Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2010, Article ID 961690, 8 pages
doi:10.1155/2010/961690

Research Article

Reference Chaser Bandwidth Controller for Wireless QoS Mapping under Delay Constraints

M. Marchese and M. Mongelli

Department of Communications, Computer and Systems Science (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy

Correspondence should be addressed to M. Marchese, mario.marchese@unige.it

Received 14 January 2010; Accepted 11 October 2010

Academic Editor: Dilip Krishnaswamy

Telecommunications networks are composed of functional layers acting in cascade. Quality of Service (QoS) derives from the action of each layer that must assure a specific level of quality to the upper layer in terms of performance parameters (e.g., loss, delay, jitter of the packets). Appropriate algorithms are needed to compute the bandwidth necessary so to assure the requested QoS when information is transferred from one layer to the next one below. This paper proposes a scheme that adapts the bandwidth to be allocated to a buffer which conveys heterogeneous traffic (both concerning traffic sources and QoS requirements) in a layer-in-cascade model. The proposal is focused on delay constraints. The proposed algorithm is based only on measures and does not use closed-form expressions, a priori information about traffic statistical properties, and assumptions about buffer dimension. Simulation results show the reliability of the approach in comparison with other techniques at the state of the art, thus corroborating the application of the algorithm for a large set of operative situations, including fading conditions.

1. Introduction

Modern telecommunication networks are composed of devices which act through layered protocol stacks. If a specific Quality of Service (QoS) is required, the interface among the layers must be able to transport the request and possibly the answer so to create a dialogue between the two layers. The overall QoS depends on the QoS achieved at each layer of the network, and it is based on the services offered at the layer interfaces. The vertical interaction between layers in cascade is defined as “Vertical QoS Mapping” [1]. The paper takes the TI-SAP (Technology Independent—Service Access Point) approach as reference [1, 2]. The original protocol architecture has been proposed by ETSI [3] for the access points to a Broadband Satellite Multimedia (BSM) network portion and specified in [4–6]. The idea proposed in [1] and developed in [2] is to extend the concept of functional independence between physical interfaces and upper layers through the separation of Technology Dependent (TD) and Technology Independent (TI) layers and the definition of a generic interface called TI-SAP (Technology Independent–Service Access Point). The aims of the TI-SAP are (1) to get a formal separation between the functional layers that use specific hardware/software solutions at layer 2 and layer 1, defined as TD layers, and often covered by patents and the layers that implement the layers above layer 2, such as IP and upper layers, defined as TI layers; (2) to establish a common interface through which TI and TD layers can communicate without affecting the specific TD layers implementation. In conformance with ETSI choices [3], TI, in this paper, is associated to the IP technology (layer 3), whose data plane and control plane should be designed independently of the solutions applied at layers 2 and 1 (TD), which depend on the specific technology in play [5, 6], but having available a proper interface (TI-SAP) composed of a set of primitives through which the TD layers provide a given service to the TI layer. In other words and in more detail on QoS, the local implementation of the QoS inside TD layers should be transparent to upper layers but TD layers should offer precise performance guarantees to TI layer through the TI-SAP. For example, when a host (e.g., an IP-based smartphone) performs a physical technology change from one wireless medium (UMTS) to another one (WiFi), the IP (TI) layer should be practically unaware of the underlying technology change and no QoS degradation should be experienced during the change.

In this generalized framework, it is important to get a model to describe the action of each layer. The proposal of this paper is to model each layer as a group of queues, as done in [4–6], so that the communication between adjacent functional layers may be described through a cascade of groups of queues. The queue model allows describing the problems of vertical QoS mapping and to formally introduce a bandwidth allocation adaptation scheme called RCBC (Reference Chaser Bandwidth Control), whose aim is to dynamically adapt the bandwidth at layer 2 so that TD layers can provide a given service in terms of delay to TI layers.

The remainder of the paper is organized as follows. Section 2 sets the application framework for the algorithm and formalizes the vertical QoS mapping as a cascade of buffers. Section 3 describes a reference framework for dynamic bandwidth adaptation for layers in cascade. Section 4 introduces the RCBC bandwidth adaptation scheme, and Section 5 outlines some possible alternatives to the algorithm proposed. Section 6 shows the simulation results. Section 7 contains conclusions and some ideas for future work.

2. Cascade-of-Queues Model for Adjacent Layers

As said, the idea is to model each layer through groups of queues, similarly as done in [4–6]. The number of queues must be large enough to support the desired QoS model. In this framework, there are three problems arising from the action of layers in cascade [1]. (1) Change of information unit, which implies additional information (overhead) and bandwidth update, when information passes from the upper to the lower layers. (2) Aggregation of heterogeneous traffic: as outlined in [7], typically the number of queues decreases from the upper (TI) to the lower (TD) layers for efficiency and speed needs. It means that the traffic may need to be aggregated when it flows down from a layer to the adjacent one. The bandwidth at lower (TD) layer must be adapted consequently. (3) Fading affect: many transmission environments, such as satellite and wireless links, need to tackle time varying channel conditions due to fading. The three problems presented above can be seen jointly. The overall cascade-of-queues model is shown in Figure 1: N buffers are available at upper layer (TI in Figure 1) and identified through the index $i = 1,\dots,N$; one queue is available at lower (TD) layer. The bandwidth assigned to each buffer so to provide a given quality of service to the flows entering the buffer is identified by $R_{\text{id}}^{\text{TI}}(t)$ at TI layer and by R^{TD} at TD layer. From the mathematical viewpoint, the fading effect may be modelled as a reduction of the bandwidth actually "seen" by the TD buffer. The reduction is represented by a stochastic process $\phi(t)$. At time t, the "real" service rate $R_{\text{real}}^{\text{TD}}(t)$ (available for data transfer) is $R_{\text{real}}^{\text{TD}}(t) = R^{\text{TD}}(t) \cdot \phi(t), \phi(t) \in [0,1]$, where time dependency is explicitly indicated to enforce the concept of time varying channel conditions. There are N traffic classes, one for each TI buffer. $a_i(t)$ is the *input rate* process of the ith traffic class and $a(t)$ the aggregate process of all $a_i(t)$, $i = 1,\dots,N$. Bandwidth measure unit is [packet/s]. The key point is bandwidth adaptation, which is very challenging both from theoretical and practical viewpoint. Referencing to Figure 1, it means to dimension the bandwidth $R^{\text{TD}}(t)$ at TD layer so that the service is transparently guaranteed to the upper layer.

A practical interpretation of bandwidth $R^{\text{TD}}(t)$, assigned to the buffer at layer 2, in a wireless system may be given by using an a priori arbitrated channel to avoid collisions as happens in TDMA, where the channel is divided into frames and each frame is divided into time slots. A master station (a network control center) manages the global slots assignment to all the stations as happens in WiFi, WiMax, and DVB-RCS/S2 technology. The number of time slots assigned to a wireless station reflects the bandwidth allocated to that station. The slots organization should change dynamically during the lifetime of sessions as a function of the traffic load, QoS requirements and channel conditions. In this context, being R^{TD} the bandwidth needed to a specific wireless station at layer 2 to assure a given performance, R^{TD} may be transmitted to the master station to communicate the minimum bandwidth need of that specific station. The master station will allocate bandwidth (slots) to the stations proportionally to the received requests. For example, if there are N wireless stations including the master, each generic station j requires R_j^{TD} bandwidth to the master. The master will provide bandwidth to each station, for instance proportionally to requests R_j^{TD} or by following another strategy, and respecting the overall bandwidth channel constraint. Obviously if the layer 2 of each station should implement multiple queues, the example still holds, but the master station should allocate bandwidth to each single queue of the jth station and the minimum bandwidth R^{TD} should be computed for each of them. In short, R_j^{TD} is the minimum bandwidth need for station j and can be used to drive bandwidth assignments within a wireless system architecture, whose details are out of the scope of this paper and possibly object of future research. The paper is focused on the computation of R_j^{TD} for one queue of one generic station j, obviously dropping the index of j.

Bandwidth allocation is a widely treated subject in the literature. Most of the schemes are based on the concept of equivalent bandwidth (EqB), which is defined as the minimum service rate to be provided to a traffic buffer to guarantee a certain degree of QoS in terms of objective parameters (e.g., packet loss, delay, jitter). EqB techniques are usually obtained analytically for homogeneous traffic trunks, with respect to a single QoS constraint, and are heavily based on the knowledge of traffic features, which are mathematically modelled. The complexity of the overall input flow process $a(t)$ entering the TD layer in the vertical QoS mapping model described above makes hardly applicable the bandwidth allocation algorithms that use mathematical models of the flow process. The flow accessing the TD queue comes from the actions of format change, traffic aggregation and fading affect, which modify the original features of the

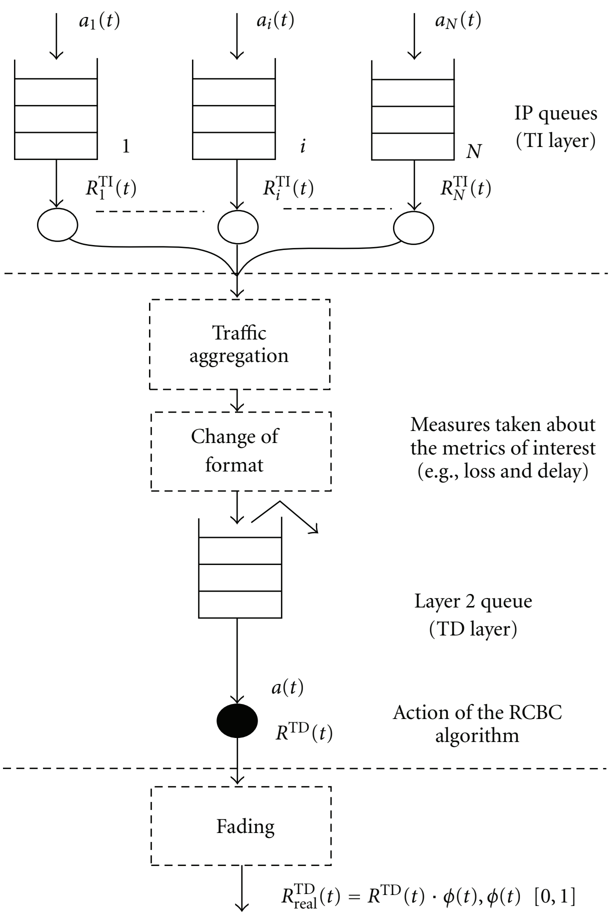

Figure 1: Cascade of queues model for adjacent layers.

flows $a_i(t)$ that enter the TI layer. The resulting flow is so complex that can hardly be analytically modelled.

3. Dynamic QoS Mapping for Layers in Cascade

The basic idea in Vertical QoS Mapping is to allocate the bandwidth periodically at TD layer after receiving the QoS constraints through TI-SAP primitives [5–7]. It may be generically applied to any decisional scheme. Time variable t_k identifies the reallocation instants. Index k is a progressive integer.

The bandwidth $R^{\mathrm{TD}}(t_k)$ allocated at the instant t_k may depend on the bandwidth allocated at previous instants $\{t_{k-1}, t_{k-2}, t_{k-3}, \dots, t_{k-D}\}$, where $D$is the depth of the allocation scheme memory, and on an information vector $I = \{i(t_k), i(t_{k-1}), i(t_{k-2}), i(t_{k-3}), \dots, i(t_{k-D})\}$. The latter may be composed of information about the TD buffer and/or, simply by the error $e(t_k)$, which is defined as the difference between the minimum bandwidth that guarantees the QoS constraints in the interval $T_k = [t_{k-1}, t_k]$, which is known at t_k, and the bandwidth allocated at t_{k-1}, which has given origin to the performance in the interval T_k.

More formally, if $R_{\mathrm{thr}}^{\mathrm{TD}}(t_k)$ is the bandwidth, computed at t_k, which would have been needed to assure the QoS constraints in T_k, the error in t_k is defined as $e(t_k) = R_{\mathrm{thr}}^{\mathrm{TD}}(t_k) - R^{\mathrm{TD}}(t_{k-1})$. The multiplicative fading parameter $\phi(t)$ shown in the previous section is not included here only to simplify the notation, but all the comments are still valid including fading, as done in the results in the following.

Being $F(\cdot)$ a generic function, a possible generic representation of the allocated bandwidth is

$$R^{\mathrm{TD}}(t_k) = F\Big(R^{\mathrm{TD}}(t_{k-1}), \dots, R^{\mathrm{TD}}(t_{k-D}), i(t_k), \dots, i(t_{k-D})\Big), \tag{1}$$

where, as said above, $i(t_k)$ may be simply $e(t_k)$, for all k. The $F(\cdot)$ law decides which and how previous allocations $[R^{\mathrm{TD}}(t_{k-1}), \dots, R^{\mathrm{TD}}(t_{k-D})]$ and information $[i(t_k), \dots, i(t_{k-D})]$ are used to obtain the new bandwidth allocation at instant t_k.

An interesting subclass of bandwidth allocations algorithms may be described through the allocation in

$$R^{\mathrm{TD}}(t_k) = F\Big(R^{\mathrm{TD}}(t_{k-1}), i(t_k), i(t_{k-1}), i(t_{k-2})\Big). \tag{2}$$

It includes the bandwidth allocations based on conventional discrete PID controller, which may be generically written as [8]:

$$\begin{aligned} R^{\mathrm{TD}}(t_k) = R^{\mathrm{TD}}(t_{k-1}) + w_k(t_k) \cdot e(t_k) \\ + w_{k-1}(t_k) \cdot e(t_{k-1}) + w_{k-2}(t_k) \cdot e(t_{k-2}). \end{aligned} \tag{3}$$

The details of the weights $w_k(t_k)$, $w_{k-1}(t_k)$, $w_{k-2}(t_k)$ and their possible computation may be found in [8] and other references related to discrete PID. To deal with nonlinear time-varying processes, also the weights may be time varying and dependent on the information vector I. A more restricted algorithm subclass is represented by the schemes where $D = 1$:

$$R^{\mathrm{TD}}(t_k) = F\Big(R^{\mathrm{TD}}(t_{k-1}), i(t_k)\Big). \tag{4}$$

A corresponding bandwidth allocation update is reported in

$$R^{\mathrm{TD}}(t_k) = R^{\mathrm{TD}}(t_{k-1}) + w_k(t_k) \cdot e(t_k). \tag{5}$$

If the requirement is that the bandwidth allocation algorithm does not use any a priori information about traffic statistical properties, any assumption about buffer dimensions, and any closed-form expression of the involved variables, a possible solution is to use only measures of the ongoing processes. The weight $w_k(t_k)$ acts either as a reducer or as an amplifier of the bandwidth need estimation and may be dynamic over time.

4. RCBC

The aim here is to dynamically dimension the weight $w_k(t_k)$ every t_k so to chase the given performance thresholds. The reference environment has been described in Section 2 and shown in Figure 1. The quantities $R_i^{\mathrm{TI}}(t)$, $R^{\mathrm{TD}}(t)$, $a(t)$, $a_i(t)$, $i = 1, \dots, N$ are defined in Section 2. So, there are N traffic classes. Traffic conveyed towards a single buffer is modelled through a *Stochastic Fluid Model* [9, 10]. $a(t)$ is supposed to be ergodic for now, so that a single realization is representative of the entire process. This assumption will be relaxed later. There is no knowledge of $a_i(t)$ processes as well as of the aggregate process $a(t)$. The only information about $a_i(t)$ and $a(t)$ may be got through real measures. The metric used here is delay; the application of the control algorithm to loss has been presented in [11]. The following additional definitions are necessary. They are all applied at the TD buffer.

$R_i^{\text{delay}}(R^{\text{TD}}(t), t)$ is the delayed packets rate process of the ith traffic class, that is, the rate of the packets which arrive with a delay over a given threshold d_{thr} [s].

$\text{Delay}_{i,\text{thr}}(t)$, which can also vary over time, is the probability that the delay for class i be over a given threshold (e.g., the probability that packets are delayed more than 50 ms). It is the performance reference that derives from the *Service Level Agreement* (SLA) of class i, and it is received from the TI layer by an internal signaling or it is set a priori by the network operator. $R_{i,\text{thr}}^{\text{delay}}(t) = a_i(t) \cdot \text{Delay}_{i,\text{thr}}(t)$ is the delayed packet rate process that can be tolerated (the delayed packet rate threshold) of the ith traffic class [packet/s].

The average value of $R_i^{\text{delay}}(R^{\text{TD}}(t), t)$ and $R_{i,\text{thr}}^{\text{delay}}(t)$ are

$$\begin{aligned} \overline{R}_i^{\text{delay}} &= \lim_{\tau \to \infty} \frac{1}{\tau} \int_{\tau} R_i^{\text{delay}}\left(R^{\text{TD}}(t), t\right) dt, \quad i = 1, \dots, N, \\ \overline{R}_{i,\text{thr}}^{\text{delay}} &= \lim_{\tau \to \infty} \frac{1}{\tau} \int_{\tau} R_{i,\text{thr}}^{\text{delay}}(t) dt, \quad i = 1, \dots, N. \end{aligned} \tag{6}$$

The aim is to provide the minimum TD buffer service bandwidth $R_{\text{opt}}^{\text{TD}}$ so that the maximum quadratic distance between $\overline{R}_i^{\text{delay}}$ and $\overline{R}_{i,\text{thr}}^{\text{delay}}$ is minimized. It corresponds to define the following optimization problem:

$$\begin{aligned} R_{\text{opt}}^{\text{TD}} &= \arg \min_{R^{\text{TD}}} R_{\Delta}^{\text{delay}}\left(R^{\text{TD}}\right), \\ R_{\Delta}^{\text{delay}}\left(R^{\text{TD}}\right) &= \operatorname*{Max}_i \left[\overline{R}_i^{\text{delay}} - \overline{R}_{i,\text{thr}}^{\text{delay}}\right]^2. \end{aligned} \tag{7}$$

Being the involved stochastic processes unknown, problem (7) is solved by taking measures over the given kth *observation horizon* (OH), $T_k = [t_{k-1}, t_k]$, $k = 1, 2, \dots$, and performing a sequence of bandwidth reallocations, $R^{\text{TD}}(t_k)$, $k = 1, 2, \dots$, for each T_k, as introduced in Section 3, based on the gradient method so that $R^{\text{TD}}(t_k) \xrightarrow{k \to \infty} R_{\text{opt}}^{\text{TD}}$. The quantities $\overline{R}_i^{\text{delay}}$ and $\overline{R}_{i,\text{thr}}^{\text{delay}}$ are averaged over each OH, giving origin to the quantities (8) and (9). Being used to solve the optimization problem (7), $\widehat{R}_i^{\text{delay},k}$and $\widehat{R}_{i,\text{thr}}^{\text{delay},k}$ must be representatives of the average values $\overline{R}_i^{\text{delay}}$ and $\overline{R}_{i,\text{thr}}^{\text{delay}}$ for all $i = 1, \dots, N$ and for all k

$$\widehat{R}_i^{\text{delay},k} = \frac{1}{T_k} \int_{T_k} R_i^{\text{delay}}\left(R^{\text{TD}}(t), t\right) dt; \quad i = 1, \dots, N, \tag{8}$$

$$\widehat{R}_{i,\text{thr}}^{\text{delay},k} = \frac{1}{T_k} \int_{T_k} R_{i,\text{thr}}^{\text{delay}}(t) dt; \quad i = 1, \dots, N. \tag{9}$$

These quantities can be easily computed in real time on the TD data plane as they correspond to the amount of actual (8) and ideal (9) delayed packets (above the SLA threshold) over the OH, for each traffic class.

Bandwidth is adapted through the algorithm in Algorithm 1, called *Reference Chaser Bandwidth Control* (RCBC). It increases the bandwidth of the weighted needs sum in case there is at least one traffic class demanding bandwidth and decreases the bandwidth of the minimum weighted excess in case all classes show they have too much bandwidth. step_k is the gradient stepsize. Modifications to RCBC are possible by using the maximum bandwidth need and bandwidth excess as well as the sum of estimated bandwidth excesses or combinations of them but, on the one hand, the performance differences among them (measured through ad hoc simulations not shown here) are not outstanding; on the other hand using the sum of bandwidth needs when adding and the minimum bandwidth excess when dropping is more conservative and safer than the alternatives.

Condition $(\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}) \geq 0$ means that the allocated bandwidth needs to be increased. Condition $(\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}) < 0$ states the opposite.

Derivatives $\partial \widehat{R}_i^{\text{delay},k} / \partial R^{\text{TD}}$, both negative (this is the motivation for the negative sign before the quantities in Algorithm 1), represent the sensitivity of loss and delay to infinitesimal variations of the rate serving the buffer. Intuitively they depend on the speed with which the system passes from an empty to a full state. They can be obtained by observing the buffer state evolution (as introduced in [9]) within each OH, which is divided into N_{T_k} busy periods identified by the variable bp. A *busy period* is simply a period of time in which the buffer is not empty.

$\partial \widehat{R}_i^{\text{delay},k} / \partial R^{\text{TD}}$ is approximated by

$$\frac{\partial \widehat{R}_i^{\text{delay},k}}{\partial R^{\text{TD}}} \cong \left\{ \begin{array}{c} -\dfrac{1}{T_k} \displaystyle\sum_{bp=1}^{N_{T_k}} \left[{}^{i}at_{T_k}^{bp}\left(R^{\text{TD}}(t_{k-1})\right) - {}^{i}ld_{T_k}^{bp}\left(R^{\text{TD}}(t_{k-1})\right)\right], \\ \text{if there at least one delayed packet within the OH} \\ 0, \ \text{otherwise} \end{array} \right\}. \tag{10}$$

$[{}^{i}at_{T_k}^{bp}(R(t_{k-1})) - {}^{i}ld_{T_k}^{bp}(R(t_{k-1}))]$ is the contribution to information delay (over the acceptable threshold) of the ith traffic class for the busy period bp within T_k, $k = 1, 2, \dots$. ${}^{i}at_{T_k}^{bp}$is the arrival time of the first packet of service class i within the busy period bp. ${}^{i}ld_{T_k}^{bp}$ is the time when the last delayed packet of class i arrives during bp. Approximation (10) is introduced in this paper. In practice bandwidth update RCBC in Algorithm 1 is in the form of (5). The derivatives multiplied by step_k may be considered a form of the weight $w_k(t_k)$. The difference $[\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}]$ is the missing (or the excess of) bandwidth, that is, a representation of the error $e(t_k)$. Like (8) and (9), the quantities in (10) are computed in real time in the TD data plane. The overall computational effort of RCBC is very small as the involved variables (number of delayed packets, size of busy periods)

```
If $(\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}) > 0$ for at least one $i$ then {
$\Delta_i^{\text{delay}}(t_k) = \begin{cases} -2 \cdot (\partial \widehat{R}_i^{\text{delay},k} / \partial R^{\text{TD}})|_{R^{\text{TD}}=R^{\text{TD}}(t_{k-1})} \cdot [\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}], & \text{if } [\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}] \geq 0 \\ 0, & \text{otherwise} \end{cases}$
$R^{\text{TD}}(t_k) = R^{\text{TD}}(t_{k-1}) + \text{step}_k \cdot \sum_{i=1}^{N} [\Delta_i^{\text{delay}}(t_k)]$
}
else {
$\Delta_i^{\text{delay}}(t_k) = -2 \cdot (\partial \widehat{R}_i^{\text{delay},k} / \partial R^{\text{TD}})|_{R^{\text{TD}}=R^{\text{TD}}(t_{k-1})} \cdot [\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}]$
$\text{Min}\,\Delta^{-}(t_k) = \Delta_j(t_k), \; j = \arg\min_i \{|\Delta_i(t_k)|\}$
$R^{\text{TD}}(t_k) = R^{\text{TD}}(t_{k-1}) + \text{step}_k \cdot \text{Min}\,\Delta^{-}(t_k)$
}
```

Algorithm 1: RCBC.

can be easily updated during the network evolution; no packet-by-packet analysis of the buffer state is needed.

5. Other Techniques for Bandwidth Allocation

The following techniques are used for performance comparison with RCBC. The aim is to highlight RCBC control reliability with respect to other mechanisms taken from the literature.

5.1. Proportional Integrative Derivative (PID) Controller. The majority of industrial processes nowadays are still regulated by PID controllers. This does not just indicate the cautious attitude of the practicing engineer towards new techniques; it reveals the rich potential of this simple control strategy for meeting various specifications for a vast variety of practical applications. The applied PID equation is a slightly modified version of (3); the only difference relies on the error function $e(\cdot)$, which must refer to the difference between the measured level of QoS and the threshold one. The weights in (3) depend on K_p, K_i, K_d (via basic algebra [8]), which are known as the proportional gain, the integral time constant, and the derivative time constant. They are set to 3.00, 1.50 and 1.25, respectively. These values guarantee the best PID performance in the following scenario and were found through accurate simulation inspection via brute force analysis.

5.2. Ideal Allocation (Ideal). An ideal allocation technique can be considered for the *Packet Delay Probability* (PDelay) control. It consists of a continuous monitoring of the buffer occupancy packet by packet. When an incoming packet experiences a delay higher than the threshold, the service rate of the buffer is instantaneously changed in order to assure the delay requirement. The operation is performed in the (1 − PDelay) percentage of the cases. Obviously Ideal is not practically implementable and may be used only as a comparison.

6. Performance Evaluation and Discussion

6.1. Variable Traffic. On-off traffic is taken as reference. Each source is an on-off process with exponentially distributed on and off time durations (mean 1.0 s and 4.0 s, resp.) and peak bandwidth of 16 kbps. Traffic enters an IP buffer whose length and service rate (set by the traffic peak bandwidth) guarantee no packet loss rate. IP traffic is encapsulated in DVB, thus generating the process $a(t)$ as output of the "Change of Format" box in Figure 1. $a(t)$ enters the DVB buffer (250 DVB cells), where the traffic delay rate in packets (of 80 bytes each) is measured every OH. The PDelay threshold is set to 5% with respect to a maximum acceptable delay of 50 ms, OH is 1 minute; RCBC gradient stepsize is set to 1.0 (no optimization of the gradient stepsize is provided, for now). The number of VoIP sources is increased of 10 from 70 to 110 each 2124 s. RCBC gradient descent is initialized by the average bandwidth of 70 sources, multiplied by the percentage DVB overhead. Figure 2 shows the resulting PDelay at the end of each OH for all the techniques, and Figure 3 shows the corresponding bandwidth allocations. From Figure 2 it is evident that both PID and RCBC sometimes produce PDelays higher than the threshold, even if close to it. In particular the number of RCBC over threshold peaks seem quite limited. The Ideal always assures under threshold delays but its average bandwidth allocation is about 7.5 Mbps (it is not reported in Figure 3 to focus on PID and RCBC allocations), which is considerably higher than PID and RCBC, whose average allocation is well below 1 Mbps. Achieving the required threshold in each OH (Ideal in Figure 2) is very bandwidth consuming.

One significant achievement arises from Figure 3: the accuracy of the RCBC computation. Just at the beginning of the simulation, RCBC rate is smoothly changed over time with much higher precision in comparison with the oscillations provided by PID. The simple observation of Figures 2 and 3 suggests that RCBC reacts quickly to traffic changes also minimizing both the bandwidth usage and bandwidth oscillations. This has an impact on the overall performance over the entire simulation horizon.

Quantitative metrics may help the interpretation of this qualitative behavior. Table 1 represents the average and standard deviation of PDelay and bandwidth over the simulation period (and noted by "Average_PDelay", "StDev_PDelay", "Average_Bw", and "StDev_Bw"), together with the percentage of the OH periods where PDelay is over threshold ("OverThr") and the average difference between

Table 1: Variable traffic: average performance.

Average_PDelay	$3.25E-02$	$5.71E-02$	$4.31E-02$
StDev_PDelay	$7.58E-02$	$1.22E-01$	$9.15E-03$
OverThr_[%]	16	32	0
AverageDiffOThr	$1.28E-02$	$3.47E-02$	$0.00E+00$
Average_Bandwidth_[Mbps]	0.761659	0.789109	7.509629
StDev_Bandwidth_[Mbps]	0.009853	0.142555	0.226046

Table 2: Variable OHs for RCBC in Table 1.

	10 m	5 m	2 m	30 s	15 s	5 s	1 s
Average_PDelay	$8.26E-02$	$5.21E-02$	$3.10E-02$	$1.03E-02$	$6.79E-03$	$8.28E-04$	$1.49E-04$
StDev_PDelay	$2.37E-01$	$1.57E-01$	$1.03E-01$	$5.92E-02$	$4.48E-02$	$2.63E-02$	$1.01E-02$
OverThr_[%]	9	4	10	5	4	0	0
AverageDiffOThr	$7.10E-02$	$3.20E-02$	$1.39E-02$	$5.22E-03$	$4.42E-03$	$7.56E-04$	$1.28E-04$
Average_Bandwidth_[Mbps]	0.812799	0.780154	0.781156	0.817441	0.839804	1.004431	0.966418
StDev_Bandwidth_[Mbps]	0.000513	0.001519	0.002621	0.003698	0.028962	0.012769	0.035661

measured PDelay and the target ("AverageDiffOThr"), for all the techniques considered.

PID does not match exactly the target on average and produces over-threshold PDelays for 32% time. Even if the average bandwidth (0.79 Mbps) assigned in this case is higher than the RCBC case, its standard deviation is quite large and, globally, PID bandwidth assignment over time is not sufficient to assure the target average performance. RCBC minimizes the bandwidth effort while assuring the average PDelay under and close to the target and minimizing bandwidth oscillations, as clear from the low bandwidth standard deviation. 16% over threshold measures may be acceptable for a large set of applications, in particular if seen jointly with the low AverageDiffOThr. Similar results can be obtained with respect to other working conditions (such as introducing other traffic categories, e.g., video streaming or changing buffer dimension).

6.2. Variable Observation Horizon (OH). The OH has an important role in estimation problems. In the case of equivalent bandwidth for loss constraints, the principle of "dominant time scale" is applied for the optimization of the OH size [12]. In practice, traffic buffers are more sensitive to traffic burstiness in the presence of small OHs, thus leading to significant oscillations of the performance measures. Table 2 extends the RCBC performance of Table 1 with variable OHs. The Average_PDelay decreases and the bandwidth increases considerably as the OH decreases. Below OH of 30 s the Average_PDelay is far below the threshold and the allocated bandwidth is overprovisioned. The motivation is that, if OH is small, PDelay is often zero but sometimes achieves high values that are significantly above the threshold. RCBC increases the assigned bandwidth very much during this lossy periods, but cannot reduce the assignments if no loss is registered in the following periods. The motivation is comprehensible directly from Algorithm 1 and formula (10): if there are no delayed packets, the gradient in (10) is zero, $\Delta_i^{\text{delay}}(t_k) = 0$ and the bandwidth assignment of the previous period is confirmed $R^{\text{TD}}(t_k) = R^{\text{TD}}(t_{k-1})$. This leads to long situations of overprovisioned bandwidth. A possible countermeasure to this may rely on setting the gradient stepsize accurately to regulate the bandwidth increases. But this would complicate the application of RCBC even more as it would need a variable stepsize as a function of the OH. In order to avoid a complicated structure of the gradient stepsize, another simple heuristic may be applied: if zero values of PDelay are registered for six consecutive times, the bandwidth is decreased of 2%. The effect of this heuristic is shown in Table 3. It guarantees quasistationary performance between 5 m and 5 s. This allows the application of RCBC for a large set of different OHs without tuning the stepsize (whose value is 1 in Tables 1–3).

6.3. Fading. The fading phenomenon is now considered. A fading process is applied over the same traffic trace used above. The employed fading process (Figure 4) has been taken from [13], where real attenuation samples are extracted from an experimental data set carried out in the Ka band on the Olympus satellite by the CSTS (Centro Studi sulle Telecomunicazioni Spaziali) Institute (Milan, Italy), on behalf of the Italian Space Agency. The Carrier/Noise Power factor is monitored at each station and, on the basis of its values, different bit and coding rates are applied to limit the BER below a chosen threshold of 10–7. Six different fading classes are defined, corresponding to combinations of channel bit and coding rate that give rise to redundancy factors $\xi_l(t)$, $l = 1, \ldots, 6$ ($\xi_l(t) \geq 1.0$); $\xi_l(t)$ represents the ratio between the Information Bit Rate (IBR) in clear sky and the IBR in specific working conditions at a given time t. The corresponding bandwidth reduction factor is $\phi(t) = (\xi_l(t))^{-1}$. With the data in [13]: $\phi(t) \in \{0.0, 0.15625, 0.3125, 0.625, 0.8333, 1.0\}$. The bandwidth

Table 3: Variable OHs with heuristic bandwidth decrease of RCBC.

	10 m	5 m	2 m	30 s	15 s	5 s	1 s
Average_PDelay	$8.26E-02$	$5.21E-02$	$3.31E-02$	$3.21E-02$	$2.70E-02$	$1.42E-02$	$5.70E-03$
StDev_PDelay	$2.37E-01$	$1.57E-01$	$1.03E-01$	$7.42E-02$	$7.03E-02$	$6.7E-02$	$5.8E-02$
OverThr_[%]	9	4	12	20	18	6	1
AverageDiffOThr	$7.10E-02$	$3.20E-02$	$1.46E-02$	$1.58E-02$	$1.63E-02$	$1.10E-02$	$4.94E-03$
Average_Bandwidth_[Mbps]	0.812799	0.780154	0.775872	0.760819	0.767617	0.807455	0.926162
StDev_Bandwidth_[Mbps]	0.000513	0.001519	0.009543	0.020729	0.031814	0.062865	0.166579

Table 4

(a) Fading: average performance, buffer 50 DVB cells

	Buffer = 50		MaxDelay = 50 ms			
	RCBC(1)	RCBC_0.5	PID	Ideal	RCBC(2)	RCBC_0.1
Average_PDelay	$1.39E-002$	$1.85E-002$	$5.69E-002$	$5.02E-002$	$8.94E-03$	$8.32E-02$
StDev_PDelay	$1.01E-001$	$1.08E-001$	$1.50E-001$	$4.26E-003$	$7.68E-02$	$2.20E-01$
OverThr_[%]	3	6	22	52	3	25
AverageDiffOThr	$1.14E-002$	$1.46E-002$	$4.35E-002$	$1.69E-003$	$6.82E-03$	$6.67E-02$
Average_Bandwidth_[Mbps]	2.84	2.76	3.26	3.54	3.17	2.52
StDev_Bandwidth_[Mbps]	2.24	2.20	2.57	1.20	2.20	1.83

(b) Fading: average performance, buffer 150 DVB cells

	Buffer = 150		MaxDelay = 50 ms			
	RCBC(1)	RCBC_0.5	PID	Ideal	RCBC(2)	RCBC_0.1
Average_PDelay	$1.39E-002$	$1.94E-002$	$5.57E-002$	$5.02E-002$	$8.21E-03$	$7.21E-02$
StDev_PDelay	$1.06E-001$	$1.12E-001$	$1.64E-001$	$4.20E-003$	$7.98E-02$	$2.27E-01$
OverThr_[%]	3	6	18	51	2	15
AverageDiffOThr	$1.22E-002$	$1.59E-002$	$4.57E-002$	$1.67E-003$	$7.17E-03$	$6.31E-02$
Average_Bandwidth_[Mbps]	2.95	2.91	3.49	10.63	3.34	2.63
StDev_Bandwidth_[Mbps]	2.41	2.32	2.73	3.61	2.38	1.91

(c) Fading: average performance, buffer 350 DVB cells

	Buffer = 350		MaxDelay = 50 ms			
	RCBC(1)	RCBC_0.5	PID	Ideal	RCBC(2)	RCBC_0.1
Average_PDelay	$1.32E-002$	$1.74E-002$	$5.53E-002$	$5.02E-002$	$7.19E-03$	$6.93E-02$
StDev_PDelay	$1.06E-001$	$1.11E-001$	$1.69E-001$	$4.26E-003$	$8.04E-02$	$2.27E-01$
OverThr_[%]	2	4	17	52	1	14
AverageDiffOThr	$1.19E-002$	$1.45E-002$	$4.57E-002$	$1.68E-003$	$6.49E-03$	$6.08E-02$
Average_Bandwidth_[Mbps]	2.98	2.89	3.52	24.81	3.39	2.64
StDev_Bandwidth_[Mbps]	2.46	2.38	2.71	8.42	2.46	1.93

reduction due to fading, denoted by $R^{\text{TD}}_{\text{real}}(t)$ (see Figure 1), can be computed as $R^{\text{TD}}_{\text{real}}(t) = \phi(t) \cdot R^{\text{TD}}(t)$; with $\phi(t) = (\xi_l(t))^{-1}$. As only the rate $R^{\text{TD}}_{\text{real}}(t)$ is available for data traffic, $R^{\text{TD}}(t)$ has to be tuned over time in order to maintain the required QoS. All system parameters are unchanged with respect to the already presented results except for the burstiness that is now 2.0 and the reallocation time period that is reduced to 10 s to tackle fading variations whose granularity is 1 minute.

The notation RCBC(s) is referred to the adoption of RCBC with a fixed stepsize set to s ($\text{step}_k = s$, for all k). RCBC_ν defines the adoption of the Vogl method (whose tunable parameter is ν) to optimize the stepsize [14]: $\text{step}_k = \nu \cdot |\widehat{R}_i^{\text{delay},k} - \widehat{R}_{i,\text{thr}}^{\text{delay},k}|$, where $\widehat{R}_i^{\text{delay},k}$ and $\widehat{R}_{i,\text{thr}}^{\text{delay},k}$ are defined in (8) and (9). Tables 4(a)–4(c) summarize all the average performance values for three cases of DVB buffer size: 50, 150, and 350 cells. Ideal is again perfect in terms of average PDelay, but with a significant bandwidth allocation. The average PDelay of RCBC(2) is too low while the one of RCBC_0.1 is always above the target. RCBC(1) and RCBC_0.5 have similar good performance, but RCBC_0.5 reveals to be the best choice in all cases as it makes use of a smaller amount of bandwidth than RCBC(1), with slightly higher PDelay, but always below the target.

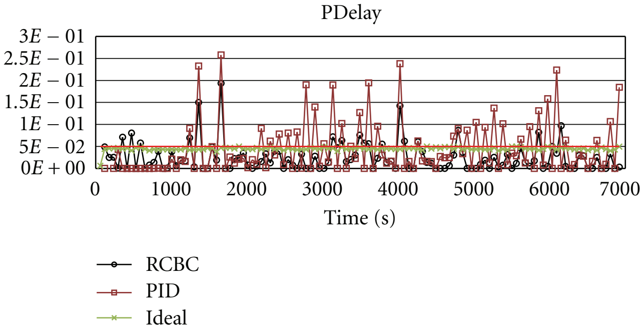

Figure 2: Variable traffic: PDelay (target $5 \cdot 10^{-2}$).

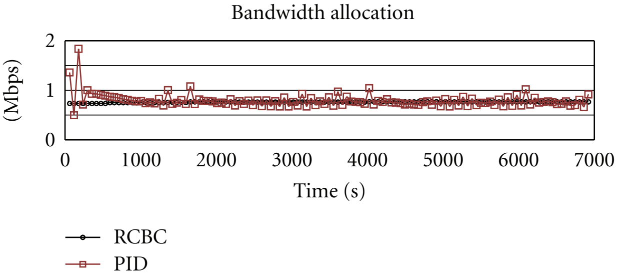

Figure 3: Variable traffic: bandwidth allocations.

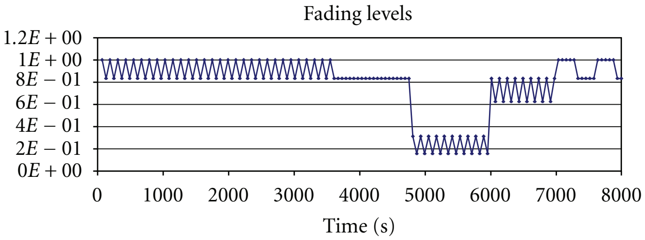

Figure 4: Fading: real fading trace (bandwidth reduction factor $\phi(t)$) used in the simulations, taken from [13].

7. Conclusions and Future Work

The paper presents a novel control scheme that adapts the bandwidth to be allocated to a buffer which conveys heterogeneous traffic (both concerning traffic sources and QoS requirements) with delay constraint in a layer-in-cascade model. The proposed algorithm is based only on measures and does not use closed-form expressions, a priori information about traffic statistical properties, and assumptions about buffer dimension. The reliability of the algorithm proposed, shown in the simulations, opens the door to future investigation involving the joint control of loss and delay together with other possible traffic categories.

References

[1] M. Marchese and M. Mongelli, "Vertical QoS mapping over wireless interfaces," *IEEE Wireless Communications Magazine*, vol. 16, no. 2, pp. 37–43, 2009.

[2] M. Marchese, *QoS over Heterogeneous Networks*, John Wiley & Sons, Chichester, UK, 2007.

[3] ETSI. Satellite Earth Stations and Systems (SES), "Broadband Satellite Multimedia, IP over Satellite," Tech. Rep. ETSI Technical Report TR 101 985, V1.1.2, 2002.

[4] ETSI. Satellite Earth Stations and Systems (SES), "Broadband Satellite Multimedia (BSM) Services and Architectures, QoS Functional Architecture," Tech. Rep. TS 102 462 V0.4.2-01, 2006.

[5] ETSI. Satellite Earth Stations and Systems (SES), "Broadband Satellite Multimedia (BSM) Services and Architectures, Interworking with RSVP-based QoS (IntServ)," Tech. Rep. TS 102 463 V0.4.2-10, 2006.

[6] ETSI. Satellite Earth Stations and Systems (SES), "Broadband Satellite Multimedia (BSM) Services and Architectures, Interworking with DiffServ QoS," Tech. Rep. TS 102 464 V0.4.1-09, 2006.

[7] ETSI. Satellite Earth Stations and Systems (SES), "Broadband Satellite Multimedia. Services and Architectures; BSM Traffic Classes. ETSI Technical Specification," Tech. Rep. TS 102 295 V1.1.1, 2004.

[8] M. Xu, S. Li, C. Qi, and W. Cai, "Auto-tuning of PID controller parameters with supervised receding horizon optimization," *ISA Transactions*, vol. 44, no. 4, pp. 491–500, 2005.

[9] C. G. Cassandras, G. Sun, C. G. Panayiotou, and Y. Wardi, "Perturbation analysis and control of two-class stochastic fluid models for communication networks," *IEEE Transactions on Automatic Control*, vol. 48, no. 5, pp. 770–782, 2003.

[10] Y. Wardi, B. Melamed, C. G. Cassandras, and C. G. Panayiòtou, "Online IPA gradient estimators in stochastic continuous fluid models," *Journal of Optimization Theory and Applications*, vol. 115, no. 2, pp. 369–405, 2002.

[11] M. Marchese and M. Mongelli, "Measurement-based computation of generalized equivalent bandwidth for loss constraints," *IEEE Communications Letters*, vol. 11, no. 12, pp. 1007–1009, 2007.

[12] S. Georgoulas, P. Trimintzios, G. Pavlou, and K. Ho, "Measurement-based admission control for real-time traffic in IP differentiated services networks," in *Proceedings of IEEE International Conference on Telecommunications (ICT '05)*, Capetown, South Africa, 2005.

[13] N. Celandroni, F. Davoli, and E. Ferro, "Static and dynamic resource allocation in a multiservice satellite network with fading," *International Journal of Satellite Communications and Networking*, vol. 21, no. 4-5, pp. 469–487, 2003.

[14] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink, and D. L. Alkon, "Accelerating the convergence of the back-propagation method," *Biological Cybernetics*, vol. 59, no. 4-5, pp. 257–263, 1988.