Flow Level Modelling of Internet Traffic in Diffserv Queuing

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

Flow Level Modelling of Internet Traffic in Diffserv Queuing

Mohamed El Hedi Boussada Mounir Frikha
Mobile Network and Multimedia
SUP'COM
Ariana, Tunisia
{med.elhadi.boussada, m.frikha}@supcom.tn

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

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 stream 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 performance of elastic traffic under Diffserv architecture. This model is based on the fluid flow approximation. We assume that Diffserv 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 model, Diffserv, Quality of Service, streaming traffic, elastic traffic

I. INTRODUCTION

Traditionally, computer communication networks (Internet for example) provide a "best-effort" service. By this, we mean that the network is not able to provide Quality of Service (QoS) to the data streams, either in time or in throughput [9]. All types of traffic passing through the network are, therefore, treated by the same manner.
For applications that circulated at the beginning of the Internet, these limitations were not troublesome for two reasons: On the one hand, these applications are insensitive to temporal variations (email or file transfer for example); on the other hand, the load of networks was limited which left enough bandwidth available for coming traffic [10].

Today, there has been interest in supporting real-time communication applications in the packet-based environments, such as interactive voice, video applications, online gaming, and videoconference applications. Therefore, we shall distinguish two broad categories of Internet traffic: stream and elastic. Stream traffic is generated by applications such as Voice over Internet protocol applications (VoIP applications), streaming video etc. 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 and results in an elastic packet transmission rate [20].
For these applications, the total amount of time required to download the file or web-page is of importance. The end-to-end delay experienced by each packet and the jitter are not of relevance. A useful abstraction in this context is to view each transfer file as a fluid elastic connection, whose rate adapts to the evolution of the number of other flows that share the same links [17], and this is the principle of the flow level modeling.

In contrast with the packet-level models, which define how the packets are generated and transported during the communication, 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 [1]. 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 [4].
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 [18]. In flow level models, a flow is seen like an end-to-end connection between two entities. 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.
To support both stream 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 [19]. Many studies have been done to perform service's differentiation. Nowadays, 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). The composition between the two policies was considered by many telecommunication equipment constructors like Cisco [8] and Huawei [11]. 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) [8].

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

II. BANDWIDTH SHARING WITH ELASTIC FLOWS ONLY

We consider a single link with capacity C (Mbits/Seconde) shared by a random number of elastic flow classes. Let E be the set of these elastic flow classes. Class- i flows, $i \in E$, arrive as an independent Poisson process with rate $\lambda_i^{(e)}$(Flows/Seconde) and have independent, exponentially distributed volumes with means σ_i(Mbits/flow). We refer to the product $\rho_i^{(e)} = \lambda_i^{(e)}\sigma_i$ (Mbits/Seconde) as the traffic intensity of class i. Let $\theta^{(e)} = \sum_{i\in E} \rho_i^{(e)}$ be the total traffic volume generated by all elastic flows. We assume that $\theta^{(e)} < C$ to ensure the stability of our system.

Each flow of a class i has a maximum bit rate $d_i^{(e)} \le C$. This is the actual rate of each flow in the absence of congestion, it means when $\sum_{i\in E} x_i^{(e)} d_i^{(e)} \le C$, with $x_i^{(e)}$ is the number of class- i flows. Congestion forces flows to reduce their rate and thus to increase their duration. We refer to the vector $x^e = \left(x_i^{(e)}\right)_{i\in E}$ as the network state.

The evolution of the system state defines a multidimensional Markov process with transition rates $\lambda_i^{(e)}$ from state x^e to state $x^e + e_i$ and $d_i^{(e)}(x^e)/\sigma_i$ from state x^e to state $x^e - e_i$ (provided $x_i^{(e)} > 0$), where $d_i^{(e)}(x^e)$ is the bit rate of class-i flows in a state x^e: $d_i^{(e)}(x^e) \le d_i^{(e)}$.

Users perceive performance essentially through the mean time necessary to transfer a document [3]. 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 $i \in E$ is related to the expected number of class- i flows in steady state, $\left(E\left[x_i^{(e)}\right]\right)$, through the relationship:

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

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 [3]. Examples of such allocations are classical max-min fairness [14] and Kelly's proportional fairness [13]. 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 [4]. It turns out that, for the flow-level dynamics we are interested in, proportional fairness can be well approximated by the slightly different notion of balanced fairness [2], [5], [6]. 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 [4]. The only required assumption is that flows arrive as a Poisson process, which is indeed satisfied in practice.

The performance of elastic traffic under Balanced Fairness allocation is treated by many studies [2] [3] [4] [5] [6]. In this section, we will suppose that all classes have the same maximum bit rate $d_i^{(e)} = d$ for all $i \in E$ and we will donate a simple and practical analytical expression for the mean number of flows of each class. This expression will be used in the next section .

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. For a single class case (with traffic intensity ρ), the arrivals and departures of flows can be modeled by the following birth-death process:

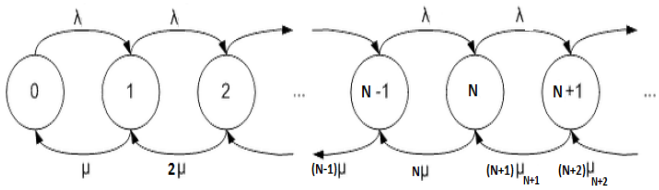

Figure 1. Birth-death process describing the arrivals and departures of connections

Where:

$$\mu = \frac{d}{\sigma} \quad (2)$$

$$\mu_x = \frac{C}{x\sigma} \quad (3)$$

(x: The number of flows in steady state).

The stationary distribution of this Markov process is given by:

$$\pi(x) = \begin{cases} \frac{\left(\frac{\rho}{d}\right)^x}{x!}\pi(0) & \text{if } x \le N \\ \frac{\rho^x}{d^N C^{x-N} N!}\pi(0) & \text{else} \end{cases} \quad (4)$$

Where: $$\pi(0) = \left(\sum_{x=0}^{N-1} \frac{\left(\frac{\rho}{d}\right)^x}{x!} + \frac{\left(\frac{\rho}{d}\right)^N}{N!}\frac{C}{C-\rho}\right)^{-1} \quad (5)$$

Therefore: $$E[x] = \frac{\rho}{d} + \frac{\left(\frac{\rho}{d}\right)^N}{N!}\frac{C}{C-\rho}\frac{\rho}{C-\rho}\pi(0) \quad (6)$$

Let B be the congestion probability on the link. It is written as follows:

$$B = \Pr[x \ge N] = \frac{\left(\frac{\rho}{d}\right)^N}{N!}\frac{C}{C-\rho}\,\pi(0) \quad (7)$$

Thus: $$E[x] = \frac{\rho}{d} + B\frac{\rho}{C-\rho} \quad (8)$$

In the case of many classes with identical maximum bit rate, an aggregation of all flows can be done to have a single class with total load $\theta^{(e)}$. The average number of flows of this class is donated by (8) and $E\left[x_i^{(e)}\right]$ can be approximated as follow:

$$E\left[x_i^{(e)}\right] = \frac{\rho_i}{\theta^{(e)}}E[x] \quad (9)$$

As compared to the exact results given by a numerical resolution of the multidimensional Markov chain. The relative error given by the ratio $\frac{\left|E\left[x_i^{(e)}\right]_{Exact} - E\left[x_i^{(e)}\right]_{Approximation}\right|}{E\left[x_i^{(e)}\right]_{Exact}}$, does not exceed 2 %. This approximation is interesting in that it simplifies the study of the performances of elastic traffic with identical maximum bit rate: A simple aggregation of flows can replace a somewhat complex resolution of the Markov chain.

All users have the same average transmission rate $\gamma_i = \gamma = \theta^{(e)}/E[x]$. The average flow throughput depends on the traffic intensity and the maximum rate per user. Fig.2 presents the evolution of the ratio γ/d in function of $\theta^{(e)}/C$ for different values of d/C.

It should be noted that for small access rate d , flow's throughput is approximately insensitive to offered load, provided $\theta^{(e)} < C$. In other words, the link is virtually transparent to the users, who perceived QoS depends much more on the access rate d.

In general, the flow's throughput deteriorates for high values of $\theta^{(e)}$, and this degradation increases when the ration d/C become higher. However, we can say that for a stable system, users always have a good transmission quality, because for links designed to support the connections of hundreds of subscribers, the ratio d/C is lower than 0.1 [7].

We conclude, therefore, that γ depends essentially on the traffic intensity. This can be a way to correctly dimension a link. In fact, Network dimensioning rules can be developed based on traffic intensity forecasts only, independently of the complex traffic structure which is continually evolving as new applications gain popularity [3]. Insensitivity is the key to simple and robust performance results.

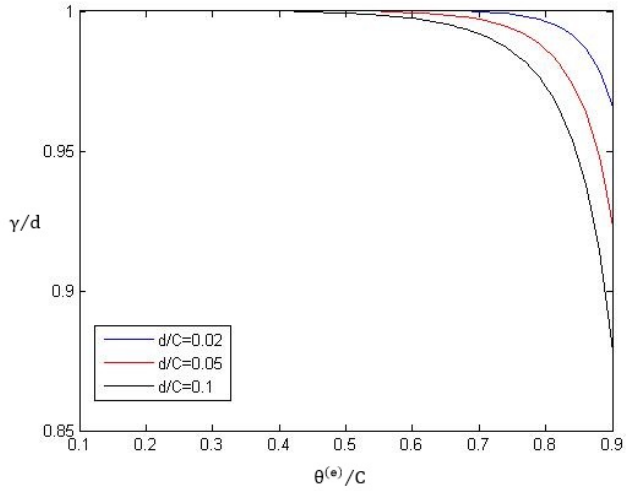

Figure 2. Evolution of the ratio γ/d in function of $\theta^{(e)}/C$ for different values of d/C.

III. INTEGRATION OF STREAMING AND DATA TRAFFIC UNDER DIFFSERV ARCHITECTURE

Little work has been devoted to evaluate the performances of elastic traffic in the existence of streaming flows. In [12], 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 is different requirements in term of quality of service, the two types of traffic cannot have the same amount of resources.

The authors of [21], and [22] interested in the performance evaluation of elastic flows in a network where streaming traffic are priority and non-adaptive. In [16] and [19] the authors justified the need for an appropriate admission control mechanism for streaming flows to guarantee a minimum rate for elastic flows.

In [20] Malhotra proposed a model with priority queues giving the high priority to stream traffic. He assumed that stream and elastic traffic have the same peak rate and the capacity left over from serving stream flows is equally divided among the elastic traffic flows. The approximation given by Malhotra to evaluate the average number of low-priority flows focus basically on the total workload and it is sensitive to the detailed characteristics of traffic. In practice, the network

traffic have not the same peak rate, which make 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-level modelling of such integration (integration between stream 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.

A. Model

We consider a flow level model of both stream and elastic traffic. Our capacity C is shared now by a random number of streaming and elastic flow classes. Let E be the set of elastic flow classes and S the set of streaming flow classes.

Elastic traffic is defined in the same way as the previous section. Streaming flows are principally 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 (Seconde).
- $d_j^{(s)}$: The rate of each flow (Mbits/Seconde).

Streaming flows arrive as an independent Poisson process with rate $\lambda_j^{(s)}$(flows/Seconde). We refer to the product $\rho_j^{(s)} = \lambda_j^{(s)} d_j^{(s)} \tau_j$ (Mbits/Seconde) as the load of a streaming class $j \in S$. We note by $\theta^{(s)} = \sum_{j\in S} \rho_j^{(s)}$ the streaming load offered to the link capacity.

Let $x_j^{(s)}$ be the number of class-j flows in progress .We note by the vector $x^s = \left(x_j^{(s)}\right)_{j\in S}$ the state of streaming classes.

To maintain the stability of the system, we assume that:

$$\theta = \theta^{(e)} + \theta^{(s)} < C \qquad (10)$$

At the entrance of the link, there is a LLQ queue combining a priority queue with a number of M WFQ queues. Let ϑ_m ,$1 \le m \le M$ the weight of the WFQ queue number m. We assume that $\sum_{m=1}^{M} \vartheta_m = 1$.

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.

Elastic traffic is distributed throughout the WFQ queues in such that all elastic flows with the same maximum bit rate pass on the same queue. We note by E_m the set of elastic flow classes passing on the queue number m and by d_m the maximum bit rate of these flows.

Rather than solving this system exactly under Markovian assumptions (this would only lead to complex calculations), the performances of TCP flows will be studied under a quasi-stationary assumption: The ratio $\lambda_j^{(s)} / \lambda_i^{(e)}$ ($j \in S, i \in E$), is assumed small enough so that, in every state of x^s, the number of elastic flows evolves rapidly and attains a stationary regime.

B. Analysis

Let n the quantity of the capacity C used by streaming flows:

$$n = \sum_{j\in S} x_j^{(s)} d_j^{(s)} \qquad (11)$$

The steady distribution of x^s is donated by:

$$\pi^{(s)}(x^s) = \pi^{(s)}(0) \prod_{j\in S} \frac{\left(\frac{\rho_j^{(s)}}{d_j^{(s)}}\right)^{x_j^{(s)}}}{x_j^{(s)!}} \qquad (12)$$

Where:

$$\pi^{(s)}(0) = \left(\sum_{0\le n\le C} \prod_{j\in S} \frac{\left(\frac{\rho_j^{(s)}}{d_j^{(s)}}\right)^{x_j^{(s)}}}{x_j^{(s)!}} \right)^{-1} \qquad (13)$$

For $n = 0..C$, we define the two following notations:

- The remaining capacity for elastic traffic:

$$C^e(n) = C - n \qquad (14)$$

- The steady probability of having n quantity of capacity link C used by streaming flows:

$$A(n) = \sum_{x^s:\sum_{j\in S} x_j^{(s)} d_j^{(s)} = n} \pi^{(s)}(x^s) \qquad (15)$$

$C^e(n)$ can be viewed as a concatenation between M virtual links of capacity $\vartheta_1 C^e(n)$, $\vartheta_2 C^e(n)$, .., $\vartheta_M C^e(n)$.

Let $\theta_m^{(e)} = \sum_{i\in E_m} \rho_i^{(e)}$ be the load offered to the virtual link m and $\psi = \max_{1\le m\le M} \frac{\theta_m^{(e)}}{\vartheta_m}$.

It is important to note that for $C - \psi \le n \le C$, there is at least one virtual link whose capacity is not enough to handle its load. Thus, if the probability $P_{Instability} = A(C - \psi \le n \le C)$ is not negligible, it will make our model "unstable" and the performance of elastic traffic unpredictable.

In the IP network design phase, we must take into account this "local instability". Remove definitely this instability requires high capacity links, or resources are expensive, so we can tolerate a local instability threshold ε that does not affect network performances. Let C^{min} is the minimum value of capacity that meets the constraint: $P_{Instability} \le \varepsilon$. Therefore,

our capacity must satisfy $C \geq C^{min}$, and C^{min} can be donated by the following algorithm:

1. $\Delta d^{(s)} = GCD(d_j^{(s)}, j \in S)$
2. $C^{min} = \theta^{(e)} + \theta^{(s)} + 1$
3. do
   $\{C^{min} = C^{min} + \Delta d^{(s)}\}$
   until $\{P_{Instability} \leq \varepsilon\}$

The virtual link of capacity $C_m^*(n) = \vartheta_m C^e(n)$ is dedicated to flows whose maximum bit rate is equal to d_m , but it can be shared among the other elastic flow classes if it remains empty. Let $\mathcal{I}(n) = \left\{1 \leq m \leq M: C_m^*(n) \leq \theta_m^{(e)}\right\}$ and $\mathcal{S}(n) = \left\{1 \leq m \leq M: C_m^*(n) > \theta_m^{(e)}\right\}$. We assume that if $m \in \mathcal{I}(n)$, this virtual link seems to be always occupied. If $\mathcal{I}(n) \neq \emptyset$ we say that there is a « local instability » on the link. Thus, the mean capacity left for a virtual link m is approximately given by:

$$\overline{C}_m^{(e)}(n) = \frac{\vartheta_m}{\vartheta_m + \sum_{\substack{k \in \mathcal{S}(n) \\ k \neq m}} \vartheta_k \left(1 - \pi_k^{(e)}(0,n)\right) + \sum_{\substack{k \in \mathcal{I}(n) \\ k \neq m}} \vartheta_k} C^e(n) \quad (16)$$

$\pi_k^{(e)}(0,n)$ is given using (5) as follows :

$$\pi_k^{(e)}(0,n) = \left(\sum_{x=0}^{N_k - 1} \frac{\left(\frac{\theta_k^{(e)}}{d_k}\right)^x}{x!} + \frac{\left(\frac{\theta_k^{(e)}}{d_k}\right)^{N_k}}{N_k!} \frac{C_k^*(n)}{C_k^*(n) - \theta_k^{(e)}} \right)^{-1} \quad (17)$$

With:

$$N_k = \left\lfloor \frac{C_k^*(n)}{d_k} \right\rfloor \quad (18)$$

For reasons of simplicity, we assume that if a virtual link m satisfies $\theta_m^{(e)} \geq \overline{C}_m^{(e)}(n)$ in a state x^s then all flows passing through this virtual link have an instantaneous 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 flows traversing a specific virtual link when the local instability occurs on this virtual link.

Let p be the virtual link satisfying $d_p = d_i^{(e)}$ with $i \in E$. For every state n satisfying $\theta_p^{(e)} \leq \overline{C}_p^{(e)}(n)$, the average number of class-i flows is donated by:

$$E\left[x_i^{(e)} | n\right] = \frac{\rho_i^{(e)}}{d_i^{(e)}} + B_p(n) \frac{\rho_i^{(e)}}{\overline{C}_p^{(e)}(n) - \theta_p^{(e)}} \quad (19)$$

$$B_p(n) = \frac{\left(\frac{\theta_p^{(e)}}{d_p}\right)^{N_p^*(n)}}{N_p^*(n)!} \frac{\overline{C}_p^{(e)}(n)}{\overline{C}_p^{(e)}(n) - \theta_p^{(e)}} \pi_p^*(0,n) \quad (20)$$

$$N_p^*(n) = \left\lfloor \frac{\overline{C}_p^{(e)}(n)}{d_p} \right\rfloor \quad (21)$$

$$\pi_p^*(0,n) = \left(\sum_{x=0}^{N_p^*(n)-1} \frac{\left(\frac{\theta_p^{(e)}}{d_p}\right)^x}{x!} + \frac{\left(\frac{\theta_p^{(e)}}{d_p}\right)^{N_p^*(n)}}{N_p^*(n)!} \frac{\overline{C}_p^{(e)}(n)}{\overline{C}_p^{(e)}(n) - \theta_p^{(e)}} \right)^{-1} \quad (22)$$

The mean flow throughput of class-i flows is:

$$\gamma_i = \sum_n \gamma_i(n)\ A(n) \quad (23)$$

With:

$$\gamma_i(n) = \begin{cases} \frac{\rho_i^{(e)}}{E\left[x_i^{(e)} | n\right]} & \text{if } \theta_p^{(e)} \leq \overline{C}_p^{(e)}(n) \\ 0 & \text{else} \end{cases} \quad (24)$$

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 in that it suggests that provisioning does not depend on the precise characteristics of applications which can change quite radically over time.

IV. SIMULATIONS AND VALIDITY OF ANALYTICAL RESULTS

To validate our results, we apply the approximation proposed in the previous section to a specific case where a capacity C is shared among two streaming flow-classes and five elastic flow-classes. TCP flows are generated by each source node according to a Poisson process. Each flow is used to transfer a stream of 1Kbyte packets representing a document of a certain size and then terminated. The application that we chose to be implemented for elastic traffic is FTP (File Transfer Protocol). Streaming traffic is generated in the form of UDP flows (UDP: User Datagram Protocol). Flow duration is drawn from an exponential distribution. We used in this level a CBR traffic model which characterized by a fixed constant rate. Table (I) donates the parameters values of traffic traversing this system.

Table I. Traffic Parameters values

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

To analyze the effect of local instability on the accuracy of our results we took C variable. At the entrance of the link there is a LLQ queue combining a priority queue with three WFQ queues. We assume that $d_1 \leq d_2 \leq d_3$ and we took $\vartheta_1 = 0.2$, $\vartheta_2 = 0.3$ and $\vartheta_3 = 0.5$. Fig. 3, Fig. 4 and Fig. 5 show the evolution of the relative error between the analytical result and the exact result of the average flow throughput as a function of $P_{Instability}$ for the three WFQ queues respectively. The probability $P_{Instability}$ is expressed in percentage ($P_{Instability}(\%) = 100 * P_{Instability}$) and the relative error is defined as:

$$\text{Relative Error} = 100\frac{|\text{Simulation Result} - \text{Analytical Result}|}{\text{Simulation Result}} \quad (25)$$

We can note that for values of $P_{Instability}$ inferior to 5%, the error rate does not exceed 3.5% for all queues, and it seems a little bit inferior to 1% when $P_{Instability}$ is less than 1.5%. That confirms our results and proves that in a stable system, where $P_{Instability}$ remains negligible, this approximation estimates very well the performance of the elastic traffic under multi-queuing system. The local instability affects badly the mean flow throughput of elastic traffic. The capacity C can be then fixed according to the total load of the network and the level of QoS that we aim to provide for elastic traffic.

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. We maintain however that fluid models provide "very valuable insight into the impact on performance of traffic characteristics" [7]. The insensitivity of average performance to the detailed statistical properties of connections is of great importance for network engineering. This property is likely to be maintained approximately even when accounting for disparities due to packet level behavior [7].

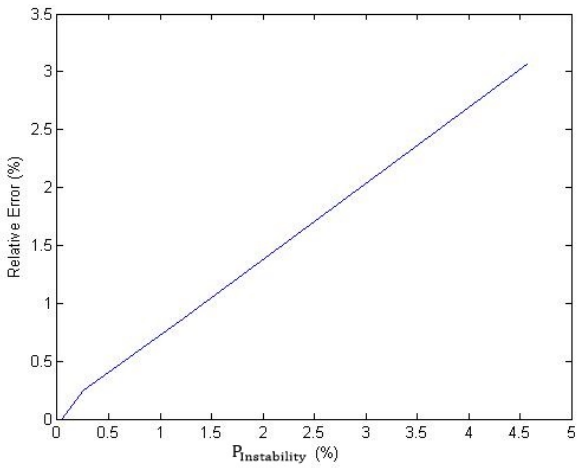

Figure 3. Relative error of the first WFQ queue in function of $P_{Instability}$

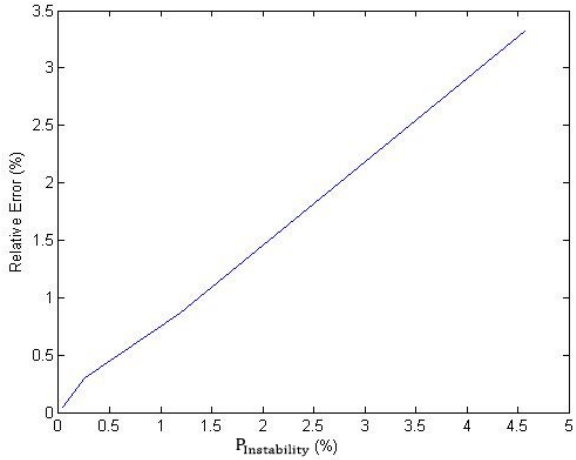

Figure 4. Relative error of the second WFQ queue in function of $P_{Instability}$

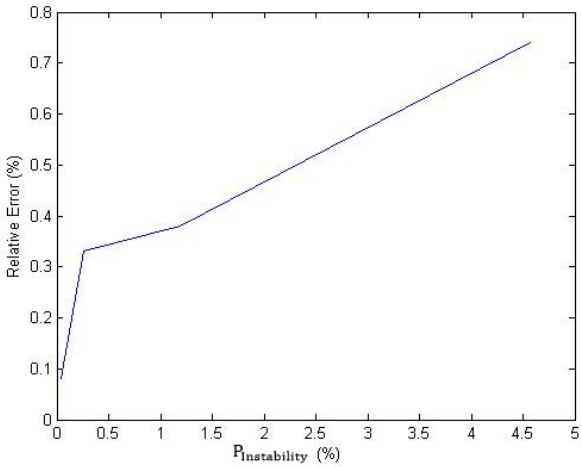

Figure 5. Relative error of the third WFQ queue in function of $P_{Instability}$

V. CONCLUSION

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 proposed a fluid model 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 to a specific weight. This amount of capacity can be viewed as a concatenation of a set of virtual links, and every virtual link is related to elastic flows with same maximum bit rate. Studying the performance of each elastic flow is, therefore, equivalent

to studying a single aggregated flow class passing on a link. So that, the result (8) is useful here in that it donates simply the mean number of flows for a single class. We had shown also that (8) can be a useful formula to approximate the mean number of flows for each class where many flow classes with identical transmission rate share a link. Detailed packet level simulations of TCP and UDP flows show that the proposed approximations work satisfactorily.

Another key result is that flow level performance metrics are 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] Brun, Olivier, Ahmad Al Sheikh, and J. Garcia. "Flow-level modelling of TCP traffic using GPS queueing networks." Teletraffic Congress, 2009. ITC 21 2009. 21st International. IEEE, 2009

[2] T. Bonald, L. Massouli´e, A. Prouti`ere, and J. Virtamo, "A queueing analysis of max-min fairness, proportional fairness and balanced fairness,"Queueing Syst. Theory Appl., vol. 53, no. 1-2, pp. 65–84, 2006

[3] Bonald, Thomas, and Jorma Virtamo. "Calculating the flow level performance of balanced fairness in tree networks." Performance Evaluation 58.1 (2004): 1-14.

[4] Bonald, T., Haddad, J. P., & Mazumdar, R. R. (2011, September). Congestion in large balanced multirate links. In Proceedings of the 23rd International Teletraffic Congress (pp. 182-189). International Teletraffic Congress.

[5] T. Bonald and A. Prouti`ere, "Insensitive bandwidth sharing in data networks," Queueing Syst. Theory Appl., vol. 44, no. 1, pp. 69–100, 2003.

[6] L. Massouli´e, "Structural properties of proportional fairness: Stability and insensitivity," Ann. Appl. Probab., vol. 17, no. 3, pp. 809–839, 2007.

[7] Bonald, Thomas, and James W. Roberts. "Congestion at flow level and the impact of user behaviour." Computer Networks 42.4 (2003): 521-536.

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

[9] Toumi, L. (2002). Algorithmes et mécanismes pour la qualité de service dans des réseaux hétérogenes (Doctoral dissertation, Institut National Polytechnique de Grenoble-INPG).

[10] Fred, S. Ben, et al. "Statistical bandwidth sharing: a study of congestion at flow level." ACM SIGCOMM Computer Communication Review. Vol. 31. No. 4. ACM, 2001.

[11] http ://www. enterprise.huawei.com

[12] Bonald,T., & Proutière, A. (2004, June). On performance bounds for the integration of elastic and adaptive streaming flows. In ACM SIGMETRICS Performance Evaluation Review (Vol. 32, No. 1, pp. 235-245). ACM.

[13] F. Kelly, A. Mauloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," J. of Oper. Res. Soc., vol. 49, no. 3, pp. 237–252, 1998.

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

[15] Kaufman, Joseph. "Blocking in a shared resource environment." Communications, IEEE Transactions on 29.10 (1981): 1474-1481.

[16] Delcoigne, Frank, Alexandre Proutiere, and Gwénaël Régnié. "Modeling integration of streaming and data traffic." Performance Evaluation 55.3 (2004): 185-209.

[17] Bonald, T., Proutiere, A., Roberts, J., & Virtamo, J. (2003). Computational aspects of balanced fairness. Teletraffic Science and Engineering, 5, 801-810.

[18] Niculae, Alexandra Mihaela. Mécanismes d'optimisation multi-niveaux pour IP sur satellites de nouvelle génération. Diss. 2009.

[19] Benameur, N., Fredj, S. B., Delcoigne, F., Oueslati-Boulahia, S., & Roberts, J. W. (2001, January). Integrated admission control for streaming and elastic traffic. In Quality of Future Internet Services (pp. 69-81). Springer Berlin Heidelberg.

[20] Malhotra, R., & van den Berg, J. L. (2006, November). Flow level performance approximations for elastic traffic integrated with prioritized stream traffic. InTelecommunications Network Strategy and Planning Symposium, 2006. NETWORKS 2006. 12th International (pp. 1-9). IEEE.

[21] Key, P., Massoulié, L., Bain, A., & Kelly, F. (2004, November). Fair Internet traffic integration: network flow models and analysis. In Annales des Telecommunications (Vol. 59, No. 11-12, pp. 1338-1352). Springer-Verlag

[22] Queija, R. N., Van den Berg, J. L., & Mandjes, M. R. H. (1999). Performance evaluation of strategies for integration of elastic and stream traffic