A Fluid Approach for Evaluating The Performance of TCP Traffic in the Presence of Real Time Traffic

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

A Fluid Approach for Evaluating The Performance of TCP Traffic in the Presence of Real Time Traffic

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*—Today, the Internet traffic is mostly dominated by an elastic data transfer. However, with the progressive development of the real time applications, it is anticipated that the streaming traffic will contribute a significant amount of traffic in the near future. By combining priority queuing with Class Based Weighted Fair Queueing (CBWFQ), The Low Latency Queueing (LLQ) is a very important router discipline that aims to provide the needed quality of service (QoS) for each traffic category. Priority queueing is used to guarantee delay constraints for real-time traffic, whereas CBWFQ is used to ensure acceptable throughput for traffic classes that are less sensitive to delay. In this paper, we focus on developing a fluid model to capture the relation between elastic traffic and the real time traffic in a LLQ system under a quasi-stationary assumption. Our analysis of the CBWFQ system is based on some numerical observations and relies on the conservation of the average total number of elastic flows carried by the elastic system. Detailed packet level simulations of TCP flows show the accuracy of our analysis. The results presented in this paper allows a rapid performance evaluation of TCP traffic circulating in the actual IP networks.**

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

I. INTRODUCTION

In the beginning, the Internet was designed for data processing applications where delays were not very important [1]. In the majority of cases, a best effort delivery service was enough, and when data was lost or corrupted, the Transmission Control Protocol (TCP) took care of the retransmission and recovery. Today these expectations have changed due to the growth of streaming and real time applications supported by this packet-based environment. Therefore, we shall distinguish two broad categories of Internet traffic: real time traffic and elastic traffic [2]. Real time traffic is generated by applications such as interactive voice, videoconference applications, online gaming and Voice over Internet Protocol applications (VoIP applications). These applications are quite sensitive for the delays and have strict bandwidth requirements for reliable operation. Elastic traffic, on the other hand, are generally transported by TCP, which results an elastic generated traffic. In fact, The TCP protocol relies on some specific mechanisms (slow start, congestion avoidance, ...) to control the congestion in the network and adapts the transmission rate to the available network resources [3]. This traffic is not only generated by web browsing and file transfer, but also by some streaming applications, which rely on TCP for transmission such YouTube for example [4,5]. There are several available mechanisms that attempt to provide the quality of service (QoS) needed for these two types of traffic. Packet traffic on the Internet could simply be handled on a First In, First Out (FIFO) policy. With a high enough bandwidth and under normal traffic conditions, this could be sufficient [6]. However, during emergency situations when the traffic demand is excessive, operators have to use some congestion management techniques to provide the differentiation between traffics categories [7]. In priority queueing (PQ), higher priority real-time 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 to each flow according to a specific weight for each flow [8]. Class-Based Weighted Fair Queueing (CBWFQ) extends weighted fair queueing to multiple user-defined traffic classes, rather than individual flows of traffic [9, 10]. Under CBWFQ, there are a buffer for each class of traffic, but, when one class of traffic is not utilizing its total bandwidth, the other classes are allowed to share this remaining bandwidth [6, 11]. The composition between these two policies of traffic management techniques (fixed priority policy and bandwidth sharing-based policy) is considered by many telecommunications equipment constructors like Cisco and Huawei [1]. 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 [12]. Priority queueing is used to satisfy the strict delay constraints for real-time traffic, whereas CBWFQ is used to ensure acceptable throughput for elastic traffic classes. The most important side of this is how the existing resources are shared during congestion times. 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 and appropriate software tools to predict and control the quality of service of different traffic flows [2].

In the literature, we can distinguish two types of models: the packet level models and flow level models [13]. The packet level defines the way in which packets are generated and transported during the communication [14]. The packet level models incorporate many details about the system (Round Trip Times, buffer size, etc.), but they generally consider a fixed number of persistent flows [15]. Flow-level models, in

978-1-5090-4663-8/17/$31.00 ©2017 IEEE

contrast, are idealized models that include random flow-level dynamics (arrivals and departures of flows) and use highly simplified models of the bandwidth sharing [15]. The complex underlying packet-level mechanisms (congestion control algorithms, packet scheduling, buffer management), at short-time scales, are simply then represented by a long-term bandwidth sharing policy between ongoing flows [16, 17]. 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]. We refer to class of flows as all flows of the same service between a source and a destination, having the same resources requirements.

This paper presents an analytical fluid approximation to estimate the performance characteristics of elastic traffics in the presence of real time traffic. In the next section, we present some related works and specify our original contribution. The third section is devoted to describe our model. In the fourth section, we present our approximations and discuss our analysis. Some results are presented and validated with NS 2 simulations in the last section.

II. Relates Works

The coexistence between TCP and real time traffic is a vexed problem [19]. Some authors have proposed that real time traffic should be TCP-friendly, so that it can fairly share network resources with elastic traffic [19, 20]. 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 [21-24], 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 streaming flows to not jeopardize the performance of ongoing flows. All these papers treated a limited system with two priority queues admitting that there is no differentiation in TCP traffic. Today, TCP is suitable to carry different traffic with different requirements of quality of service, and then, the traffic carried by TCP should be not treated in the same manner. A multi-queuing system combining priority queues for real time traffic and CBWFQ queues for elastic traffic is treated using simulation in [25]. The integration of real time and elastic traffic, under one packet-based environment, was mainly treated under a quasi-stationary assumption [21, 22, 24]. Under this assumption, we proposed in [1-3] an analytical model to evaluate the performance of elastic traffic under a LLQ system. The CBWFQ system is treated in these works under a decoupling approach, where the CBWFQ system was approximated by independent working servers with variable service rates.

In this paper, we aim to differently treat the CBWFQ queues under this LLQ system. The originality of this paper is to deal with the CBWFQ system as a one entity by exploiting some numerical results. The performance of TCP traffic are always studied under a quasi-stationary assumption

III. Model

We consider a single link with capacity $C(Mbit/Second)$shared by a random number of elastic and streaming flow classes. Let E be the set of these elastic flow classes, and S be the set of streaming flow classes. Streaming flows are mainly defined by their rate and their mean holding-time. For each streaming class-j flows ($j \in S$), we define:

- $\tau_j(Second)$:The mean holding-time of flows.
- $\mathrm{d}_j^{(s)}(Mbit/Second)$:The rate of each flow.

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

- $\sigma_i(Mbit/Flow)$:The mean volume transferred by flows.
- $\mathrm{d}_i^{(e)}(Mbit/Second)$: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 flows and $\lambda_i^{(e)}(Flows/Second)$ for elastic class-i flows. We refer to the product $\rho_i^{(e)} = \lambda_i^{(e)}\sigma_i(Mbit/Second)$ 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 $\rho_j^{(s)} = \lambda_j^{(s)}\tau_j(Flows)$.

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) . Let us denote by the vector $x^{(s)} = (x_j^{(s)})_{j\in S}$ (respectively $x^{(e)} = (x_i^{(e)})_{i\in E}$) the state of streaming classes (respectively the state of elastic classes). Let $\theta^{(e)} = \sum_{i\in E}\ \rho_i^{(e)}$ (respectively $\theta^{(s)} = \sum_{j\in S}\ d_j^{(s)}\rho_j^{(s)}$) be the elastic load (respectively the streaming load) offered to the capacity C.

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, at the entrance of the link, there is a LLQ queue combining a priority queue with a number of M CBWFQ queues. Let v_m, $1 \leq m \leq M$, be the weight of the CBWFQ queue number m. We assume that $\sum_{m=1}^{M}\ v_m = 1$.

The priority queue is devoted to streaming flows, which have strict bandwidth and delay requirements that can be met only if the link capacity is completely allocated to them. Streaming flows whose requirements cannot be guaranteed will be blocked rather than allow them into the system and jeopardize the performance of 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 [1, 22]. Elastic traffic is distributed throughout the CBWFQ queues. In this paper, we will suppose that all elastic classes have the same maximum bit rate:$d_i^{(e)} = d$ for all $i \in E$.

The performances of elastic traffic are essentially evaluated through the mean time necessary to transfer a document [17, 26, 27]. In the following, we evaluate performance in terms of average throughput per flow, 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 mean throughput of a flow for any class $i \in E$ is related to

the expected mean number of class- i flows in steady state, $(E[x_i^{(e)}])$, through the relationship [1]:

$$\gamma_i = \frac{\rho_i^{(e)}}{E[x_i^{(e)}]} \tag{2}$$

IV. Analysis

A. Quasy-stationnary approach

The quasi-stationary assumption supposes that the ratio $\lambda_j^{(s)}/\lambda_i^{(e)} (\forall i \in E, \forall j \in S)$ is small enough so that, in every state of $x^{(s)}$, the number of elastic flows evolves rapidly and attains a stationary regime.
Let n the quantity of the capacity C used by streaming flows in a state $x^{(s)}$:

$$n = \sum_{j \in S} x_j^{(s)} d_j^{(s)} \tag{3}$$

The steady distribution of $x^{(s)}$ is given by:

$$\pi(x^{(s)}) = \frac{1}{G} \prod_{j \in S} \frac{{\rho_j^{(s)}}^{x_j^{(s)}}}{x_j^{(s)}!} \tag{4}$$

With G is the normalization constant.

For each state n, we define the two following notations:

The remaining capacity for elastic traffic:

$$C^{(e)}(n) = C - n \tag{5}$$

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

$$A(n) = \sum_{x^{(s)}: \sum_{j \in S} x_j^{(s)} d_j^{(s)} = n} \pi(x^{(s)}) \tag{6}$$

The average throughput for each queue m is then given by:

$$\gamma_m = \sum_n \gamma_m(n) A(n) \tag{7}$$

Where:

$$\gamma_m(n) = \frac{\theta_m^{(e)}}{E_m^{WFQ}(n)} \tag{8}$$

Where $\theta_m^{(e)} = \sum_{i \in E_m} \rho_i^{(e)}$ is the elastic load offered to the queue m, $E_m^{WFQ}(n)$ is the mean number of flows traversing the queue m in a state n and assuming that $\theta^{(e)} < C^{(e)}(n) \ \forall n$.

B. Studying the CBWFQ system

The capacity $C^{(e)}(n)$ can be viewed as a concatenation between M virtual links. We note by E_m then the set of elastic flow classes traversing the virtual link number m.

The capacity $C_m^{(*)}(n) = v_m C^{(e)}(n)$ is principally dedicated to the flow classes of E_m , but if a part of this capacity remains no used, it will be shared between the other virtual links that need more resources to transmit its traffic in a better condition. Therefore, $C_m^{(*)}(n)$ can be viewed as the minimum capacity available to the virtual link m. In [11], we were studying this coupling approach by exploiting some numerical results to evaluate the performance of TCP traffic, having the same bit rate, under two CBWFQ queues. We showed that the average number of flows traversing a queue m (or the virtual link number m) could be approximated by:

$$E_m^{WFQ} = \frac{a}{v_m^{\alpha} + b} + c \tag{9}$$

Where:
α is numerically adjusted to 3 and:

$$a = \left(\frac{\theta_m^{(e)}}{\theta^{(e)}} E^{BE} - c\right)\left(\left(\frac{1}{M}\right)^{\alpha} + b\right) \tag{10}$$

$$b = \frac{\theta_m^{(e)}\left(\frac{B_m}{C^{(e)}(n) - \theta_m^{(e)}} - \frac{B}{C^{(e)}(n) - \theta^{(e)}}\right)}{(M^{\alpha}-1)\theta_{m^*}^{(e)}\left(\frac{B_{m^*}}{C^{(e)}(n) - \theta_{m^*}^{(e)}} - \frac{B}{C^{(e)}(n) - \theta^{(e)}}\right) - \theta_m^{(e)}\left(\frac{B_m}{C^{(e)}(n) - \theta_m^{(e)}} - \frac{B}{C^{(e)}(n) - \theta^{(e)}}\right)} \tag{11}$$

$$c = \frac{\theta_m^{(e)}}{\theta^{(e)}} E^{BE} + \frac{1+b}{1 - (\frac{1}{M})^{\alpha}} \theta_m^{(e)} \left(\frac{B_m}{C^{(e)}(n) - \theta_m^{(e)}} - \frac{B}{C^{(e)}(n) - \theta^{(e)}}\right) \tag{12}$$

With:
E^{BE}:The average total number of flows traversing the system.
M:The number of CBWFQ queues (In this case $M = 2$).
B_m:The congestion probability when v_m tends to 1.
B:The congestion probability in Best Effort system.
$m^* = (m \bmod 2) + 1$

In this section, we aim to provide a generalization of this approximation in order to calculate the performance of TCP traffic with more than two CBWFQ queues.

1) Evaluating the average total number of flows:

As we showed in [11], all our numerical results prove that, when the system is stable, the average total number of flows traversing the system approximately remains the same with or without the use of the CBWFQ mechanism.

Let $N(n) = \lfloor \frac{C^{(e)}(n)}{d} \rfloor$ be the maximum number of flows that can be have their maximum bit rate d in a state n. Above this limit, congestion occurs and flows equally share the capacity $C^{(e)}(n)$.
In each state n ,the average total number of flows is given then by:

$$E^{BE} = \frac{\theta^{(e)}}{d} + B \frac{\theta^{(e)}}{C^{(e)}(n) - \theta^{(e)}} \tag{13}$$

Where:

$$B = \frac{(\frac{\theta^{(e)}}{d})^{N(n)}}{N(n)!} + \frac{C^{(e)}(n)}{C^{(e)}(n) - \theta^{(e)}} \pi(0) \tag{14}$$

And:

$$\pi(0) = (\sum_{x=0}^{N(n)-1} \frac{(\frac{\theta^{(e)}}{d})^x}{x!} + \frac{(\frac{\theta^{(e)}}{d})^{N(n)}}{N(n)!} \frac{C^{(e)}(n)}{C^{(e)}(n) - \theta^{(e)}})^{(-1)} \quad (15)$$

Under Best Effort system, the average number of flows for each elastic class of flows $i \in E$ is given by:

$$E_i^{BE} = \frac{\rho_i^{(e)}}{\theta^{(e)}} E^{BE} \quad (16)$$

2) Evaluating the averge total number of flows for each queues:

When the weight of a CBWFQ queue number m, $m = 1..M$, tends toward 1, it can be considered as a priority queue [28]. So that, the load $\theta_m^{(e)}$ will be supposed to exploit all the residual capacity for the elastic traffic in a state n. Let $E^{WFQ}_{(m/v_m \to 1)}$ the average total number of flows traversing this queue in this case.

When $v_m = 1/M$,$m = 1..M$, the system equally share its resources. The system can be seen as a Best-Effort system. So that, we have:

$$E^{WFQ}_{(m/v_m \to 1/M)} = E_m^{BE} \quad (17)$$

With E_m^{BE} is the average total number of flows for all flow classes traversing the queue number m in a Best Effort system.

Another numerical key result is that when the weight of a queue m tends to zero, the average number of flows passing through this queue is obtained by:

$$E^{WFQ}_{(m/v_m \to 0)} = E^{BE} - K \quad (18)$$

Where K is the average total number of flows traversing the system regardless the traffic of the queue number m.

Exploiting (16), (17) and (18), the average total number of flows traversing a queue m,$m = 1..M$, can be approximated by (9) where:

$$a = \left(E_m^{BE} - c\right)\left(\left(\frac{1}{M}\right)^{\alpha} + b\right) \quad (19)$$

$$b = \frac{E_m^{BE} - E^{WFQ}_{(m/v_m \to 1)}}{(M^{\alpha} - 1)(E^{BE} - K - E_m^{BE}) - (E_m^{BE} - E^{WFQ}_{(m/v_m \to 1)})} \quad (20)$$

$$c = E_m^{BE} - \frac{1+b}{1 - (\frac{1}{M})^{\alpha}}(E_m^{BE} - E^{WFQ}_{(m/v_m \to 1)}) \quad (21)$$

E^{BE} is given by (13), E_m^{BE} is given by (16) by replacing $\rho_i^{(e)}$ with $\theta_m^{(e)}$, K is given by (13) by replacing $\theta^{(e)}$ with $\theta^{(e)} - \theta_m^{(e)}$ and $E^{WFQ}_{(m/v_m \to 1)}$ is given by (13) by replacing $\theta^{(e)}$ with $\theta_m^{(e)}$. It is easy to verify that for $M = 2$, we obtain the same expressions of a , b and c proposed in [11].

The average number of flows for each class traversing a queue m, $m = 1..M$, is obtained exploiting the relation (16) as follows:

$$E[x_i^{(e)}]_{(i \in E_m)} = \frac{\rho_i^{(e)}}{\theta_m^{(e)}} E_m^{WFQ} \quad (22)$$

3) Testing the accuracy of our analysis:

To examine the accuracy of the proposed analysis for the CBWFQ system, we consider a scenario of capacity $C = 20(Mbit/Second)$ shared by three TCP flow classes having the same maximum bit rate $d = 4(Mbit/Second)$. Let $M = 3$ and we assume then that each queue, is traversed by a class flows.

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 $x^{(e)}$: $d_i^{(e)}(x^{(e)}) \leq d_i^{(e)}$.

Figure (1) compares our approximation (9) with the exact result given by the numerical resolution of the Markov chain in term of E_1^{WFQ} in function of the weight assigned to it for different values of θ , for two different values of $\theta_1^{(e)}/\theta$.

For the numerical resolution, we assume that $v_2 = 3v_3 = 3/4(1 - v_1)$. For the second case when $\theta_1^{(e)}/\theta = 0.6$, we assume that the other elastic classes have the same contribution in the total elastic traffic.

In general, our approximation gives very good results. However, it seems to be more accurate when the distribution of the load is almost the same. It should be noted also that the results given by our approximation are more accurate when the system is far for the instability regime. When the load is close to the link capacity, the assumption of the conversation of the average total number of flows is not very accurate, and then the error rate increases a little bit.

V. Simulations and validity of analytical results

In this section, we aim to compare our analytical analysis with the real behavior of TCP traffic. We simulate then with NS 2 the case of a link of capacity $C = 100(Mbit/Second)$ 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 traversed by a single TCP class flows. Let $v_1 = 0.85$,$v_2 = 0.15$ and $v_3 = 0.05$.

Let $d_1^{(s)} = 10(Mbit/Second)$, $d_2^{(s)} = 20(Mbit/Second)$ and $d_3^{(s)} = 5(Mbit/Second)$. We assume that elastic traffic constitutes 50% of the system capacity and the streaming traffic varies from 10% to 30% of the capacity. The two streaming classes have the same load. In the same manner, we assume

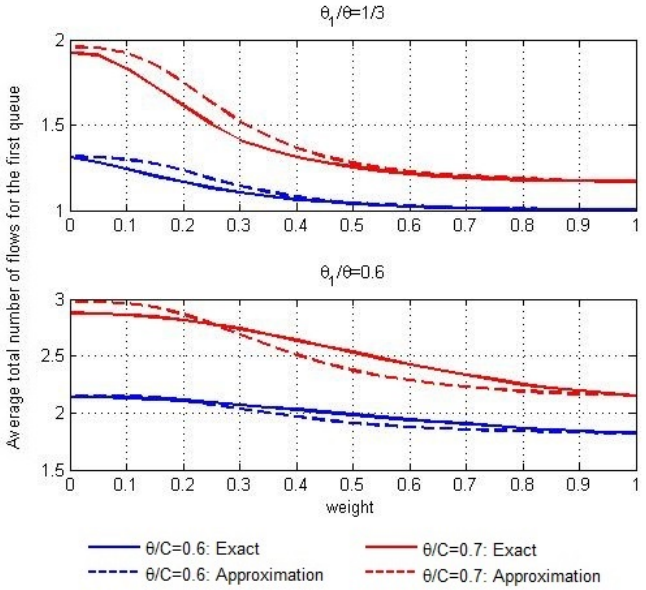

Fig. 1. Comparison between the analytical and the exact result of the averge number of flows for the first queue in function of the weight assigned to it

that all elastic flow classes have the same contribution in the total elastic load. For our simulation, we took $\lambda_j^{(s)}/\lambda_i^{(e)} = 0.1$ $(\forall i \in E, \forall j \in S)$ to guarantee the quasi-stationary assumption.

Each simulation point is the average of 10 simulation runs. In each simulation run, we simulate an hour and a half of dynamic arrival and departure of flows, to ensure that our system reaches the stability phase. Figure 2 evaluates the error rate between the approximate and simulation results in function of the contribution of streaming traffic in the total traffic circulating in the network. The error rate is defined as:

$$Error\ Rate = \frac{|Exact\ Result - Approximate\ Resul|}{Exact\ Result} \tag{23}$$

The error rate doesnt exceed 6% in all cases, which confirms our analysis. However, our approximation seems to be more accurate when the weight of the queue tends to one (first queue) or when it tends to zero (third queue). In fact, for these two limit cases, we have more accurate analytical results (see section IV.B.2)

It is important to note that the presented model is an ideal model. In practice, link bandwidth is not shared as precisely as assumed in the fluid models. The congestion control under TCP protocol actually relies on some complex algorithms (Slow Start, Congestion Avoidance...) that restrict the throughput of flows. However, we maintain that fluid models provide very valuable insight into the impact on performance of traffic characteristics [29]. The insensitivity of performance metrics, to the detailed statistical properties of traffic, provided under these models is of great importance for network engineering. This property is likely to be adopted even with these disparities due to packet level behavior [29].

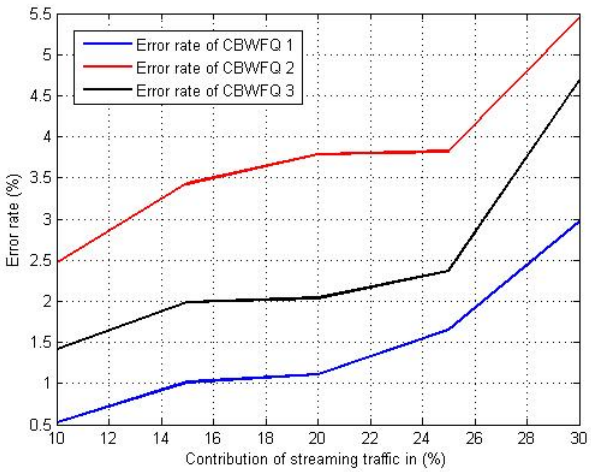

Fig. 2. Evaluation of the error rate between the approxiamte and simulation result for the three CBWFQ queues

VI. Conclusion

The performance of elastic traffic (TCP traffic) is principally studied in flow level. In this paper, 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 real time traffic, the remaining capacity is shared between the elastic traffic according to a specific weight.

The main contribution of this paper is the new methodology presented to treat the LLQ system, and more precisely the new approach given to study the CBWFQ system dedicated for the elastic traffic. In this context, this elastic system is mainly studied basing on some numerical observations for the evolution of the average number of flows for each queue in function of the weight assigned to it. The important numerical result is that the average total number of flows traversing the system remains the same with or without the use of the CBWFQ policy. Therefore, the results proven for a Best Effort system are very helpful to study the CBWFQ case. Detailed packet level simulations of TCP and UDP flows show that the proposed analysis work satisfactorily.

The approximation given for the CBWFQ system express the average total number of flows for a queue independently to the weights assigned to the other queues. Although that this approximation gives a good results, further work is needed to improve the accuracy of the proposed model.

Another key result is that the expressions of the performance metrics are insensitive to detailed traffic characteristics. This is particularly important for data network engineering since performance can be only predicted from a global estimate of the traffic circulating in the network. These results directly lead to simple rules of traffic engineering and robust methods of performance evaluation needed to control the current multimedia networks.

References

[1] Boussada, M. E. H., Frikha, M., Garcia, J. M. (2015, November). Flow level modelling of Internet traffic in Diffserv queuing. In Communica-

tions and Networking (COMNET), 2015 5th International Conference on (pp. 1-7). IEEE.

[2] 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: 219-33.

[3] Boussada, M. E. H., Garcia, J. M., and Frikha, M. 2016. Evaluation des Performances Bout en Bout du Trafic TCP Dans une Architecture de Reseau Multi Files Dattente. Presented at the 17th Congress of the French Society of Operations Research and Decision Support (ROADEF), February, Compiegne/France.

[4] Hsiao, P. H., Kung, H. T., Koan-Sin, T. A. N. (2003). Streaming video over TCP with receiver-based delay control. IEICE transactions on communications, 86(2), 572-584.

[5] Rao, A., Legout, A., Lim, Y. S., Towsley, D., Barakat, C., Dabbous, W. (2011, December). Network characteristics of video streaming traffic. In Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies (p. 25). ACM.

[6] Fischer, M. J., Masi, D. M. B. (2007). A Quantitative Analysis of the Voice and Data Quality of Service Problem. REVIEW, 66.

[7] Balogh, T., Medvecky, M. (2011). Performance evaluation of WFQ, WF 2 Q+ and WRR queue scheduling algorithms. In 2011 34th International Conference on Telecommunications and Signal Processing (TSP).

[8] Balogh, T., Medveck, M. (2012). Average bandwidth allocation model of WFQ. Modelling and Simulation in Engineering, 2012, 39.

[9] Al-Sawaai, A., Awan, I., and Fretwell, R. 2009. Performance Evaluation of Weighted Fair Queuing System Using Matrix Geometric Method. In NETWORKING 2009 (pp. 66-78). Berlin Heidelberg: Springer.

[10] Fischer, M. J., Masi, D. M. B., and Shortle, J. F. 2008. Simulating the Performance of a Class-Based Weighted Fair Queueing System. In Proceedings of the 40th Conference on Winter Simulation Winter Simulation Conference, 2901-8.

[11] Boussada, M. E. H., Garcia, J. M., Frikha, M. (2016). Numerical Key Results for Studying the Performance of TCP Traffic under Class-Based Weighted Fair Queuing System. Journal of Communication and Computer, 13, 195-201.

[12] Szilgyi, S., Almsi, B. (2012). A Review of Congestion Management Algorithms on Cisco Routers. Journal of Computer Science and Control Systems, 5(1), 103.

[13] Venkataramanan, R., Jeong, M. W., Prabhakar, B. A Flow-and Packet-level Model of the Internet.

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

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

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

[17] Bonald, T., Massoulie, 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.

[18] Alexandra Mihaela, N. (2009) Mcanismes d’optimisation multi-niveaux pour IP sur satellites de nouvelle gnration. Diss.

[19] Key, P., Massouli, 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.

[20] Bonald, T., Proutire, 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.

[21] Delcoigne, F., Proutiere, A., Rgni, G. (2004). Modeling integration of streaming and data traffic. Performance Evaluation, 55(3), 185-209.

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

[23] Malhotra, R., Van Den Berg, J. L. (2006, November). Flow level performance approximations for elastic traffic integrated with prioritized stream traffic. In Telecommunications Network Strategy and Planning Symposium, 2006. NETWORKS 2006. 12th International (pp. 1-9). IEEE.

[24] Ben Cheikh, H. (2016, January). Integration of streaming and elastic traffic: Modeling and Performance Analysis. In Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools (pp. 65-68). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[25] Geyer, F. (2015). End-to-End Flow-Level Quality-of-Service Guarantees for Switched Networks (Doctoral dissertation, Universittsbibliothek der TU Mnchen).

[26] Garcia, J. M., and Boussada, M. E. H. 2016. Evaluation des Performances Bout en bout du Trafic TCP Sous le rgime. Equite Equilibre. In Proceedings of the 11th Performance Evaluation Workshop, March, LAAS-CNRS, Toulouse/France. (sciencesconf.org:aep11: 92416).

[27] Bonald, T. and Prouti‘ere, A. (2003) Insensitive Bandwidth Sharing in Data Networks. Queueing Systems, Theory and Applications, 44, 69-100.

[28] Bockstal, C., Garcia, J., and Brun, O. 2004. Approximation Du Rgime Stationnaire Dun Systeme wfq. 6eme Rencontres Francophones Sur les Aspects algorithmiques des telecommunications (ALGOTEL2004), Batz-sur-Mer (France), 75, 76.

[29] Bonald, T., and James, W. R. 2003. Congestion at Flow Level and the Impact of User Behaviour. Computer Networks 42 (4): 521-36.