Measurement-Based Computation of Generalized Equivalent Bandwidth for Loss Constraints

Mario Marchese, *Senior Member, IEEE,* and Maurizio Mongelli, *Member, IEEE*

***Abstract*— This letter proposes a novel measurement-based Equivalent Bandwidth technique that computes the bandwidth to be allocated to a buffer which conveys heterogeneous traffic (both concerning traffic sources and QoS requirements), without using any closed-form expression. The effectiveness of the algorithm is checked through simulation analysis.**

***Index Terms*— QoS, equivalent bandwidth, measurement control.**

I. Introduction

Equivalent bandwidth (EqB), is defined as the minimum service rate to be provided to a traffic buffer to guarantee a certain degree of *Quality of Service* (QoS) in terms of objective parameters (packet loss, delay, jitter). EqB techniques are obtained analytically for homogeneous traffic trunks, with respect to a single QoS constraint. Modern network solutions often imply the aggregation of service classes with different QoS constraints, thus generating heterogeneous trunks from the point of view of both traffic sources and QoS requirements. This situation leads to the need to develop new equivalent bandwidth techniques so as to match heterogeneity.

II. Assumptions

There are N traffic classes. $a_i(t)$ is the *input rate* process of the i-th traffic class and $a(t)$ the aggregate process of all $a_i(t), i = 1, ..., N$. Traffic is conveyed towards a single buffer, modelled through a *Stochastic Fluid Model* [1, 2]. $a(t)$ is supposed 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 $a(t)$. Additionally, aggregation may involve also buffering and encapsulation operations as typically done in real network nodes. It makes $a(t)$ analytical modelling virtually impossible to get, also in case of full knowledge of $a_i(t)$ processes. The only information about $a_i(t)$ and $a(t)$ may be got through real measures. The service rate of the buffer is $R(t)$. $l_i(R(t), t)$ is the *loss rate* process of the i-th traffic class, measured in [bps]. The average value of the loss rate is defined in (1) for $i = 1, ..., N$. No analytical expression for $l_i(R(t), t)$ is supposed available. Information about its behaviour is got by measures. The entire system model is reported in Fig. 1.

The SLA (*Service Level Agreement*) for each traffic class is composed of a *Packet Loss Probability* threshold (PLP_i^*). It means that the amount of feasible loss rate must be limited in any time instant by the process $l_i^*(t) = PLP_i^* \cdot a_i(t)$, measured in [bps], whose average value is contained in (2) for $i = 1, ..., N$.

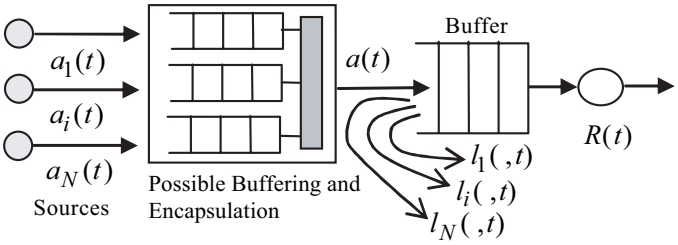

Fig. 1. System model.

$$\bar{l}_i = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{\tau} l_i(t)\ dt \tag{1}$$

$$\bar{l}_i^* = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{\tau} l_i^*(t)\ dt \tag{2}$$

III. Problem Definition and Solution

The aim is to provide the minimum buffer service rate so that the maximum quadratic distance between $\bar{l}_i$ and $\bar{l}_i^*$ is minimized. It corresponds to define the optimization problem in (3), introduced in this letter and identified as *Generalized Equivalent Bandwidth* (GEqB).

$$R^* = \arg\min_R\ l_\Delta(\cdot, R), \quad l_\Delta(\cdot, R) = \max_i\ \left\{\bar{l}_i - \bar{l}_i^*\right\}^2 \tag{3}$$

Being the involved stochastic processes unknown, GEqB problem is solved by taking measures over a given k-th *observation horizon* (OH), $T_k = [t_{k-1}, t_k]$, $k = 1, 2, ...$ and performing a sequence of bandwidth reallocations $R(t_k), k = 1, 2, ...$, each T_k, based on the gradient method. The loss rate $l_i(R(t_k), t)$ and the feasible loss rate $l_i^*(t)$ processes are averaged over each OH, giving origin to the quantities $\widehat{l}_i^k$ in (4) and $\widehat{l}_i^{*,k}$ in (5). Being used to solve the GEqB problem, $\widehat{l}_i^k$ and $\widehat{l}_i^{*,k}$ must be representative of the average values $\bar{l}_i$ and $\bar{l}_i^*$, $\forall i = 1, ..., N$ and $\forall k$.

$$\widehat{l}_i^k = \frac{1}{T_k} \int_{T_k} l_i(t)\ dt \tag{4}$$

$$\widehat{l}_i^{*,k} = \frac{1}{T_k} \int_{T_k} l_i^*(t)\ dt \tag{5}$$

The bandwidth allocation at each instant t_k is ruled by the algorithm introduced in Fig. 2 and called *Gradient-based Generalized Equivalent Bandwidth* (G^2EqB) algorithm. $step_k$

Manuscript received June 17, 2007. The associate editor coordinating the review of this letter and approving it for publication was Prof. Carla-Fabiana Chiasserini.

The authors are with the Department of Communications, Computer and Systems Science, University of Genoa, Italy (e-mail: {mario.marchese, maurizio.mongelli}@unige.it).

Digital Object Identifier 10.1109/LCOMM.2007.070982.

a) if $\hat{l}_i^k - \hat{l}_i^{*,k} \geq 0$ for at least one i

$$\Delta_i(t_k) = \begin{cases} 2\left.\dfrac{\partial \hat{l}_i(R)}{\partial R}\right|_{R=R(t_{k-1})} \left[\hat{l}_i^{*,k} - \hat{l}_i^k\right], & \text{if } \hat{l}_i^k - \hat{l}_i^{*,k} \geq 0 \\ 0 & otherwise \end{cases}$$

$$\Delta(t_k) = \max_i \left|\Delta_i(t_k)\right|, \quad R(t_k) = R(t_{k-1}) + step_k \cdot \Delta(t_k)$$

b) if $\hat{l}_i^k - \hat{l}_i^{*,k} < 0, \ \forall i; \ \Delta_i(t_k) = 2\left.\dfrac{\partial \hat{l}_i(R)}{\partial R}\right|_{R=R(t_{k-1})} \cdot \left[\hat{l}_i^k - \hat{l}_i^{*,k}\right]$

$$\Delta(t_k) = \min_i \left|\Delta_i(t_k)\right|, \quad R(t_k) = R(t_{k-1}) - step_k \cdot \Delta(t_k)$$

Fig. 2. G^2EqB algorithm.

is the gradient stepsize. Condition **a)** means that the allocated bandwidth needs to be increased. Condition **b)** states the opposite. Derivatives $\dfrac{\partial \hat{l}_i(R)}{\partial R}$ represent the sensitivity of the loss 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 within each OH, which is divided into N_{T_k} busy periods (i.e., where the buffer is not empty) identified by the variable bp. If there is one traffic class, the derivative exact form is presented in [1], but it is still unknown in the multiple class case when the service rate is the control variable. This letter introduces the approximation in (6). (6) is equality in case of single class, as proved in [1]. $\left\{ {}^i at_{T_k}^{bp}(R(t_{k-1})) - {}^i ll_{T_k}^{bp}(R(t_{k-1}))\right\}$ is the contribution to information loss of the i-th traffic class for the busy period bp within $T_k, k = 1, 2, ...$ ${}^i at_{T_k}^{bp}$ is the arrival time of the first packet of service class i within the busy period bp. ${}^i ll_{T_k}^{bp}$ is the time when the last loss of class i occurs during bp.

$$\left.\frac{\partial \hat{l}_i(R)}{\partial R}\right|_{R=R(t_{k-1})} \cong -\frac{1}{T_k}\sum_{bp=1}^{N_{T_k}} \left\{ {}^i at_{T_k}^{bp}(R(t_{k-1})) - {}^i ll_{T_k}^{bp}(R(t_{k-1}))\right\} \quad (6)$$

IV. Algorithm Convergence

The technical conditions for convergence to global optimum ($\lim_{k\to\infty} R(t_k) = R^*$) are: **1)** ergodic stochastic processes, **2)** decreasing behaviour of $step_k$, **3)** gradient bounded within the control domain $R(t) \in \Re^+$, $\forall t$, and **4)** non-existence of local optima. **1)** and **2)** are assumptions. Concerning **3)**, the lengths of buffer busy periods are bounded by OH size; measured loss rate at the end of each OH cannot be infinite. Concerning **4)**, the loss rate of a traffic queue can be reasonably assumed to be continuous, differentiable, with a negative derivative with respect to the service rate, so the cost function is also continuous, differentiable with unique minimum. The convergence speed depends on the length of OH and on the gradient stepsize. The length of OH is important because, on one hand, it must be long enough to assure that $\hat{l}_i^k$ and $\hat{l}_i^{*,k}$ are representative of the average values $\bar{l}_i$ and $\bar{l}_i^*$, $\forall i = 1, ..., N$ and $\forall k$, but, on the other hand, it must be short enough to assure quick convergence. In this context, the assumption of process ergodicity may be relaxed and limited to the time that the sequence of bandwidth allocations $R(t_k), k = 1, 2, ...$ needs to converge to R^*. When $a(t)$ changes its statistical behaviour, a new GEqB problem is solved by supposing $a(t)$ ergodic at least within the convergence time and by starting the G^2EqB algorithm again. In consequence, tuning OH length is important also to get fast reactions to traffic variations. Concerning gradient stepsize dimension, tests not reported here have shown that it is not a critical parameter for convergence. Actually, convergence condition **2)** may be relaxed: setting a proper constant value of the gradient stepsize, as done in [2] and in this paper, does not affect convergence. For example if $step_k = 1$ in the tests reported below the algorithm converges, but stepsize length adaptation helps improve convergence speed and limit bandwidth oscillations.

V. Performance Analysis and Discussion

Due to the complexity of the overall input rate process $a(t)$, equivalent bandwidth approaches which use complex mathematical descriptors may be hardly applied in real time. The approach in [3] (called EqB in the following) is applicable in this context and used here to make a comparison with G^2EqB. $m_a(t_k)$ and $\sigma_a(t_k)$ are the measured *mean* and *standard deviation* of $a(t)$ over the k-th OH. Bandwidth is assigned at time $t_k, k = 1, 2, ...$ as in (7). PLP_{EqB}^* is the allowed PLP upper bound and is defined as the most stringent PLP requirement out of N SLAs.

$$R(t_k) = m_a(t_k) + \sigma_a(t_k) \cdot \sqrt{-2ln(PLP_{EqB}^*) - ln(2\pi)} \quad (7)$$

A. G^2EqB versus EqB: rate provision and convergence

VoIP SLA is considered. Each source is an on-off process. Mean on and off time durations are exponentially distributed with mean 1.008 s and 1.587 s, respectively. Peak bandwidth is 16 kbps. VoIP 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 ATM (via AAL5) so generating the process $a(t)$ as output of the "Buffering and Encapsulation" box in Fig. 1. $a(t)$ enters the ATM buffer (1600 bytes), where the VoIP loss rate $\hat{l}^k$ in IP packets is measured each T_k. PLP_{VoIP}^* is set to $2 \cdot 10^{-2}$; G^2EqB OH to 30 s; gradient stepsize to 6.0. EqB OH is either fixed to 30 s or tuned through an approximation of the *Dominant Time Scale* principle that computes the optimal OH size to get a reliable estimate of EqB statistics. The two alternatives are respectively identified as "EqB OH 30s" and "EqB OH DTS", in the following. Fig. 3 and Fig. 4 show, respectively, PLP and corresponding allocated bandwidth of G^2EqB and EqB. The number of VoIP sources is increased of 10 from 70 to 110 each 3000 seconds. Average PLP results are: $4.40 \cdot 10^{-3}$ for G^2EqB and $1.18 \cdot 10^{-2}$ for EqB OH 30s. EqB OH DTS assures null packet loss. Average allocated bandwidths are: 0.842 Mbps for G^2EqB, 0.867 Mbps for EqB OH 30s, and 1.31 Mbps for EqB OH DTS. Even if the average PLP values seem to be satisfying for all the schemes, the simple observation of Figs. 3 and 4 suggests that G^2EqB reacts quickly to traffic changes also minimizing bandwidth oscillations. EqB OH DTS always

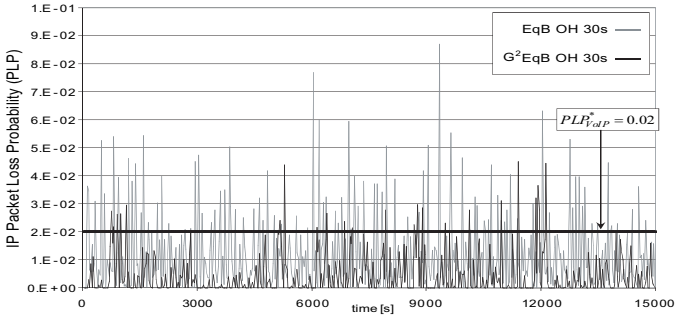

Fig. 3. G^{2}EqB and EqB: PLP.

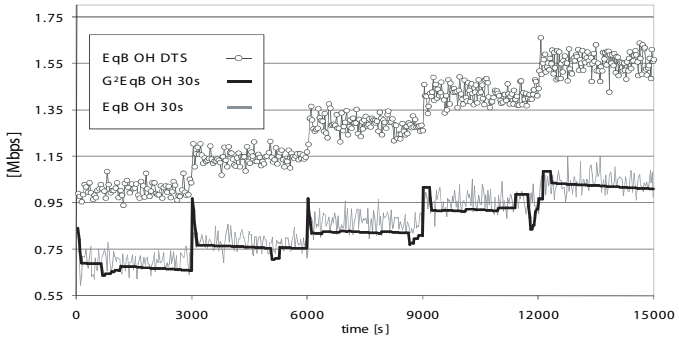

Fig. 4. G^{2}EqB and EqB: bandwidth allocation.

matches PLP^*_{VoIP} request but implies a relevant bandwidth waste; EqB OH 30s often fails to guarantee PLP^*_{VoIP} and introduces wide bandwidth and PLP oscillations. Quantitative metrics may help the interpretation of this qualitative behaviour. PLP standard deviation is $7.4 \cdot 10^{-3}$ for G^{2}EqB and $1.33 \cdot 10^{-2}$ for EqB OH 30s. The percentage of the OH periods where PLP is over threshold is 5% for G^{2}EqB and 18.6% for EqB OH 30s. The average difference value between measured PLP and PLP^*_{VoIP} selecting the OH periods where PLP is over threshold is $4.22 \cdot 10^{-4}$ for G^{2}EqB and $2.77 \cdot 10^{-3}$ for EqB OH 30s. G^{2}EqB allows minimizing the distance between measured and threshold values.

B. G^{2}EqB versus EqB: heterogeneous traffic

A real video trace ("Jurassik park" from [4]) is added to the VoIP scenario used in section V.A. Peak and average rate are 1.418 and 0.280 Mbps. Video enters an IP buffer whose length and service rate (set to 1.418 Mbps) guarantee no loss. Both VoIP and video traffic are encapsulated over DVB (packets of 188 bytes) at the exit of the IP buffers and generate the process $a(t)$. PLP^*_{video} is set to $5 \cdot 10^{-3}$. Fig. 5 contains the average measured video and VoIP PLPs together with the allocated bandwidth to the DVB buffer in [Mbps] for G^{2}EqB OH 3min (G^{2}EqB in Fig. 5), EqB OH 3min, and EqB OH DTS. DVB buffer dimension is changed as well as the number of VoIP calls. Each single test simulates 107 overall minutes. The average G^{2}EqB PLP is always close to but below the threshold of the most restrictive requirement (PLP^*_{video}). G^{2}EqB is adaptive to buffer length because its behaviour depends only on loss measures. EqB-based schemes do not adapt to buffer length: EqB OH 3min underestimates the bandwidth and fails to match video requirements. EqB OH DTS behaves similarly for short buffer length while overestimates the bandwidth for larger buffer dimensions.

Number VoIP Calls	Buffer Length [bytes]	Allocated Bandwidth **G^{2}EqB**/EqB OH 3 min/EqB OH DTS	Video PLP **G^{2}EqB**/ EqB OH 3 min/ EqB OH DTS	VoIP PLP **G^{2}EqB**/ EqB OH 3 min/ EqB OH DTS
30	9400	**0.94** / 0.79 / 0.87	**2.62 10^{-3}** / 8.26 10^{-2} / 1.05 10^{-2}	**1.20 × 10^{-4}** / 5.58 10^{-3} / 5.93 10^{-4}
30	18800	**0.84** / 0.79 / 0.87	**3.20 × 10^{-3}** / 2.96 10^{-2} / 0.0	**2.22 × 10^{-4}** / 2.13 10^{-3} / 0.0
30	28200	**0.82** / 0.79 / 0.87	**1.78 × 10^{-3}** / 1.34 10^{-2} / 0.0	**1.24 × 10^{-4}** / 9.14 10^{-4} / 0.0
60	9400	**1.47** / 1.26 / 1.41	**3.94 × 10^{-3}** / 8.26 10^{-2} / 1.15 10^{-2}	**2.26 × 10^{-4}** / 7.77 10^{-3} / 5.63 10^{-4}
60	18800	**1.41** / 1.26 / 1.41	**2.23 × 10^{-3}** / 5.13 10^{-2} / 1.54 10^{-3}	**1.55 × 10^{-4}** / 3.96 10^{-3} / 8.17 10^{-5}
60	28200	**1.39** / 1.26 / 1.41	**1.38 × 10^{-3}** / 3.0 10^{-2} / 4.20 10^{-4}	**1.04 × 10^{-4}** / 2.32 10^{-3} / 8.17 10^{-5}
90	9400	**1.97** / 1.73 / 2.00	**4.13 × 10^{-3}** / 1.07 10^{-1} / 1.58 10^{-2}	**2.38 × 10^{-4}** / 8.11 10^{-3} / 8.28 10^{-4}
90	18800	**1.87** / 1.73 / 2.0	**3.26 × 10^{-3}** / 5.33 10^{-2} / 4.35 10^{-3}	**1.94 × 10^{-4}** / 3.97 10^{-3} / 2.37 10^{-4}
90	28200	**1.84** / 1.73 / 2.0	**1.73 × 10^{-3}** / 3.07 10^{-2} / 2.02 10^{-3}	**1.02 × 10^{-4}** / 2.35 10^{-3} / 2.37 10^{-4}

Fig. 5. G^{2}EqB and EqB: packet loss and bandwidth allocation.

VI. Conclusions

A novel equivalent bandwidth algorithm to automatically adapt the rate assigned to a buffer which conveys heterogeneous traffic is presented. It is based only on measures and does not use closed-form expressions, a-priori information about traffic statistical properties, and assumptions about buffer dimension.

References

[1] Y. Wardi, B. Melamed, C. G. Cassandras, and C. G. Panayiotou, "Online IPA gradient estimators in stochastic continuous fluid models," *J. Optimization Theory and Applic.*, vol. 115, no. 2, pp. 369–405, Nov. 2002.

[2] 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 Trans. Automat. Contr.*, vol. 48, no. 5, pp. 770–782, May 2003.

[3] R. Gurin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 7, pp. 968-981, Sept. 1991.

[4] http://www-tkn.ee.tu-berlin.de/research/trace/trace.html.