_{C_l} can be called the local instability probability of the link *l*.

The performances of TCP flows will be studied under a quasi-stationary assumption: For every state of x^s, the number of flows for each elastic class evolves rapidly and attains a stationary regime.

3.2.1. First Case: Elastic Flows with the Same Maximum Bit Rate Using the Same Queue

Elastic traffic is distributed throughout these links in such that all flows with the same maximum bit rate pass on the same virtual link. Let $d^{(e)}_{l,m}$ be the maximum bit rate for the virtual link number m of the link l. Virtual links of different links crossed by flows with the same maximum bit rates have the same Code Point. Without loss of generality, we assume that $CP_{l,m}=d^{(e)}_{l,m}$.

In the same way as in [1], if there is no flow crossing the capacity $C^*_{l,m}(n_l)$, this capacity will be shared on the other virtual links according to their weight. Let $\pi^{(l,k)}(0,n_l)$ be the probability that the virtual link number k of the link l is empty when streaming flows used a quantity of resources equal to n_l on this link. $\pi^{(l,k)}(0,n_l)$ is given using (12) as follows:

$$\pi^{(l,k)}(0,n_l)=\left(\sum_{x=0}^{N_{l,k}-1}\frac{\left(\dfrac{\theta^{(e)}_{l,k}}{d^{(e)}_{l,k}}\right)^x}{x!}+\frac{\left(\dfrac{\theta^{(e)}_{l,k}}{d^{(e)}_{l,k}}\right)^{N_{l,k}}}{N_{l,k}!}\frac{C^*_{l,k}(n_l)}{C^*_{l,k}(n_l)-\theta^{(e)}_{l,k}}\right)^{-1} \tag{23}$$

with:

$$N_{l,k}=\frac{C^*_{l,k}(n_l)}{d^{(e)}_{l,k}} \tag{24}$$

Example 1:

We assume that for a link l, $M_l=2$ and for each m, $1\le m\le M_l$, we have $C^*_{l,m}(n_l)>\theta_{l,m}$. The mean capacity for the first virtual link of this link l is given by:

$$\bar{C}^{(e)}_{l,1}(n_l)=\vartheta_{l,1}C^e(n_l)\left(1-\pi^{(l,2)}(0,n_l)\right)+C^e(n_l)\pi^{(l,2)}(0,n_l) \tag{25}$$

Example 2:

We assume that for a link l, $M_l=3$ and for eachm, $1\le m\le M_l$, we have $C^*_{l,m}(n_l)>\theta_{l,m}$. The mean capacity for the first virtual link of this link l is given by:

$$\begin{aligned}\bar{C}^{(e)}_{l,1}(n_l)&=\vartheta_{l,1}C^e(n_l)\left(1-\pi^{(l,2)}(0,n_l)\right)\left(1-\pi^{(l,3)}(0,n_l)\right)\\&+\left[\vartheta_{l,1}C^e(n_l)+\frac{\vartheta_{l,1}}{\vartheta_{l,1}+\vartheta_{l,3}}\vartheta_{l,2}C^e(n_l)\right]\pi^{(l,2)}(0,n_l)\left(1-\pi^{(l,3)}(0,n_l)\right)\\&+\left[\vartheta_{l,1}C^e(n_l)+\frac{\vartheta_{l,1}}{\vartheta_{l,1}+\vartheta_{l,2}}\vartheta_{l,3}C^e(n_l)\right]\pi^{(l,3)}(0,n_l)\left(1-\pi^{(l,2)}(0,n_l)\right)\\&+C^e(n_l)\pi^{(l,2)}(0,n_l)\pi^{(l,3)}(0,n_l)\end{aligned} \tag{26}$$

The expression of $\bar{C}^{(e)}_{l,1}(n_l)$ will be more complex for values of M_l higher than 3. A simple approximation can be given to calculate the mean capacity of each virtual link m of a link l as follows:

$$\bar{C}^{(e)}_{l,m}(n_l)=\frac{\vartheta_{l,m}}{\vartheta_{l,m}+\sum_{k\ne m}\vartheta_{l,k}\left(1-\pi^{(l,k)}(0,n_l)\right)}C^e(n_l) \tag{27}$$

This approximation is based on numerical observations: we compared the exact solution of the mean capacity for each virtual link and the value given by the approximation (27) for many cases and the error rate doesn't exceed 5% for a very low traffic and it is negligible for medium and high traffic. If we take into account the instability of some virtual links, the mean capacity left for a virtual link m of a link l is approximately given by [1]:

$$\bar{C}_{l,m}^{(e)}(n_l)=\frac{\vartheta_{l,m}}{\vartheta_{l,m}+\sum_{\substack{k\in\mathcal{S}_l(n_l)\\ k\neq m}}\vartheta_{l,k}\left(1-\pi^{(l,k)}(0,n_l)\right)+\sum_{\substack{k\in\mathcal{I}_l(n_l)\\ k\neq m}}\vartheta_{l,k}}C^{e}(n_l) \tag{28}$$

Let p_l be the virtual link of each link l satisfying $CP_{l,p_l}=d_i^{(e)}$ with $i\in E$. For every state x^s satisfying $\theta_{l,p_l}^{(e)}\le\bar{C}_{l,p_l}^{(e)}(n_l),\forall l\in R_i$, the average number of class-*i* flows is given using (9) by:

$$E\left(x_i\mid x^s\right)=\frac{\rho_i^{(e)}}{d_i^{(e)}}+\sum_{l\in\mathcal{L}}a_{i,l}\pi\left(B_{n_l}^{(l,p_l)}\right)\frac{\rho_i^{(e)}}{\bar{C}_{l,p_l}^{(e)}(n_l)-\theta_{l,p_l}^{(e)}} \tag{29}$$

where $B_{n_l}^{(l,p_l)}=\left\{x^{(l,p_l)}:N_{l,p_l}^{*}<x^{(l,p_l)}\right\}$ representing all the congestion states on the virtual link p_l of the link l, $\forall l\in\mathcal{L}$, with $x^{(l,p_l)}$ is the number of elastic flows in progress on this virtual link and:

$$N_{l,p_l}^{*}=\frac{\bar{C}_{l,p_l}^{(e)}(n_l)}{d_{l,p_l}^{(e)}} \tag{30}$$

$\pi\left(B_{n_l}^{(l,p_l)}\right)$ is given by:

$$\pi\left(B_{n_l}^{(l,p_l)}\right)=\frac{\left(\dfrac{\theta_{l,p_l}^{(e)}}{d_{l,p}^{(e)}}\right)^{N_{l,p_l}^{*}}}{N_{l,p_l}^{*}!}\frac{\bar{C}_{l,p_l}^{(e)}(n_l)}{\bar{C}_{l,p_l}^{(e)}(n_l)-\theta_{l,p_l}^{(e)}}\pi^{(l,p_l)*}(0,n_l) \tag{31}$$

where:

$$\pi^{(l,p_l)*}(0,n_l)=\left(\sum_{x=0}^{N_{l,p_l}^{*}-1}\frac{\left(\dfrac{\theta_{l,p_l}^{(e)}}{d_{l,p_l}^{(e)}}\right)^{x}}{x!}+\frac{\left(\dfrac{\theta_{l,p_l}^{(e)}}{d_{l,p_l}^{(e)}}\right)^{N_{l,p_l}^{*}}}{N_{l,p_l}^{*}!}\frac{\bar{C}_{l,p_l}^{(e)}(n_l)}{\bar{C}_{l,p_l}^{(e)}(n_l)-\theta_{l,p_l}^{(e)}}\right)^{-1} \tag{32}$$

3.2.2. Second Case: Elastic Flows with Different Maximum Bit Rate Using the Same Queue

Flows with different maximum bit rate can be passed through the same queue. The elastic traffic is differentiated then according the service's type and no according to the maximum bit rate of the flows.

The Equation (28) that gives the mean capacity left for a virtual link m of a link l doesn't change. Nevertheless, $\pi^{(l,k)}(0,n_l)$ is given now by:

$$\begin{aligned}\pi^{(l,k)}(0,n_l)=&\left(\sum_{0\le n_{l,k}^{(e)}\le C_{l,k}^{*}(n_l)}\prod_{i\in E}\frac{\left(\dfrac{a_{i,l,k}\rho_i^{(e)}}{d_i^{(e)}}\right)^{x_i^{(e)}}}{x_i^{(e)}!}\right.\\&\left.+\frac{1}{C_{l,k}^{*}(n_l)-\theta_{l,k}^{(e)}}\sum_{i\in E}a_{i,l,k}\rho_i^{(e)}\sum_{C_{l,k}^{*}(n_l)-d_i^{(e)}<n_{l,k}^{(e)}\le C_{l,k}^{*}(n_l)}\prod_{f\in E}\frac{\left(\dfrac{a_{f,l,k}\rho_f^{(e)}}{d_f^{(e)}}\right)^{x_f^{(e)}}}{x_f^{(e)}!}\right)^{-1}\end{aligned} \tag{33}$$

where $a_{i,l,k}=1$ if the class-*i* flows pass through the virtual link number k on the link l and $a_{i,l,k}=0$ otherwise,

$\forall l\in\mathcal{L}$, $\forall i\in E$, and $n_{l,k}^{(e)}=\sum_{i\in E}a_{i,l,k}x_i^{(e)}d_i^{(e)}$.

Let $i\in E$. We assume that the flows of this class pass among its path on the virtual link number p_l on each link *l*.

For every state x^s satisfying $\theta_{l,p_l}^{(e)}\le\bar{C}_{l,p_l}^{(e)}\left(n_l\right),\forall i\in R_i$, the average number of class-*i* flows is given by:

$$E\left(x_i\mid x^s\right)=\frac{\rho_i^{(e)}}{d_i^{(e)}}+\sum_{l\in\mathcal{L}}a_{i,l}\pi\left(B_{(i,n_l)}^{(l,p_l)}\right)\frac{\rho_i^{(e)}}{\bar{C}_{l,p_l}^{(e)}\left(n_l\right)-\theta_{l,p_l}^{(e)}} \tag{34}$$

with $B_{(i,n_l)}^{(l,p_l)}=\left\{x^e:\bar{C}_{l,p_l}^{(e)}\left(n_l\right)-d_i^{(e)}<n_{l,p_l}^{(e)}\right\}$ and:

$$\pi\left(B_{(i,n_l)}^{(l,p_l)}\right)=\frac{1}{\bar{C}_{l,p_l}^{(e)}\left(n_l\right)-\theta_{l,p_l}^{(e)}}\sum_{f\in E}a_{f,l,p_l}\rho_f^{(e)}\pi\left(W_{(f,n_l)}^{(l,p_l)}\right)+\pi\left(W_{(i,n_l)}^{(l,p_l)}\right) \tag{35}$$

The set $W_{(i,n_l)}^{(l,p_l)}$ and its probability $\pi\left(W_{(i,n_l)}^{(l,p_l)}\right)$ are defined in the same manner as the Section 2.2.2 by replacing C_l by $\bar{C}_{l,p_l}^{(e)}\left(n_l\right)$, $a_{k,l}$ by a_{k,l,p_l}, $\theta_l^{(e)}$ by $\theta_{l,p_l}^{(e)}$ and $n_l^{(e)}$ by $n_{l,p_l}^{(e)}$, $\forall l\in\mathcal{L}$, $\forall i\in E$.

For reasons of simplicity, we assume that if a virtual link m of a link l satisfyies $\theta_{l,m}^{(e)}\ge\bar{C}_{l,m}^{(e)}\left(n_l\right)$ then all flow passing through this virtual link have an end-to-end throughput equal to zero. In fact, the "local instability" deteriorates the throughput of flows passing through this virtual link. So these flows will continue their path with very low rates. With the effect of congestion on other links of the network, it can be assumed that these flows will almost arrive with a throughput equal to zero.

This assumption admits that our capacity is really divided into different independent links and the quality of service seems to be very bad for all elastic classes when a local instability occurs on a specific virtual link.

The mean flow throughput of class-*i* flows is:

$$\gamma_i=\sum_{x^s}\gamma_i\left(x^s\right)\pi^{(s)}\left(x^s\right) \tag{36}$$

with:

$$\gamma_i\left(x^s\right)=\begin{cases}\dfrac{\rho_i^{(e)}}{E\left(x_i\mid x^s\right)} & \text{if } \forall l\in R_i:\theta_{l,p_l}^{(e)}\le\bar{C}_{l,p_l}^{(e)}\left(n_l\right)\\ 0 & \text{else}\end{cases} \tag{37}$$

The approximation proposed is completely insensitive to both the service time distribution of stream traffic and the file size distribution of elastic traffic. This is an extremely useful property for operators in that it suggests that provisioning does not depend on the precise characteristics of applications which can change quite radically over time.

4. Validation of the Analytical Model by Means of Simulations

To validate our results, we apply the approximation proposed in the previous section to two specific network topologies: Linear network and tree network. In all graphs below, we plot a comparison between the analytical model and the exact model of the average flow throughput. The accuracy of our approximation is verified from the relative error defined as:

$$\text{RelErr}=100\frac{\left|\text{Simulation Result}-\text{Analytical Result}\right|}{\text{Simulation Result}} \tag{38}$$

In the following, the probability P_{C_l} is expressed in percentage ($P_{C_l}\left(\%\right)=100*P_{C_l}$). We note by q_{li} the queue situated at the entrance of the link of capacity C_i.

4.1. Linear Network

We consider the linear network presented in **Figure 1**. All queues considered are LLQ queues and their configurations are illustrated in **Table 1**. Two streaming flow-classes and five elastic flow-classes compete for the resources of the network. **Table 2** gives the parameters values of traffic carried by this network.

While C_2 is variable, the two other capacities C_1 and C_3 are fixed in such that $P_{C_1} = 0.47\%$ and $P_{C_3} = 0.31\%$.

For this network, we assume that flows with the same maximum bit rate use the same queue, and we assume that for each link $l \in \mathcal{L}$:

$$d_{l,1}^{(e)} \le d_{l,2}^{(e)} \le \cdots \le d_{l,M_l}^{(e)}$$

Figure 2 plots a comparison between the analytical results and the simulation results of the average flow throughput as a function of the percentage of P_{C_2} for the first and the third class flows. As expected, the local instability affects badly the average flow throughput for each class. The capacity of links can be then fixed according to the total load passing through it and the level of QoS that we aim to provide for elastic traffic.

We can note that for values of P_{C_2} inferior to 11%, the two results are very close: the error rate does not exceed 3% for both classes, and it seems a little bit inferior to 1% when P_{C_2} is less than 5%. This observation confirms our results and proves that in a stable system where P_{C_2} remains negligible, our approximation estimates very well the performance of the elastic traffic under multi-queuing system.

4.2. Tree Network

Now let us consider the tree network illustrated in **Figure 3**. Five streaming flow-classes and ten elastic flow–classes compete for the resources of the network. **Table 3** gives the parameters values of traffic crossing this network. All queues are LLQ and their configurations are illustrated in **Table 4**.

C_1, C_2, C_3, C_4 and C_5 are fixed in such that $P_{C_1} = 0.02\%$, $P_{C_2} = 0.02\%$, $P_{C_3} = 0.09\%$, $P_{C_4} = 0.06\%$ and $P_{C_5} = 0.09\%$. The capacity C_6 is variable. For this network, we assume that flows with different maximum bit rate can pass on the same queue. Along its path, each flow passes on queues having the same code

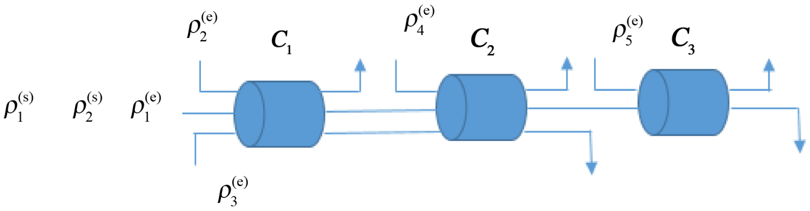

Figure 1. Linear network.

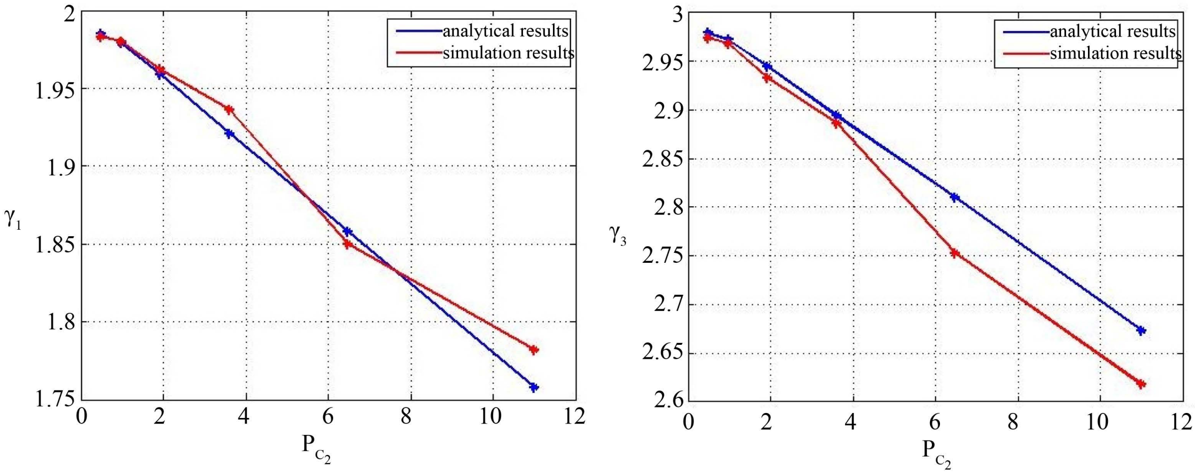

Figure 2. Comparison between the analytical result and the simulation result of the average flow throughput as a function of the percentage of P_{C_2} for the first and the third class.

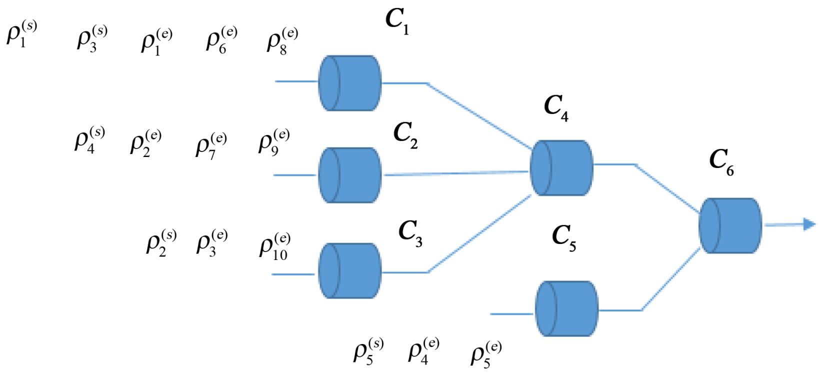

Figure 3. Tree network.

Table 1. Queues configurations values of linear network.

q_{l_i}	M_{l_i}	$\vartheta_{l,m}, 1 \le m \le M_{l_i}$
q_{l_1}	3	$\vartheta_{l_1,1} = 0.2\vartheta_{l_1,2} = 0.3\vartheta_{l_1,3} = 0.5$
q_{l_2}	2	$\vartheta_{l_2,1} = 0.4\vartheta_{l_2,2} = 0.6$
q_{l_3}	2	$\vartheta_{l_3,1} = 0.4\vartheta_{l_3,2} = 0.6$

Table 2. Traffic parameters values for the linear network.

Streaming classes			
Class	$\lambda^{(s)}$	τ	$d^{(s)}$
Class 1	0.1	10	10
Class2	0.1	20	5
Elastic classes			
Class	$\lambda^{(e)}$	σ	$d^{(e)}$
Class 1	2	1	2
Class 2	2	1	1
Class 3	3	1	3
Class 4	1	1	2
Class 5	1	1	1

point. We assume that the flows of class 1, 7 and 10 pass on the queues having a Code Point equal to 20, the flows of class 5, 6 and 9 pass on queues having a Code Point equal to 30 and the rest of class-flows passes on queues having a Code Point equal to 40.

A comparison between the analytical results and the simulation results of the average flow throughput as a function of the percentage of P_{C_6} for the first and the fifth class flows is respectively shown in **Figure 4**.

In a stable zone, where P_{C_6} is less than 12%, we observe that the relative error is less than 2% for both classes which confirms the good behavior of our approximation.

In practice, link bandwidth is not shared as precisely as assumed in the fluid models. TCP uses some algorithms (Slow Start, Congestion Avoidance…) to control congestion inside the network and restrict the throughput of flows. However, for large scale networks we maintain that fluid models provide "very valuable insight into the impact on performance of traffic characteristics" [26]. The insensitivity of average performance to the detailed statistical properties of connections is of great importance for network engineering. This property is likely

Table 3. Traffic parameters values for the tree network.

Streaming classes			
Class	$\lambda^{(s)}$	τ	$d^{(s)}$
Class 1	0.1	10	10
Class 2	0.1	20	5
Class 3	0.1	10	10
Class 4	0.1	20	5
Class 5	0.1	20	5
Elastic classes			
Class	$\lambda^{(e)}$	σ	$d^{(e)}$
Class 1	2	1	2
Class 2	2	1	1
Class 3	3	1	3
Class 4	1	1	2
Class 5	1	1	1
Class 6	2	1	2
Class 7	1	1	2
Class 8	3	1	3
Class 9	1	1	3
Class 10	2	1	1

Table 4. Configurations of LLQ Queues for the tree network.

q_{l_i}	M_{l_i}	$\vartheta_{l,m}--CP_{l,m}, 1 \le m \le M_{l_i}$
q_{l_1}	3	0.2--200.3--30 0.5--40
q_{l_2}	3	0.2--200.3--30 0.5--40
q_{l_3}	2	0.4--200.6--40
q_{l_4}	3	0.2--200.3--30 0.5--40
q_{l_5}	2	0.4--300.6--40
q_{l_6}	3	0.2--200.3--30 0.5--40

to be maintained approximately even when accounting for disparities due to packet level behavior [26].

5. Conclusions

A key design objective of traffic control schemes in communication networks is to ensure maximum stability. Performance is generally much better and more predictable if the system is uniformly stable, having no or negligible periods of local instability. In this sense, we have derived a good approximation to evaluate the average end-to-end throughput of elastic traffic under multi-queuing system using a quasi-stationary approximation. Assuming priority service for streaming traffic, the remaining capacity is shared between the elastic traffic according their weight. This remaining capacity can be viewed as a concatenation of a set of virtual links, and every virtual link is related to a specific elastic flow classes. Studying the performance of each elastic flow is, therefore, equivalent to studying a single flow class passing on a set of links. So that, the results (5) and (13) are useful here in that they simply give the mean number of flows for a single class. Detailed packet level simulations show that the proposed formulas yield good results.

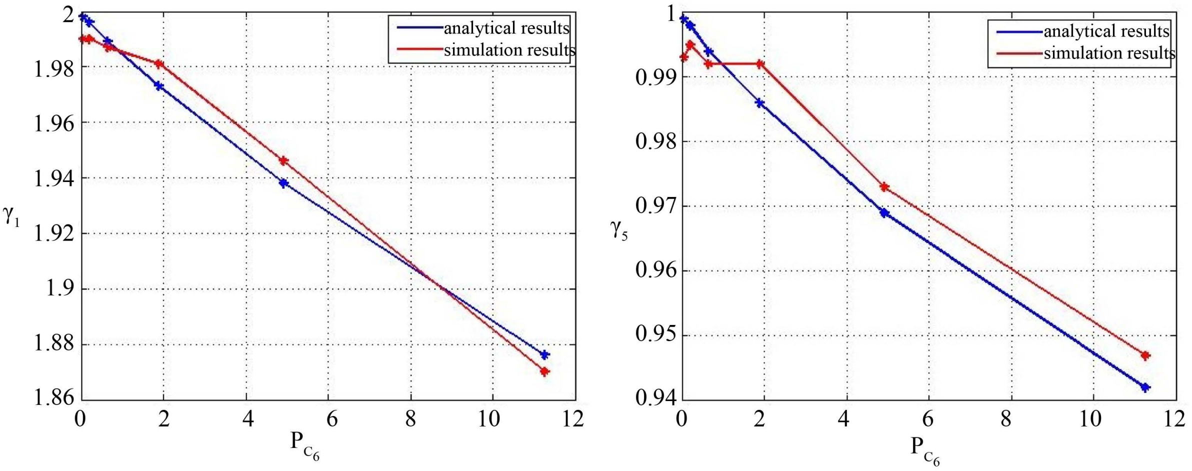

Figure 4. Comparison between the analytical results and the simulation results of the average flow throughput as a function of the percentage P_{C_2} for the first and the fifth class.

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 using a quasi-stationary assumption is precise and allows a generalization for large networks with a reasonable computation time.

Another key result is that the approximation proposed is insensitive to detailed traffic characteristics. 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 data networks.

References

[1] Boussada, M.E.H., Garcia, J.M. and Frikha, M. (2015) Flow Level Modelling of Internet Traffic in Diffserv Queuing. *Proceedings of the* 5*th International Conference on Communications and Networking*, Hammamet, November 2015.

[2] Garcia, J.M. and Boussada, M.E.H. (2016) Evaluation des performances bout en bout du trafic TCP sous le régime «Équité Équilibrée». *Proceedings of the* 11*th Performance Evaluation Workshop*, Toulouse, March 2016, 15.

[3] Boussada, M.E.L., Garcia, J.M. and Frikha, M. (2016) Evaluation des performances bout en bout du trafic TCP dans une architecture de réseau multi files d'attente. *The* 17*th Congress of the French Society of Operations Research and Decision Support* (*ROADEF*), Compiegne, February 2016, 72.

[4] http://www.cisco.com/en/US/tech/tk543/tk544/tk399/tsd_technology_support_sub-protocol_home.html

[5] http://www.enterprise.huawei.comilink/enenterprise/download/HW_U_149185

[6] http://www.journaldunet.com

[7] http://www.e-commercons.com

[8] http://www.lemonde.fr/technologies/article/2012/10/11/2-3-milliards-d-internautes-dans-le-monde_1774055_651865.html

[9] Cheikh, H.B. (2015) Evaluation et optimisation de la performance des flots dans les réseaux stochastiques à partage de bande passante. PhDThesis, INSA, Toulouse.

[10] Olivier, B., Al Sheikh, A. and Garcia, J.M. (2009) Flow-Level Modelling of TCP Traffic Using GPS Queueing Networks. 21*st International Teletraffic Congress*, Paris, 2009, 1-8.

[11] Alexandra Mihaela, N. (2009) Mécanismes d'optimisation multi-niveaux pour IP sur satellites de nouvelle génération. Diss.

[12] Thomas, B. and Virtamo, J. (2004) Calculating the Flow Level Performance of Balanced Fairness in Tree Networks. *Performance Evaluation*, **58.1**, 1-14. http://dx.doi.org/10.1016/j.peva.2004.03.001

[13] Bertsekas, D. and Gallager, R. (1992) Data Networks. 2nd Edition, Prentice Hall, Englewood Cliffs.

[14] Kelly, F., Mauloo, A. and Tan, D. (1998) Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability. *Journal of the Operational Research Society*, **49**, 237-252. http://dx.doi.org/10.1057/palgrave.jors.2600523

[15] Bonald, T., Haddad, J.P. and Mazumdar, R.R. (2011) Congestion in Large Balanced Multirate Links. *Proceedings of the* 23*rd International Teletraffic Congress*, San Francisco, 2011, 182-189.

[16] Bonald, T., Massouli´e, L., Prouti`ere, A. and Virtamo, J. (2006) A Queuing Analysis of Max-Min Fairness, Proportional Fairness and Balanced Fairness. *Queueing Systems*, *Theory and Applications*, **53**, 65-84. http://dx.doi.org/10.1007/s11134-006-7587-7

[17] Bonald, T. and Prouti`ere, A. (2003) Insensitive Bandwidth Sharing in Data Networks. *Queueing Systems*, *Theory and Applications*, **44**, 69-100. http://dx.doi.org/10.1023/A:1024094807532

[18] Massouli´e, L. (2007) Structural Properties of Proportional Fairness: Stability and Insensitivity. *The Annals of Applied Probability*, **17**, 809-839. http://dx.doi.org/10.1214/105051606000000907

[19] Bonald, T., Penttinen, A. and Virtamo, J. (2006) On Light and Heavy Traffic Approximations of Balanced Fairness. 2006 *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, Saint Malo, 26-30 June 2006, 109-120. http://dx.doi.org/10.1145/1140103.1140291

[20] Bonald, T. and Proutière, A. (2004) On Performance Bounds for the Integration of Elastic and Adaptive Streaming Flows. *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*, New York, 12-16 June 2004, 235-245. http://dx.doi.org/10.1145/1012888.1005716

[21] Delcoigne, F., Proutiere, A. and Régnié, G. (2004) Modeling Integration of Streaming and Data Traffic. *Performance Evaluation*, **55**, 185-209. http://dx.doi.org/10.1016/S0166-5316(03)00115-9

[22] Key, P., Massoulié, L., Bain, A. and Kelly, F. (2004) Fair Internet Traffic Integration: Network Flow Models and Analysis. *Annals of Telecommunications*, **59**, 1338-1352.

[23] Queija, R.N., Van den Berg, J.L. and Mandjes, M.R.H. (1999) Performance Evaluation of Strategies for Integration of Elastic and Stream Traffic. Report-Probability, Networks and Algorithms, No 3, 1-17.

[24] Benameur, N., Fredj, S.B., Delcoigne, F., Oueslati-Boulahia, S. and Roberts, J.W. (2001) Integrated Admission Control for Streaming and Elastic Traffic. In: Smirnov, M.I., Crowcroft, J., Roberts, J. and Boavida, F., Eds., *Quality of Future Internet Services*, Springer, Berlin, 69-81. http://dx.doi.org/10.1007/3-540-45412-8_6

[25] Malhotra, R. and van den Berg, J.L. (2006) Flow Level Performance Approximations for Elastic Traffic Integrated with Prioritized Stream Traffic. 12*th International Telecommunications Network Strategy and Planning Symposium*, New Delhi, November 2006, 1-9. http://dx.doi.org/10.1109/netwks.2006.300367

[26] Thomas, B. and Roberts, J.W. (2003) Congestion at Flow Level and the Impact of User Behaviour. *Computer Networks*, **42**, 521-536. http://dx.doi.org/10.1016/S1389-1286(03)00200-7