Resource Allocation for Video Streaming in Wireless Environment

Shahrokh Valaee and Jean-Charles Gregoire

Abstract— **This paper focuses on the development of a new resource allocation scheme for video streaming in wireless networks. The technique utilizes the ϵ-weak burstiness curve which has been defined as the $(1-\epsilon)$-percentile of the maximum burstiness curve. We show that the ϵ-weak burstiness curve is a convex non-increasing function of the service rate. We also devise a "water-filling" algorithm to obtain the empirical ϵ-weak burstiness curve. The technique has been applied to MPEG-4 encoded video traces. The numerical studies show that if the ϵ-weak burstiness curve is used for resource provisioning, a considerable percentage of bandwidth will be saved.**

Keywords: Video streaming, burstiness curve, water-filling, single FIFO queue, VBR traffic, MPEG-4 trace.

I. Introduction

The next generation wireless systems should support multimedia streaming applications. Multimedia streams are usually modelled by variable-bit-rate (VBR) traffics. Video streams are generated by applying an encoder — such as an MPEG-4 encoder — to digitized video traces. The result is a periodic sequence of packets with nonuniform lengths. The size of each packet is related to the information content of the frame and its location inside the group of frames.

To support video applications, the network should provide appropriate resources — in terms of guaranteed bandwidth and buffer size — for the encoded trace. The assigned resources are frequently measured in the stochastic or deterministic paradigm. In the stochastic approach, a canonical probability distribution function is assigned to input traffic and the performance is measured in terms of the mean and percentiles of the network behavior. A mismatch between the selected model and the underlying data structure might produce significant errors in the estimated network behavior. In the deterministic network provisioning, the input traffic is replaced by its worst-case behavior — *greedy* source [1] [2] [3]. This approach usually creates conservative solutions for resource allocation.

Recent studies of the Internet traffic indicate that the classical approaches to traffic modelling, such as the Poisson and Markov processes, do not provide appropriate models for LAN and WAN traffics [4] [5]. It has been shown that, unlike the traditional traffic models, the Internet traffic has long-range dependence. This observation has been the motivating force for devising self-similar models for Internet traffic [6]. It has also been shown that MPEG coded video traces also reveal long-range dependence and that self-similar models can also be employed for video streams [7] [8]. Unfortunately, self-similar models are very complicated and do not provide simple solutions for network provisioning.

In this paper, we propose an alternative approach. We devise a technique for non-conservative resource allocation. Nevertheless, we would like to avoid the necessity of selecting a probability model for input traffic. We assume that the network has a single bottleneck node and model the network with a simple first-in-first-out queue. This is a valid assumption if we consider that the bandwidth is usually the scarce commodity in the wireless hops as compared to the backbone wired network. We measure the network performance in terms of percentile of the queue size. The proposed approach is stochastic; nonetheless, no explicit model for input traffic is imposed.

An acceptable level of quality-of-service (QoS) for the transmitted stream will be provided if the wireless network can guarantee a minimum bandwidth for the encoded video stream. For a real-time application, an acceptable bandwidth would be the peak rate. If a bandwidth equal to the peak rate is allocated to the video stream, each packet will arrive at the receiver early enough so that it can be played back in its proper time instant. Although using the peak rate will guarantee a real-time playback of the video stream, it will require a large allocated bandwidth and will lead to a conservative network design.

On the other hand, if the allocated bandwidth is smaller than the peak rate, some packets will not arrive in time for playback. In such cases, a buffer will be used to store the packets. The packets will be played back after an appropriate amount of traffic is accumulated in the buffer. The packets will be collected in the playback buffer and will be read into the decoder in a constant rate. In practice, there exists a relationship between the size of the playback buffer, the induced delay, the speed of transmission, and the percentage of data that will be buffered. In this paper, we will formulate this relationship and use it to estimate the performance of video streaming applications in a wireless environment. Due to lack of space, we will not provide the proofs of lemmas and theorems; for proofs refer to [9].

The completion of this research was made possible in parts thanks to Bell Canada's support through its Bell University Laboratories R & D program.

S. Valaee is with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, 10 King's College Road, University of Toronto, Toronto, Ontario, Canada, M5S 3G4. Email: valaee@comm.utoronto.ca.

J. C. Gregoire is with INRS-Telecommunications, Universite du Quebec, Place Bonaventure, 900 Rue de la Gaucheiere Ouest, Niveau C, Montreal, Quebec, Canada, H5A 1C6. Email: gregoire@inrs-telecom.uquebec.ca

II. Problem Formulation

A video stream is represented by the sequence $\{\ldots, r_{-1}, r_0, r_1, \ldots\}$ where r_i is the size of the encoded video packet at time $t_i = i\Delta$, with $i = \ldots, -1, 0, 1, \ldots$, and Δ the time difference between two consecutive frames. We assume that each frame is encoded in a single packet. The aggregated input traffic over the interval $[s, t]$ is represented by

$$R(s,t) = \sum_{i=\lceil s \rceil}^{\lfloor t \rfloor} r_i \tag{1}$$

where $\lceil s \rceil$ is the smallest integer larger than or equal to s and $\lfloor t \rfloor$ is the largest integer smaller than or equal to t. Throughout this study, we assume that the input traffic is generated by a discrete stationary stochastic process with

$$\sup_i \frac{r_i}{\Delta} = \rho_{\scriptscriptstyle M} \tag{2}$$

$$\lim_{t\to\infty} \frac{R(s,t)}{t-s} = \bar{\rho} \quad \text{uniformly in } s. \tag{3}$$

$\rho_{\scriptscriptstyle M}$ is the maximum rate of the traffic and $\bar{\rho}$ is the average rate.

Backlogged traffic at time t is defined as

$$Q_t \triangleq \sup_{-\infty \le s \le t} \{R(s,t) - S(s,t)\} \tag{4}$$

where $S(s,t)$ is the amount of traffic served in the interval $[s,t]$. For a server with the constant rate ρ, the total service given over the interval $[s,t]$ will be $S(s,t) = (t-s)\rho$. The backlogged traffic will then be represented by

$$Q_t(\rho) = \sup_{-\infty \le s \le t} \{R(s,t) - (t-s)\rho\}. \tag{5}$$

Since the system is stationary, one might be able to represent the queue size at the origin by

$$Q_0(\rho) = \sup_{t \ge 0} \{R_{-t} - t\rho\} \tag{6}$$

where R_{-t} is the aggregated traffic over the interval $[-t, 0]$.

Definition 1: For a given $0 \le \epsilon \le 1$, the *ϵ-weak burstiness curve* is defined as

$$b(\epsilon, \rho) \triangleq \inf\Big\{ b \,|\, \Pr(Q_0(\rho) \ge b) \le \epsilon \Big\}. \tag{7}$$

The ϵ-weak burstiness curve is, in fact, the $(1-\epsilon)$-percentile of the burstiness curve [10] [11]

$$b(0, \rho) = \sup_t \Big\{ Q_t(\rho) \Big\}. \tag{8}$$

Proposition 1: The ϵ-weak burstiness curve satisfies the following properties:
(i) For any $0 \le \epsilon \le 1$, $b(\epsilon, \rho)$ is a non-increasing convex function of ρ;
(ii) For any $\rho \ge 0$, $b(\epsilon, \rho)$ is a non-increasing function of ϵ.
(iii) If $Q_t(\rho) \le Q'_t(\rho)$ for all t, then $b(\epsilon, \rho) \le b'(\epsilon, \rho)$ where $b(\epsilon, \rho)$ and $b'(\epsilon, \rho)$ are, respectively, the ϵ-weak burstiness curves associated to $Q_t(\rho)$ and $Q'_t(\rho)$.

The following theorem shows that a mix of traffics will have an ϵ-weak burstiness curve smaller than the sum of the ϵ-weak burstiness curves of individual flows.

Theorem 1: Consider the traffic of L users being multiplexed over a single link with the constant rate ρ (normalized over Δ seconds). Let the aggregated input traffic for user i over the interval $[-t, 0]$ be represented by $R^{(i)}_{-t}$. The corresponding backlog at the origin will be indicated by $Q^{(i)}_0(\rho) = \sup_{t\ge 0}\{R^{(i)}_{-t} - \rho t\}$. Define $b^{(i)}(\epsilon, \rho) = \inf\{b \,|\, \Pr(Q^{(i)}_0(\rho) \ge b) \le \epsilon\}$ for $i = 1, \ldots, L$. Let the aggregated multiplexed traffic over the interval $[-t, 0]$ be represented by $R^{\Sigma}_{-t} \triangleq \sum_{i=1}^{L} \lambda_i R^{(i)}_{-t}$ where $\sum_{i=1}^{L} \lambda_i = 1$. Also define $Q^{\Sigma}_0(\rho) = \sup_{t\ge 0}\{R^{\Sigma}_{-t} - \rho t\}$, and $b^{\Sigma}(\epsilon, \rho) = \inf\{b \,|\, \Pr(Q^{\Sigma}_0(\rho) \ge b) \le \epsilon\}$. Then

$$b^{\Sigma}(\epsilon, \rho) \le \sum_{i=1}^{L} \lambda_i b^{(i)}(\epsilon, \rho). \quad \square \tag{9}$$

This theorem shows that multiplexing can reduce the ϵ-weak burstiness curve. In fact, the total burstiness of the multiplexed traffic is smaller than the summation of the burstiness of all users.

III. Empirical Approach

In the previous section, we defined the ϵ-weak burstiness curve assuming stationary traffic over an infinite interval. The assumption of infinite interval is however unrealistic in practice. In the present section, we investigate the case of finite intervals.

We consider the discrete time observation over the grid $i\Delta$, for $i = 0, 1, \ldots$. Without loss of generality, in the sequel, we assume $\Delta = 1$. The queue size for a constant rate server in a discrete time setting satisfies the Lindley's equation,

$$q_m = [q_{m-1} - \rho]^+ + r_m \tag{10}$$

where q_m is the backlog at time m, ρ is the service given over a unit time (Δ), r_m is the amount of traffic arriving at time m, and $[a]^+ = max\{a, 0\}$.

For a given threshold σ one is able to define

$$\mu_n(\sigma) \triangleq \frac{1}{n} \sum_{i=0}^{n-1} 1\Big\{ q_i > \sigma \Big\} \min\Big\{ \frac{q_i - \sigma}{\rho}, 1 \Big\} \tag{11}$$

where $1\{.\}$ is the *indicator* function — $1\{\mathcal{A}\} = 1$ if the predicate $\mathcal{A}$ is true, and $1\{\mathcal{A}\} = 0$ if the predicate $\mathcal{A}$ is false. Note that $\mu_n(\sigma)$ is also a function of ρ (for simplicity of notation we have dropped this parameter). Let

$$Q_t(\rho) = q_m - (t-m)\rho, \quad \text{for } m \le t < m+1. \tag{12}$$

If q_m is the sample of the queue size $Q_t(\rho)$ at the mth time instant, then $\mu_n(\sigma)$ will be the proportion of time that the queue size stays above the threshold, σ (see Fig. 1 for an illustration). Therefore, $\mu_n(\sigma)$ is, in fact, the extent of the time for which $Q_t(\rho) > \sigma$, normalized over the whole window of observation.

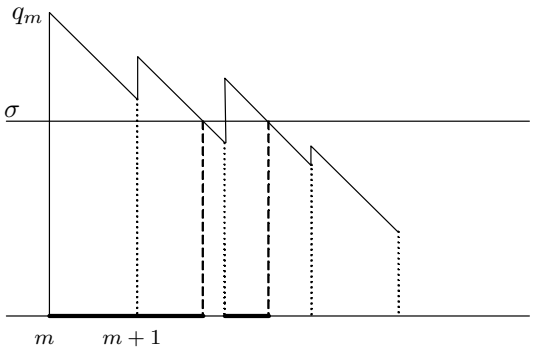

Fig. 1. The queue size in a discrete input setting and constant service rate. The time extent over which $Q_t(\rho)$ is larger than σ has been illustrated in heavy lines.

Definition 2: The *empirical ϵ-weak burstiness curve* for the observations over the interval $[0, n]$ is defined as a function $b_n(\epsilon, \rho)$ that satisfies

$$\mu_n\Big(b_n(\epsilon, \rho)\Big) = \epsilon. \tag{13}$$

Using (12), we can prove the following lemma.

Lemma 1: For all $n \geq 1$,

$$\frac{1}{n}\int_0^n 1\{Q_t(\rho) > b_n(\epsilon, \rho)\}dt = \epsilon. \quad \square \tag{14}$$

Lemma 2: For any fixed $\bar{\rho} < \rho < \rho_{_M}$, the empirical ϵ-weak burstiness curve, $b_n(\epsilon, \rho)$, is a non-increasing function of ϵ. $\square$

Theorem 2: For any $0 \geq \epsilon \leq 1$, the empirical ϵ-weak burstiness curve, $b_n(\epsilon, \rho)$, is a convex decreasing function of ρ. $\square$

Theorem 3: The empirical ϵ-weak burstiness curve is a consistent estimator of the ϵ-weak burstiness curve, that is

$$\lim_{n\to\infty} b_n(\epsilon, \rho) = b(\epsilon, \rho). \quad \square \tag{15}$$

In the sequel, we will propose a water-filling algorithm to obtain the empirical ϵ-weak burstiness curve that satisfies (15).

A. Water-filling

If $\mu_n\Big(b_n(\epsilon, \rho)\Big) \leq \epsilon$ for all n, then one can guarantee that in the limit

$$\lim_{n\to\infty} \Pr\{q_n > b_n(\epsilon, \rho)\} \leq \epsilon. \tag{16}$$

We would like to obtain the smallest $b_n(\epsilon, \rho)$ that satisfies (16). Assume that the empirical ϵ-weak burstiness curve is selected such as to satisfy $\mu_n(b_n(\epsilon, \rho)) = \epsilon$ for all n. The solution to this problem is with "water-filling".

Consider the case $n = 1$ and define $\delta \triangleq \epsilon\rho$. It is straightforward to notice that (13) is satisfied for $b_1(\epsilon, \rho) = q_0 - \delta$. Now let $n = 2$. Without loss of generality, assume $q_0 > q_1$. There are two cases:

(i) $q_1 \leq q_0 - 2\delta$;

(ii) $q_1 > q_0 - 2\delta$.

The solution for case (i) is

$$b_2(\epsilon, \rho) = q_0 - 2\delta \tag{17}$$

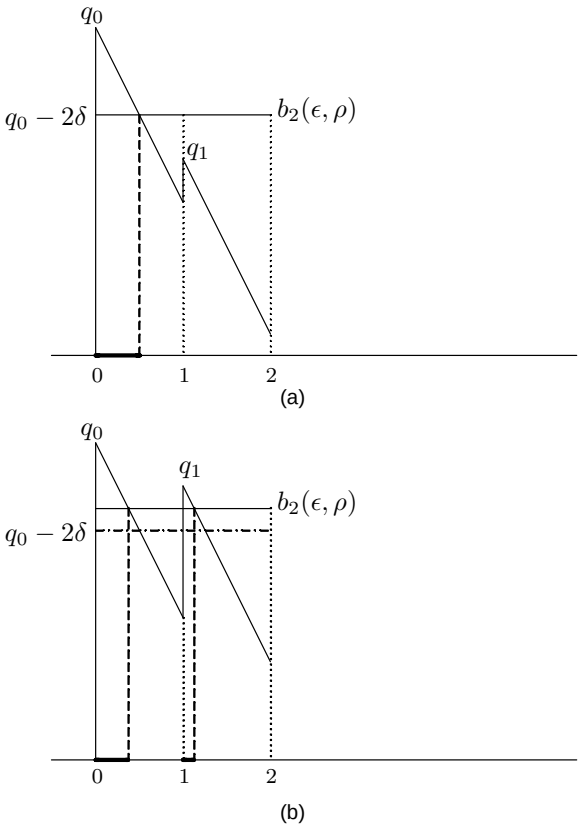

Fig. 2. The queue size and the empirical 0.25-weak burstiness thresholds for two consecutive samples of the queue size for: (a) $q_0 \geq q_1 + 2\delta$ and (b) $q_0 < q - 1 + 2\delta$.

and for case (ii) is

$$b_2(\epsilon, \rho) = \frac{q_0 + q_1}{2} - \delta. \tag{18}$$

Fig. 2 illustrates the two cases with their corresponding solutions for $\epsilon = 0.25$.

We continue this process by letting $n = 3$. Note first that any percentile of the queue size is independent of the ordering at which the queue sizes q_0, q_1, and q_2 arrive. One might visualize the queue size by considering the quadrangles with the maximum height q_i and the decreasing slope of the upper line ρ. The quadrangles can be arranged in any arbitrary order. Therefore, without loss of generality, one can assume $q_0 \geq q_1 \geq q_2$. There exist three cases:

(i) $q_2 \leq q_1 \leq q_0 - 3\delta$;

(ii) $q_1 \geq q_0 - 3\delta$ and $q_0 - 3\delta \leq q_2 \leq \frac{q_0+q_1-3\delta}{2}$;

(iii) $q_1 \geq q_0 - 3\delta$, and $q_2 \geq \frac{q_0+q_1-3\delta}{2}$.

For case (i), the empirical ϵ-weak burstiness curve is

$$b(\epsilon, \rho) = q_0 - 3\delta. \tag{19}$$

For case (ii), we have

$$b(\epsilon, \rho) = \frac{q_0 + q_1 - 3\delta}{2}. \tag{20}$$

And for case (iii), we have

$$b(\epsilon, \rho) = \frac{q_0 + q_1 + q_2 - 3\delta}{3}. \tag{21}$$

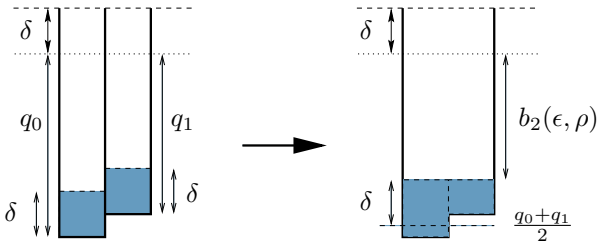

Fig. 3. The water-filling procedure for $n = 2$ with $q_1 < q_2 + 2\delta$.

This procedure can be continued for other values of n. We now devise a water-filling algorithm to obtain the empirical ϵ-weak burstiness curve at a fixed ρ.

The solution of the above problems can indeed be interpreted in terms of the concept of water-filling (see Fig. 3). Let the queue size over the interval $[i, i+1)$ be the linear decreasing function $Q_t(\rho) = q_i - (t-i)\rho$. Each time instant is represented by a container holding $\delta = \epsilon\rho$ units of liquid. The height of the container is equal to $q_i + \delta$. The liquid in the container, δ, is obtained by solving

$$\int_i^{i+1} 1\{Q_t(\rho) > q_i - \delta\}dt = \epsilon. \tag{22}$$

The whole liquid is distributed over all containers with the water-filling process. In the water-filling algorithm, the height of the empty portion of the container at any time instant n represents the empirical ϵ-weak burstiness curve $b_n(\epsilon, \rho)$.

IV. Numerical Results

In this section, we represent the simulation results. We apply the concept of empirical ϵ-weak burstiness curve to video traces. We use an MPEG-4 encoded video trace of the movie Star Wars IV which was taken from [12]. Fig. 4 illustrates the trace as a function of time. The total number of frames is 54000 with 30 frames arriving in each second. We assume that each frame is encapsulated in a single packet. Therefore, $\Delta = 1/30$ seconds. The maximum packet size is 9370 bytes and is located at the 154th frame.

The queue size for a single constant rate server has been illustrated in Fig. 5. The service rate is $\rho = 60$ KByte/second. The prominent peak in the queue size is due to the large size of the earlier packets of the trace.

The empirical ϵ-weak burstiness curve for $\epsilon = 0, 0.1, 0.2, 0.3, 0.4$ and 0.5 has been illustrated in Fig. 6. The curves have been obtained for $n = 54000$. As expected, the empirical ϵ-weak burstiness curve is a convex non-increasing function of the service rate ρ.

The empirical ϵ-weak burstiness curve of the movie has been shown in Fig. 7 in a semilog plot along with the delay lines corresponding to 10 msec and 100 msec delay limits. Different curves correspond to $\epsilon = 0, 0.01, 0.02, 0.03, 0.04$ and 0.05. As illustrated, there exists a fairly large difference between the empirical maximum burstiness curve and the empirical 0.01-weak burstiness curve. Therefore, using $\epsilon = 0.01$ instead of $\epsilon = 0$ will save the bandwidth. In this example, using $\epsilon = 0.01$ will save approximately 43% of

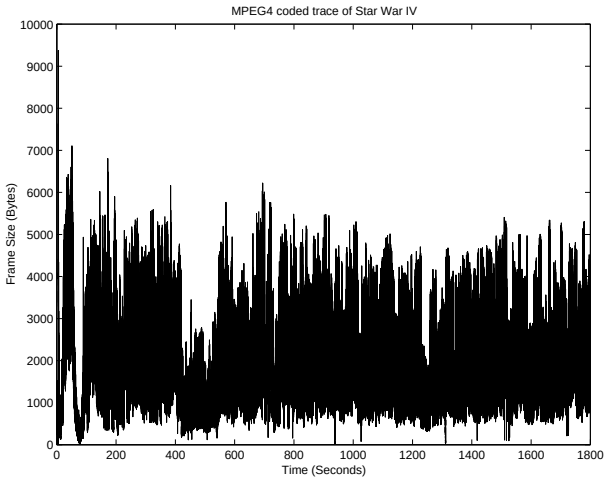

Fig. 4. The MPEG-4 encoded trace of the movie Star Wars IV.

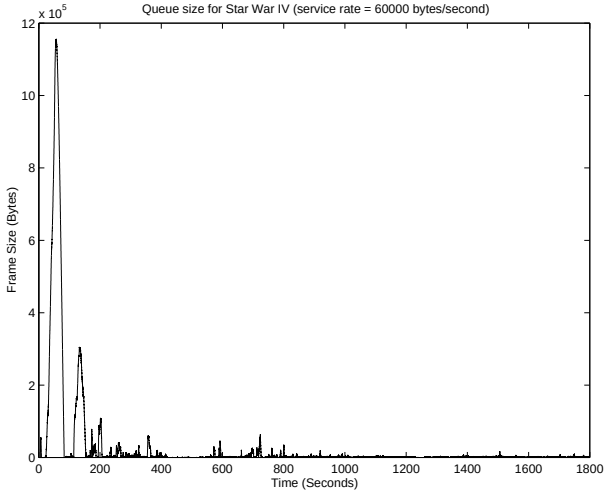

Fig. 5. The queue size of a constant rate server for the movie Star Wars IV. The service rate is 60000 bytes/second.

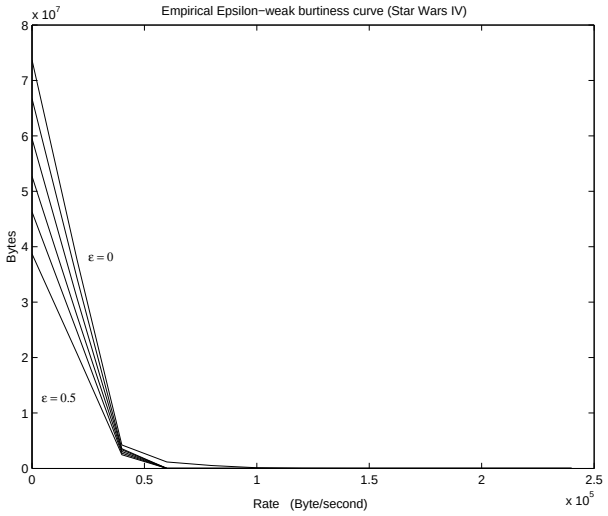

Fig. 6. The empirical ϵ-weak burstiness curve of the movie Star Wars IV for $\epsilon = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

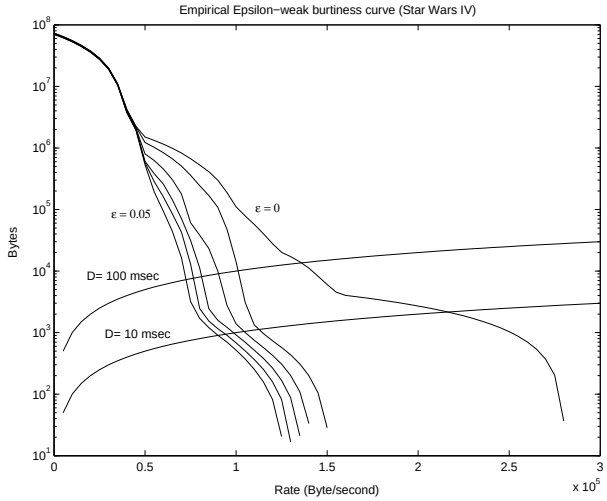

Fig. 7. Intersection of the empirical ϵ-weak burstiness curve of the movie Star Wars IV for $\epsilon = 0, 0.01, \ldots, 0.05$ and the delay lines of 10 msec and 100 msec.

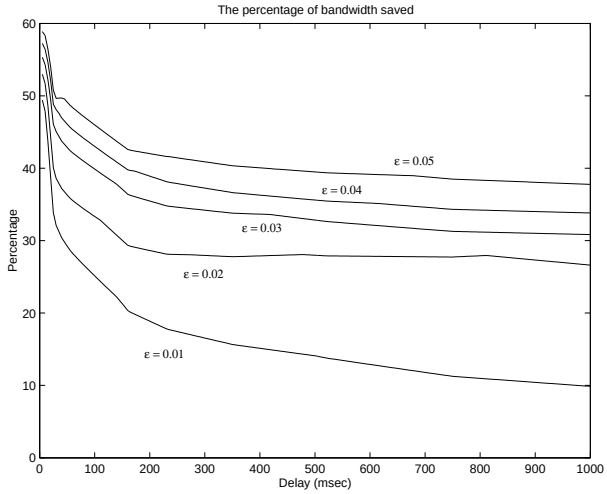

Fig. 8. The percentage of bandwidth saved, when the empirical ϵ-weak burstiness curve is used, as a function of the maximum delay.

the required bandwidth reducing from 220 KByte/second to 125 KByte/second for a maximum delay of $D = 10$ msec. In other words, with the service rate of 125 KByte/second for the trace, only 1% of the packets will be delayed beyond 10 msec.

Fig. 8 illustrates the percentage of bandwidth that can be saved if the empirical maximum burstiness curve is replaced by the empirical ϵ-weak burstiness curve. The curves have been found by locating the intersection of the empirical ϵ-weak burstiness curve and the delay lines. The percentage depends on the parameter ϵ. It is obvious that the required bandwidth decreases with ϵ.

V. Conclusion

In this paper, we have focused on the development of a new resource allocation scheme for video streaming in wireless networks. The proposed technique is based on the ϵ-weak burstiness curve. The ϵ-weak burstiness curve is the $(1-\epsilon)$-percentile of the queue size for a constant rate server. We have shown that the ϵ-weak burstiness curve is a convex non-increasing function of the service rate and also that the ϵ-weak burstiness curve of a multiplexed traffic is smaller than the summation of the corresponding curves of individual flows. Therefore, multiplexing can reduce the ϵ-weak burstiness curve.

We have devised a "water-filling" algorithm to obtain the empirical ϵ-weak burstiness curve, which provides a sequence of convex decreasing curves that converge to the true ϵ-weak burstiness curve. The proposed approach has been applied to MPEG-4 encoded video traces. It has been shown that, for these traces, considerable saving in terms of bandwidth and allocated buffer size can be obtained if the maximum burstiness curve is replaced by 0.01-weak burstiness curve.

References

[1] R. L. Cruz, "A calculus for network delay, part I: network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–131, Jan. 1991.

[2] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services network: the single node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 334–357, June 1993.

[3] D. E. Werge, E. W. Knightly, H. Zhang, and J. Liebeherr, "Deterministic delay bounds for VBR video in packet-switching networks: Fundamental limits and practical tradeoffs," *IEEE/ACM Trans. Networking*, vol. 4, pp. 352–362, June 1996.

[4] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," in *ACM SIGCOMM*, pp. 183–193, 1993.

[5] V. Paxson and S. Floyd, "Wide-area traffic: The failure of poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226–244, June 1995.

[6] I. Norros, "Studies on a model for connectionless traffic, based on fractional Brownian motion," *Conference on Applied Probability in Engineering, Computer and Communication Sciences*, June 1993.

[7] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," *ACM SIGCOMM*, September 1994.

[8] B. K. Ryu and A. Elwalid, "The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities," *ACM SIGCOMM Computer Communications Review*, vol. 26, pp. 3–14, October 1996.

[9] S. Valaee and J.-C. Gregoire, "An estimator of the regulator parameters in a stochastic paradigm," *in preparation*.

[10] S. Low and P. Varaiya, "A simple theory of traffic resource allocation in ATM," in *Proceeding IEEE Globecom*, pp. 1633–1637, 1991.

[11] S. Valaee, "A recursive estimator of worst-case burstiness," *IEEE/ACM Trans. Networking*, vol. 9, pp. 211–222, April 2001.

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