A QoS Index for IP Services to Effectively Support Usage-based Charging

Dario Di Sorte, Mauro Femminella, Gianluca Reali

Abstract—The goal of this letter is to present a model to compute a Quality of Service (QoS) index to characterize IP services. Then, we show how such a score may be used in a clear and flexible way for defining advanced usage-based tariff criteria to charge QoS guaranteed network services to address the dynamics of the expected future telecommunications scenario.

Index Terms—guaranteed IP service, usage-based tariff, numerical analysis.

I. INTRODUCTION

Today, the most common charging method in the Internet is based on the flat-rate model, i.e., subscribers pay a fee for accessing the network, independent of the actual use of the service. If the Internet is to become a network supporting differentiated application and transfer services, the deployment of advanced and expensive architectures to efficiently support Quality of Service (QoS), charging, billing, and accounting must be carried out [1]. In addition, the price definition is expected to greatly influence the selection of the network operator. For these reasons, usage-based tariffs must be applied. In fact, the flat-rate model will likely be inefficient from an economic point of view, since it does not enable charges to be made according to the type and to the perceived QoS. A number of pricing models have been proposed so far [2].

In this letter, we start from the tariff model to charge for IP guaranteed services already sketched in [3], which has the following characteristics:

- it charges the edge-to-edge service of a single administrative domain for a given class of traffic, described by Dual Leaky Bucket (DLB) parameters [4];
- it is on a per-call basis;
- it depends on (i) the QoS level of the transfer service offered, (ii) the amount of network resources afforded by the domain for such a service;
- it is based on the duration of the connection and the amount of traffic volume exchanged.

This tariff is based on the concept of virtual delay [3]. The virtual delay value is a comprehensive, all-inclusive appraisal of the QoS parameters characterizing the edge-to-edge transfer service within a domain. The novelty of this letter is the definition of a law to compute the virtual delay. Then, we show how it depends on the QoS level and on the amount of available network resources, and how it affects the tariff.

II. COMPUTATION OF THE VIRTUAL DELAY

In this section, our goal is to compute the value of the virtual delay that characterizes an edge-to-edge service offered by a domain, from technical considerations at the IP layer, and to show how such a value may be used for charging purpose.

We start from the following assumptions:

- flows entering the domain are regulated by DLBs, with parameters: peak rate, $P_S$, sustainable rate, $r_S$, and burst tolerance, $B_T$. These traffic parameters define the extreme traffic profile associated with the considered class of traffic that users are not allowed violating;
- the QoS of the port-to-port IP service provided to the specific traffic class is described by the following service parameters: the maximum transfer delay, $D_{M}$, the maximum delay jitter, $D_{jitter}$, and the loss probability, $P_L$, due to buffer overflow. We assume that these parameters can be negotiated between network operators and users;
- other QoS parameters, such as channel reliability, resilience and connection set-up time (network parameters), characterize the intrinsic quality of the network, do not depend on the specific flow, and cannot be negotiated;
- resource reservation and admission control functions are implemented to respect the service level agreement. Consequently, an amount of network resources is reserved for the flow from the ingress to the egress point of the domain; we are not interested in the QoS architecture, nor in the specific admission control scheme and resource reservation strategy, implemented within the network.

The total edge-to-edge delay includes transmission time at the source, propagation delay, processing and transmission time, and queuing delay at network nodes. Since delay jitter is mainly caused by queuing in nodes, it is possible to assume that $D_{jitter}$ is a component of the maximum transfer delay. In other words, the maximum queuing delay is equal to the maximum delay jitter. Moreover, in an equivalent “virtual” model, it is possible to compensate the packet loss probability, due to buffer overflow in nodes, by increasing the amount of buffer allocated to the flow. From this point of view, loss probability may be traded for queuing delay, and therefore it...
may be represented by a contribution in the virtual delay.

Our idea is that a network service may be modeled as a hypothesised equivalent service with the virtual delay, \( d \), which gives a measure of the QoS level: the higher the level of the service, the lower the value of \( d \). Consequently, at this point, our goal is to find the component of the virtual delay related to the loss probability, \( D_p \). Once this law has been found, the value of virtual delay is given by

\[
d(D_M, P_L) = D_M + D_p(P_L) .
\]

(1)

To find \( D_p(P_L) \), let us proceed associating a port-to-port service described by \( (D_M, D_{jitter}, P_L) \) with an equivalent node that introduces an (almost) constant delay \( D_C = D_{jitter} \) (jitter), a maximum queuing delay equal to \( D_{jitter} \) and a loss probability equal to \( P_L \). We call \( C \) the amount of capacity from the input port to the output port, and consequently \( B = C D_{jitter} \) is the buffer of the equivalent node.

In packet networks, the value of the bandwidth reserved in a node to a flow is generally called effective bandwidth, and its typical range lies within the interval \([r_S, P_S]\). It depends on traffic characteristics, the amount of network resources, and the requested performance [4]. The maximum buffer occupancy of a single flow, characterized by a set of DLB traffic descriptors, which feeds a buffer with unlimited size served at a transmission capacity \( c \) is given by:

\[
b = B_{TS} (P_S - c)/(P_S - r_S) .
\]

(2)

Since, in our case, \( D_{jitter} = B/C \), the pair effective buffer and effective bandwidth, \((b_p, c_p)\), to be assigned to each flow at the equivalent node in order to guarantee a service characterized by a maximum queuing delay equal to \( D_{jitter} \) and no packet losses, are given by [4]

\[
c_p = \frac{P_S B_{TS}}{D_{jitter} (P_S - r_S) + B_{TS}} ,
\]

\[
b_p = c_p D_{jitter} .
\]

(3)

Then, let \( c_p \) be the value of effective bandwidth and let \( b_p \) be the effective buffer corresponding to a service characterized by a queuing delay equal to \( D_{jitter} \) and by a value of packet loss probability equal to \( P_L \). We undertake to calculate the \((c_p, b_p)\) pair by using, without loss of generality, the novel approach illustrated in [4].

Our goal is to associate the network service with losses to a virtual service without losses, by increasing the effective buffer space allocated to the flow. From (2), once the amount of bandwidth \( c_p \) reserved to the flow has been determined, the amount of buffer needed to avoid losses would be equal to

\[
b_p = B_{TS} (P_S - c_p)/(P_S - r_S) .
\]

(4)

It means that the buffer allocated to the flow should be increased by a value equal to \( b_p - b_p \). This would imply an additional queuing delay. Consequently, the virtual component of delay associated to the loss probability is equal to

\[
D_p = (b_p - b_p)/c_p .
\]

(5)

This leads to the final equation of the virtual delay:

\[
d(D_M, P_L) = D_C + D_{jitter} + D_p = D_C + \frac{P_S - c_p}{P_S - r_S} B_{TS} .
\]

(6)

This QoS index depends on (i) the QoS level of the service, (ii) the traffic parameters describing the flow, and (iii) the amount of network resources (included in \( c_p \)). It might be also possible to differently weight the different contributions \( (D_C, D_{jitter}, D_p) \), depending on the pricing policy of the domain or the type of application service to be supported.

The virtual delay is the starting point to define a pricing law [3]. We consider a monotonic, not increasing function of the virtual delay, \( f(d) \), which associates the port-to-port transfer of an information unit with a technical measure, expressed in commodity units; the better the service the higher the number of commodity units. Each domain is free to choose the function which best fits its own requirements. The cost (or value) of the transfer of an information unit from an edge port \( A \) to an edge port \( B \) with a virtual delay \( d \) is

\[
S(d) = \alpha_{A\to B} f(d) ,
\]

where \( \alpha_{A\to B} \) is the cost of each commodity unit, which depends on (i) network parameters; (ii) the two points \( A \) and \( B \) (e.g., their distance); (iii) the policies of the relevant domain. Clearly, when the commodity is on the market, its price can fluctuate according to factors that are beyond technical considerations. For this reason, we define the price of the transfer of an information unit as the quantity

\[
P(d) = \gamma S(d) = \gamma \alpha_{A\to B} f(d) ,
\]

where \( \gamma \) is a price variation factor that accounts for market fluctuations, and \( \beta = \alpha_{A\to B} \) is the market commodity price, that is the price per commodity unit.

Let \( T \) be the duration of a connection and \( t_0 \) its starting time. The per-call tariff to charge for the service offered to a flow entering the domain with an instant bandwidth \( B_{inst}(t) \) is

\[
Q = \beta f(d) \int_{t_0}^{t_0+T} \max[B_{res}(t) - B_{res}, 0] dt + \beta f(d) B_{res} T ,
\]

(7)

where \( B_{res} \) is the bandwidth value charged on a per-time basis. The tariff consists of a component depending on the duration time of the connection and a component depending on the amount of traffic volume exchanged. The weights of the two components can be arbitrarily set, by varying \( B_{res} \), which, in this view, may be regarded as a tunable knob. It is worth noting that if the value of \( B_{res} \) increases, the price charged increases as well, unless \( B_{inst}(t) \geq B_{res}, \forall t \). The value of \( B_{res} \) may range from zero to the peak rate of the flow. It is our opinion that the most reasonable choice is to set \( B_{res} \) equal to the amount of reserved bandwidth \( c_p \), since, in this way, operators charge users in proportion to the reserved resources, and protect themselves against unfair users behavior.

Then, it is clear that the tariff strongly depends on the instant bandwidth of the flow entering the domain (note that we are assuming that the accounting devices operate at the ingress of the domain). From (7), it is easy to verify that the maximum tariff (relevant to ON/OFF DLB output process) and the minimum tariff (when \( B_{inst}(t) \leq c_p, \forall t \)) are, respectively:
\[ Q_{\text{max}} = T[\beta (c_p (1 - r_3 / P_3) + r_2)] f(d) ; \]  
\[ Q_{\text{min}} = T[\beta c_p] f(d) . \]  

It is worth noting that the tariff model is such that the charge is higher for a bursty transmission rate, and this is correct since it is well known that bursty flows stress network resources more than flows with a smoothed transmission rate.

### III. Numerical Results

In this section, our goal is to provide numerical results regarding the sensitivity of the model to system parameters.

In our analysis, we assume the function \( f(d) \) to be equal to an exponential function with parameter \(-m\), with \( m \geq 0 \). The price per-commodity unit \( \beta \) is set to \( 1/100000 \) $, where $ is a given currency unit (e.g., cents of US Dollar or Euro).

For the specific case of a Voice over IP (VoIP) service, we assume that flows are described by the following DLB parameters: \( P_S = 32 \text{ Kbps} \), \( r_S = 13.6 \text{ Kbps} \), and \( B_{150} = 5300 \text{ bytes} \). The target performance level of the guaranteed transfer service is specified as follows: \( D_M = 150÷200 \text{ ms} \), \( D_M = 140 \text{ ms} \) and \( D_M = 10÷60 \text{ ms} \), and \( P_L = 10^{-3} \). The minimum system edge-to-edge capacity is set at 2 Mbps.

In Fig. 1(a), we show the virtual delay as a function of packet loss probability, with \( D_M \) set as a parameter. As expected, a better service in terms of lower values of packet loss probability is mapped into a smaller value of \( d \). Since \( f(d) \) is decreasing, and the effective bandwidth clearly increases when the packet loss probability decreases, from (8) and (9), it results that improved performance has a higher price. In Fig. 1(b), the maximum and minimum per-time charges are plotted as a function of the virtual delay (evaluated in Fig. 1(a)) with \( m \) as a parameter; \( D_M \) is fixed to 175 ms, and \( C \) is equal to 2 Mbps. Clearly, the value of \( m \) in \( f(d) \) must be set in order to tune the influence of the QoS level (i.e., the virtual delay) on the tariff charged to network users. In other words, \( m \) has to be increased or decreased to accentuate or mitigate, respectively, the price difference between different network services.

When the transmission capacity \( C \) between two edge nodes varies, we have to analyze the effects of system scaling on the virtual delay and then on the applied tariffs. In Fig. 2(a), we obtain the behavior of the virtual delay as a function of \( C \) ranging from 2 to 20 Mbps, with \( D_M \) set as a parameter, and \( P_L = 10^{-3} \). It is evident that the virtual delay is a function that increases with the value of the system capacity. Consequently, as shown in Fig. 2(b), the maximum and minimum tariffs decrease with the value of \( C \). To explain these results, let us consider the common knowledge that increasing the system capacity means increasing the statistical multiplexing gain. This means that the effective bandwidth \( c_p \) associated with a flow decreases. The effect of this phenomenon is evident in both the virtual delay, which increases when \( c_p \) decreases (refer to (6)), and in the per-commodity unit per-time unit expressions within square brackets in (8) and (9), which decrease with \( C \). Therefore, since both \( f(d) \) and \( c_p \) decrease when \( C \) increases, then an operator which owns a bunch of transmission resources can lower the tariff for a given service quality. Therefore, this result is consistent with both the "economies of scale" concept (economic point of view), and the fact that the statistical multiplexing gain increases with \( C \) (technical point of view).

### IV. Conclusions

We have shown an innovative way to charge for improved IP services, based on a QoS index, namely virtual delay. By using this parameter, we have built a per-call, usage-based tariff. The dependence of the tariff on the QoS level may easily be tuned, and an operator with a large amount of resources can lower costs in a flexible and clear way.

### References
