MinMax Bandwidth Allocation for Time-slotted Systems with Internet Traffic

J. Aracil[*], D. Morato, E. Magana and M. Izal
Dep. Automatics and Computer Science
Universidad Pública de Navarra
Campus Arrosadía s/n, 31006 Pamplona, SPAIN[†]

Abstract

The MinMax scheduling algorithm has been proved to be fair and efficient for dynamic bandwidth allocation in wireless scenarios. In this paper, MinMax is extended to the case of Internet traffic, which is bursty at all timescales. To do so, MinMax is equipped with a minimum square error estimate in order to improve bandwidth allocation accuracy. The proposed MinMax extension is named MinMaxPred*, which has been developed following a novel methodology that considers the traffic prediction and bandwidth scheduling problem jointly while, traditionally, they were considered as isolated problems.*

1. Introduction and problem statement

In the past, wireless networks, specially those satellite-based, were devoted to voice and video broadcasting, which demanded a fixed bandwidth. However, the increasing volume of best-effort traffic is making the standards change, with the aim of supporting best-effort Internet services. Basically, a fixed bandwidth allocation is not adequate for best-effort traffic due to its inherent burstiness. Alternatively, Demand-Assignment Multiple Access (DAMA) techniques are gaining increasing importance in a wide range of wireless networks scenarios, both terrestrial and satellite-based. These techniques are based on resource reservation on-demand, i. e. resource reservation messages are sent from the stations to the bandwidth scheduler, and the best-effort bandwidth is assigned per station according to the traffic demand. For example, the current Digital Video Broadcast (DVB) standard has undergone significant changes recently, with the introduction of a return channel from the user stations (DVB-RCS). In this way, the user stations (Very Small Aperture Terminals -VSATs-) are provided with reservation slots in the upstream frame and on-demand bandwidth allocation is made possible. In addition, the VSATs may act as bridges for LAN interconnection via satellite [7]. Furthermore, a growing number of satellites incorporate On-Board Processing (OBP) and remultiplexing, thus making it possible to reduce propagation delay to that of a single satellite hop.

However, an open issue is the bandwidth scheduling and multiple access technique to be adopted. Since traffic is expected to be bursty at all timescales (due to the self-similar features of Internet traffic) the stations must continuously report the traffic demand to the bandwidth scheduler using reservation mini-slots. Nevertheless, the number of reservation minislots is usually less than the number of user stations, in order to improve channel efficiency and network scalability. We will assume that stations have the chance to produce a reservation request once per RTT to the scheduler, i.e. a new bandwidth allocation request will not be sent before the response from the previous bandwidth allocation message has been received[1]. Figure 1 shows the scenario under analysis.

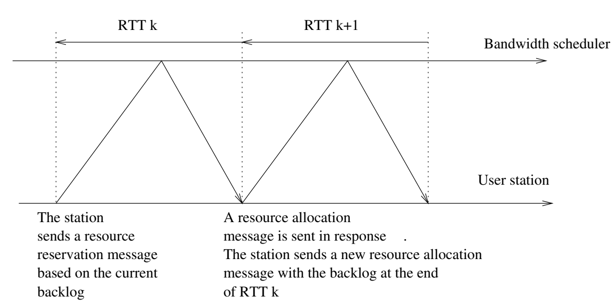

Figure 1. Network scenario

Note that the above scenario does not preclude that the frame duration may be less than a Round-Trip Time (RTT). This is usual for satellite networks, for instance. However, the number of reservation slots per station is reduced to one per RTT, thus enabling to shorten the control part of the

[*] On leave to Universidad Autónoma de Madrid

[†] This work was funded by Plan Nacional de I+D

[1] Processing time at the scheduler is neglected

frame. On the other hand, the upstream frames are reconfigured (due to bandwidth allocations) only once per RTT and both network control and synchronization are greatly simplified. In fact, since the processing power at the bandwidth scheduler is limited (specially in OBP scenarios) frequent frame reconfigurations have a negative impact on performance.

Despite of the many variants of DAMA protocols, the following approach can be adopted for modeling purposes [3, 8] (figure 1): Time is slotted in RTT-slots and, at the beginning of the current RTT-slot k, a reservation request is sent from the user station through reservation minislots, possibly involving contention among the stations. Then, a bandwidth allocation message is received in response from the bandwidth scheduler, with the allocated bandwidth for the next RTT-slot $k+1$ (figure 1). Since only traffic already backlogged at the beginning of RTT-slot k is covered by the reservation request message, traffic arriving *during* RTT-slot k is necessarily buffered until transmission in RTT-slot $k+2$[2]. Thus, in order to reduce access latency and increase channel utilization, the bandwidth allocation for the next RTT-slot should also include resources for traffic arriving at the current RTT-slot. However, in that case, bandwidth scheduling can only be performed with an *estimate* of the incoming traffic, since the reservation request message is sent prior to the arrival of packets for which the reservation is performed. This issue has been identified in a number of papers (see for instance [8]). Consequently, both bandwidth scheduling and traffic prediction algorithms are key to bandwidth allocation performance. Let λ_k be the amount of traffic (bytes) arriving during RTT-slot k. The simplest estimate $\hat{\lambda}_k$ is $\hat{\lambda}_k = \lambda_{k-1}$, i.e. traffic at the next RTT-slot will be the same as that in the current RTT-slot. However, Internet traffic is bursty at all timescales and abrupt load changes are expected to occur at the RTT-slot timescale, leading to inaccurate bandwidth allocation. An alternate approach, that is proposed in this paper, is to take advantage of the long-range dependence features of Internet traffic and employ a Linear Minimum Square Error Estimate (LMSEE). Such estimate has also been used to effectively tackle the traffic prediction problem in a number of network scenarios [1, 2, 11, 10]. However, traditionally, traffic prediction and bandwidth scheduling are performed in a sequential fashion. First, an estimate of the incoming traffic is provided by the traffic prediction algorithm. Then, such estimate is made available to the bandwidth scheduling algorithm. In the context of MinMax bandwidth scheduling and Internet traffic prediction, a departure from the traditional methodology is pursued in this paper. The *MinMax-Pred* algorithm is presented as a joint traffic prediction and bandwidth scheduling technique that provides, in a single step, the same performance that a traffic prediction technique and a bandwidth scheduling technique applied sequentially would yield. Furthermore, if sources have the same variance in the marginal distribution, a closed-form analytical expression for the bandwidth allocation per station can be obtained.

[2]The reservation request message covering traffic arriving during RTT-slot k is sent at the beginning of RTT-slot $k+1$.

1.1. MinMax scheduling

Minmax scheduling has been proposed in [3] as a fair algorithm that outperforms previous scheduling techniques in the context of DAMA schemes (Wireless ATM). Figure 2 ([3]) shows the dynamic capacity allocation model, a DAMA system supporting N sources on a Λ Mbps uplink. Let us assume that the channel is slotted and packets are constant length. Let $\lambda_k^i, i = 1, \ldots, N$ be the demand (bytes) from each source at the end of slot k. Let $x_k^i, i = 1, \ldots, N$ be the buffer occupancy at the end of slot k and let $F_k^i, i = 1, \ldots, N$ be the allocated bandwidth to source i in RTT-slot k. MinMax is based on the assumption that traffic rates change with a larger time constant in comparison with the RTT-slot duration, namely the demand for RTT-slot k is considered to be equal to the demand for RTT-slot $k-1$ ($\lambda_k^i = \lambda_{k-1}^i$). Then, the buffer dynamics at the source are given by

$$x_k^i = \max\{0, x_{k-1}^i + \lambda_{k-1}^i - F_k^i\} = \max\{0, x_{k-1}^i + \lambda_k^i - F_k^i\} \tag{1}$$

for $i = 1, \ldots, N$ where $\sum_{i=1}^N F_k^i = \Lambda$. The MinMax bandwidth allocation policy can be defined as follows.

Bandwidth allocation policy I (Min Max): At each RTT-slot, the task is to find $F_k^i, i = 1, \ldots, N$ values so that the quantity

$$\max_i \{(x_{k-1}^i + \lambda_{k-1}^i - F_k^i)^+\} \tag{2}$$

is minimized, where $(y)^+ = \max\{0, y\}$ for all y, and subject to

$$\sum_{i=1}^N F_k^i = \Lambda, \qquad F_k^i \geq 0 \tag{3}$$

It has been shown [3] that the MinMax allocation strategy provides very good performance in certain congestion scenarios. However, while the approximation $\lambda_k^i = \lambda_{k-1}^i$ is valid during congestion epochs [3] it may not hold if the incoming traffic is highly bursty, like in the Internet. In this paper, the MinMax algorithm is extended to the case of traffic changing at all timescales. First, an extension of MinMax is proposed which makes use of a LMSEE for self-similar traffic (Bandwidth allocation policy II). This is an

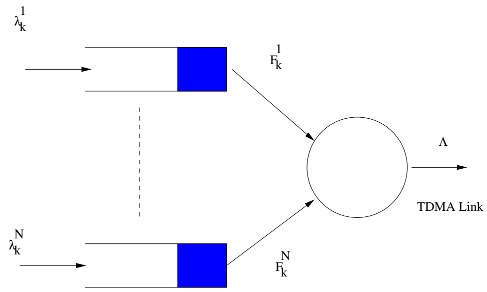

Figure 2. Bandwidth allocation model

improvement to MinMax, that results in smaller buffer occupancy in comparison to a MinMax bandwidth scheduler with no prediction (2). Secondly, the *MinMaxPred* resource allocation rule is presented (Bandwidth allocation policy III), that is equivalent to the bandwidth allocation policy II but has been developed with a novel approach that considers resource allocation and traffic prediction jointly. While the LMSEE $\hat{\lambda}_k^i, i = 1, \ldots, N$ is used as an *input* to the MinMax algorithm in bandwidth allocation policy II, the starting point of MinMaxPred is the application of the MinMax criteria to the a-posteriori random vector of traffic demands. Consequently, MinMaxPred introduces a novel methodology that serves to effectively bridge the gap between traffic prediction and resource allocation. Furthermore, MinMaxPred allows to obtain a closed analytical formula for the F_k^i if sources have the same variance.

The rest of the paper is organized as follows: section 2 presents the methodology and section 3 is devoted to analysis. Then, results and discussion are presented in section 4, followed by the conclusions that can be drawn from this research.

2. Methodology

Our methodology relies on two fundamental assumptions: First, the traffic marginal distribution is (truncated)Gaussian[3], secondly, the *Hurst parameter* remains constant. In order to better understand the rationale behind these assumptions let us briefly present the concept of self-similarity. Let $\{Z(t), t \in R\}$ be the continuous time process of number of bytes transmitted in the interval $[0, t)$ and consider the discrete-time process $\{\lambda_k = Z(k\delta) - Z((k-1)\delta), k \in N, k \geq 1, Z(0) = 0\}$, with δ being the frame duration. Note that λ denotes the (stationary) discrete process of number of bytes per frame. Now, consider the *aggregated* process

$$\lambda_{i(n)} = \frac{1}{n} \sum_{k=n(i-1)+1}^{ni} \lambda_k, \qquad n > 1, i \geq 1 \tag{4}$$

and let $\rho^{(n)}(j)$ with $j > 1$ be the autocorrelation function of $\{\lambda_{i(n)}, i = 1, 2, \ldots\}$. The process λ_k is *asymptotically second-order self-similar* if

$$\lim_{n \to \infty} \rho^{(n)}(j) = \frac{1}{2}((j+1)^{2H} - 2j^{2H} + (j-1)^{2H}) \tag{5}$$

where H is the Hurst (or self-similarity) parameter. For $1/2 < H < 1$ the autocorrelation function in (5) decays slowly, thus being not summable, and we call $\{\lambda_k, k \geq 1\}$ *long-range dependent*. Note that the index n in (4) defines a traffic timescale. On the other hand, equation 5 states that self-similarity is an asymptotic property, namely, it only happens when $n \to \infty$. In practice, there is a cutoff timescale beyond which the traffic behaves approximately as a stationary Gaussian self-similar process with constant H parameter [9], while the short timescales show complex, multifractal behavior. This behavior has been clearly identified in a number of recent studies [5] that confirm that there is no single characterization for traffic at all timescales. Intuitively, the number of packets per time interval can be arbitrarily small if we select a timescale small enough. Hence, for a very short timescale the marginal distribution of the arrival process is not Gaussian but discrete. As we increase the timescale, by the Central Limit Theorem, the statistical multiplexing of packets coming from a larger number of sources results in a Gaussian process. As the network bandwidth increases more packets from different sources can be accommodated in smaller timescales and for timescales beyond a certain cutoff value the number of bytes per interval are well modeled by a Fractional Gaussian Noise (FGN)[4]. Thus, for packet-switched networks the traffic dynamics at low timescales are relevant, specially at low or intermediate load [4]. However, in our case study, we are concerned with *the number of bytes per RTT-slot only*. Since a large number of packets can be accommodated in a slot duration (RTT) we may safely assume that the *number of bytes per RTT-slot* can be characterized as a FGN. On the other hand, it will also be assumed that sources have the same variance in the marginal distribution and same H parameter.

2.1. LMSEE for self-similar traffic

In this subsection, we provide the LMSEE that will be used along the paper. Our goal is to estimate $\lambda_k^i, i = 1, \ldots N$ with the information provided by $\lambda_{k-1}^i, \ldots, \lambda_{k-n}^i$,

[3] In what follows we will assume a truncation in the Gaussian distribution to the positive values

[4] An FGN is defined as the increments of a Fractional Brownian Motion [9].

with n being the process history that is used to forecast the next sample. Due to the process stationarity, the problem is equivalent to finding an estimate for λ_{n+1}^i provided that $\lambda_1^i, \ldots, \lambda_n^i$ are known. On the other hand, since $\{\lambda_k^i, k > 0\}$ is a Gaussian process any finite set of the random variables λ_k^i's is a multivariate Gaussian random variable with mean vector $\mu = (\mu^1, \ldots, \mu^N)$, variance σ^2 and covariance matrix $\mathbf{\Sigma} = \{S_{pq}\}$ with $p, q = 1, \ldots n$. For a FGN, the covariance matrix of the Gaussian multivariate random variable $(\lambda_1^i, \ldots, \lambda_n^i)$ is defined as follows:

$$S_{pq} = \frac{1}{2}\sigma^2 \left[(|p-q|+1)^{2H} - 2|p-q|^{2H} + (|p-q|-1)^{2H}\right] \tag{6}$$

with $p, q = 1 \ldots n$. The LMSEE $\hat{\lambda}_{n+1}^i$ is obtained as the mean of the a-posteriori distribution of $\lambda_{n+1}^i | (\lambda_1^i, \ldots, \lambda_n^i)$ [6, Theorem 3.3.1].

$$\hat{\lambda}_{n+1}^i = \mu^i + \mathbf{\Psi}_{21}{\mathbf{\Psi}_{11}}^{-1}(\lambda_1^i - \mu^i, \ldots, \lambda_n^i - \mu^i)' \tag{7}$$

where $\mathbf{\Psi}_{21} = (S_{(n+1)1}, \ldots, S_{(n+1)n})$ and $\mathbf{\Psi}_{11} = \{S_{pq}\}$ for $p, q = 1, \ldots n$.

3. Analysis

First, the bandwidth allocation policy II is presented (MinMax equipped with LMSEE for self-similar traffic), followed by the MinMaxPred algorithm (bandwidth allocation policy III).

Bandwidth allocation policy II (Min Max with LMSSE): At each RTT-slot, find $F_k^i, i = 1, \ldots, N$ values so that the quantity

$$\max_i \{(x_{k-1}^i + \hat{\lambda}_k^i - F_k^i)^+\} \tag{8}$$

is minimized subject to the restriction indicated by (3).

The former bandwidth scheduling algorithm outperforms MinMax simply because incoming bytes are estimated more accurately. Note that the development of bandwidth allocation policy II follows the traditional approach that considers traffic prediction and bandwidth scheduling as sequential procedures. In fact, this bandwidth allocation policy is the result of applying bandwidth allocation policy I (MinMax) to the LMSEE $\hat{\lambda}_k^i$ (7).

In order to define MinMaxPred, let $Y_k = (Y_k^1, \ldots, Y_k^N)$ be the *unsatisfied demand at RTT-slot* k, namely $Y_k^i = x_{k-1}^i + \hat{\lambda}_k^i - F_k^i, i = 1, \ldots, N$. On the other hand, let $Z_k = \max(Y_k^1, \ldots, Y_k^N), k = 1, 2, \ldots$.

Bandwidth allocation policy III (*MinMaxPred*): At each RTT-slot, find F_k^i values so that the quantity

$$F_{Z_k}(c) = P(Z_k \leq c) = P(\max(Y_k^1, \ldots, Y_k^N) < c) \tag{9}$$

is maximized, where c is a positive real value.

Note that Z_k is a random variable, which is equal to the Nth ordered statistic of $Y_k^1, \ldots, Y_k^N$. On the other hand, Y_k^i are independent Gaussian random variables with mean $x_{k-1}^i + \hat{\lambda}_k^i - F_k^i$, for $i = 1, \ldots, N$. The following lemma constitutes the cornerstone for MinMaxPred.

Lemma 1 *If all sources have the same variance in the marginal distribution and H parameter, the bandwidth allocation policies II and III are equivalent.*

The proof is given in the Appendix. The previous lemma states that MinMaxPred is the MinMax bandwidth allocation policy with LMSEE. However, MinMaxPred is developed from the a-posteriori random vector of demands $(Y_k^1, \ldots, Y_k^N)$. This methodology allows to obtain a closed expression for the bandwidth allocations $F_k^i, i = 1, \ldots, N$, as stated in the following lemma.

Lemma 2 *If all sources have the same variance and H parameter then, for each RTT-slot k, bandwidth allocation policies II and III yield*

$$F_k^i = \left(x_{k-1}^i + \hat{\lambda}_k^i + \frac{1}{N}\left(\Lambda - \sum_{k=1}^{N}\left(x_{k-1}^i + \hat{\lambda}_k^i\right)\right)\right)^+ \tag{10}$$

The proof is given in the Appendix.

4. Results and discussion

A simulation model of the system depicted in figure 2 has been used in order to i) verify the analytical results and ii) evaluate the performance of the different bandwidth allocation policies. The simulator is based on (1) and has been written in Scilab script language. Sources have been modeled as independent Fractional Gaussian Noises (FGNs). The FGN parameters are set to those inferred from the *Bellcore traces* (Hurst parameter $H = 0.78$), which have also been used in other studies [9, 11, 10]. The coefficient of variation $c_v^2 = \sigma^2/\mu^2$ is set to 0.2 and 0.3, according to previously reported figures. Variance is left constant and the mean μ is equal to 1.3 and 1.55 Mbps respectively. The different load conditions are achieved by varying the link bandwidth accordingly, with a fixed number of sources equal to 10.

The average number of bytes in queue is depicted in figure 3. The results showed that the use of bandwidth allocation policies II/III is advantageous with respect to MinMax. This improvement was expected since the LMSEE

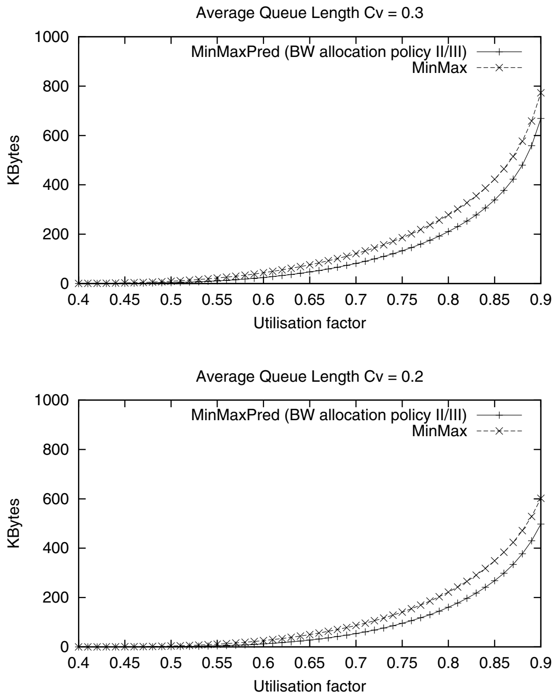

Figure 3. Average queue length

(7) was used to forecast incoming traffic, whereas MinMax assumed that the traffic rate was constant from one RTT-slot to the following. On the other hand, and most importantly, the bandwidth allocations F_k^i were obtained with the closed expression (10), thus alleviating the processing burden at the bandwidth scheduler. Finally, the results confirmed that bandwidth allocation policies II and III(MinMaxPred) are equivalent[5].

5. Conclusions

In this paper, a novel BW allocation policy (MinMax-Pred) has been introduced that extends MinMax to the case of bursty Internet traffic. Furthermore, MinMaxPred has been developed using a methodology which departs from traditional schemes and considers traffic prediction and scheduling jointly. We believe such methodology will be most useful in the development of future bandwidth scheduling algorithms.

A. Proof of lemma 1

The following preliminary lemma will be used to show lemma 1.

[5]Simulation results for both BW allocation policies are exactly the same. Thus, only one curve is displayed in figure 3.

Lemma 3 *Let $F(x_1, \ldots, x_N) = \sum_{i=1}^{N} log(x_i)$, with $x_i > 0, i = 1, \ldots, N$. Subject to the restriction $\sum_{i=1}^{N} \delta_i = K, K > 0$, the N-tuple $(\delta_1, \ldots, \delta_N)$ that makes $F(x_1 + \delta_1, \ldots, x_N + \delta_N)$ reach a maximum is the same that makes $\min_i\{x_i + \delta_i\}$ reach a maximum.*

Proof. The problem is to maximize a concave objective function under a single linear constraint. Hence, a simple exchange argument shows that in order for $\sum_{i=1}^{N} log(x_i)$ to be maximized, for given $x_i > 0, i = 1, \ldots, N$ it must be the case that if $x_i + \delta_i > x_j + \delta_j$ then $\delta_i = 0$, i. e., $\min_i\{x_i + \delta_i\}$ is maximized.

Now, lemma 1 will be proved. Since sources are assumed to be independent then:

$$P(Z_k \leq c) = P(\bigcap_{i=1}^{N}\{Y_k^i \leq c\}) = \prod_{i=1}^{N} P(Y_k^i \leq c) \quad (11)$$

for all $k > 0$. Since the logarithm is a monotonically increasing function the maximum of $P(Z_k \leq c)$ is attained at the same value that $log(P(Z_k \leq c)) = \sum_{i=1}^{N} log(P(Y_k^i \leq c))$. By lemma 3 the N-tuple $(\delta_1, \ldots, \delta_N)$ that makes $\sum_{i=1}^{N} log(P(Y_k^i \leq c))$ reach a maximum is the same that makes $\min_i\{P(Y_k^i \leq c) + \delta_i\}$ reach a maximum. Consequently, in order to maximize $P(Z_k \leq c)$ (BW allocation policy III) the maximum variation is obtained by maximizing the minimum of $P(Y_k^i \leq c), i = 1, \ldots N$. Since the variables Y_k^i are Gaussian with the same variance and mean $x_{k-1}^i + \hat{\lambda}_k^i - F_k^i$ this is equivalent to finding the F_k^i values so that the quantity

$$\max_i\{x_{k-1}^i + \hat{\lambda}_k^i - F_k^i\} \quad (12)$$

is minimized, which is precisely the bandwidth allocation policy II.

B. Proof of lemma 2

Lemma 1 shows that bandwidth allocation policies II and III are equivalent. Let us take bandwidth allocation policy III. The problem is to find, for all k, the bandwidth allocation $F_k^i, i = 1, \ldots, N$ that maximizes

$$log(P(Z_k \leq c)) = \sum_{i=1}^{N} log(P(Y_k^i \leq c)) \quad (13)$$

subject to

$$\begin{aligned} \sum_{i=1}^{N} F_k^i &= \Lambda \\ F_k^i &\geq 0 \end{aligned} \quad (14)$$

We use the Lagrange Multipliers Theorem to find the maximum. Note that the critical points are those that make the partial derivatives of

$$G(F_k^1,\ldots,F_k^N)=log(P(Z_k\le c))-\psi\left[\sum_{i=1}^{N}F_k^i-\Lambda\right] \tag{15}$$

be equal to zero, with $\psi > 0$, i.e.,

$$\frac{\partial G(F_k^1,\ldots,F_k^N)}{\partial F_k^i}=0 \qquad i=1,\ldots N \tag{16}$$

Since the function in (15) is continuous and the region defined by (14) is compact then a maximum is necessarily attained in the region. First, it will be shown that (16) has only one solution that fulfills (14). Since

$$P(Y_k^i\le c)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\frac{c-(x_{k-1}^i+\hat{\lambda}_k^i-F_k^i)}{\sigma}} e^{-\frac{x^2}{2}}dx \tag{17}$$

then (16) can be written as

$$\frac{\partial log(P(Y_k^i\le c))}{\partial F_k^i}=\frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{c-(x_{k-1}^i+\hat{\lambda}_k^i-F_k^i)}{\sigma}\right)^2}}{\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\frac{c-(x_{k-1}^i+\hat{\lambda}_k^i-F_k^i)}{\sigma}} e^{-\frac{x^2}{2}}dx}=\psi \tag{18}$$

for $i=1,\ldots,N$. On the other hand, the functions

$$\begin{aligned} f(y) &= \frac{e^{-\frac{y^2}{2}}}{\int_{-\infty}^{y} e^{-\frac{x^2}{2}}dx} \qquad (19)\\ g_k^i(y) &= \frac{c-(x_{k-1}^i+\hat{\lambda}_k^i-y)}{\sigma} \end{aligned}$$

are monotonically decreasing (f) and monotonically increasing (g_k^i) then $\frac{\partial log(P(Y_k^i\le c))}{\partial F_k^i} = (f\circ g_k^i)(c), i = 1,\ldots,N$ is monotonically decreasing for $i=1,\ldots N$ and, thus, the solution of system (18) is unique and can be found by inspection to be

$$\frac{c-(x_{k-1}^i+\hat{\lambda}_k^i-F_k^i)}{\sigma}=\frac{c-(x_{k-1}^1+\hat{\lambda}_k^1-F_k^1)}{\sigma} \tag{20}$$

for $i = 1,\ldots,N$. Apply the boundary condition $\sum_{i=1}^{N}F_k^i=\Lambda$ (14) to obtain

$$F_k^i=\left(x_{k-1}^i+\hat{\lambda}_k^i+\frac{1}{N}\left(\Lambda-\sum_{k=1}^{N}\left(x_{k-1}^i+\hat{\lambda}_k^i\right)\right)\right)^+ \tag{21}$$

which is precisely equation 10.

References

[1] A. Adas. Supporting real time VBR video using dynamic reservation based on linear prediction. In *IEEE INFOCOM 96*, San Francisco, CA, 1996.

[2] J. Aracil and D. Morató. Characterizing Internet load as non-regular multiplex of TCP streams. In *Proceedings of IEEE ICCCN 00*, Las Vegas, NV, October 2000.

[3] S. Biswas and R. Izmailov. Design of a fair bandwidth allocation policy for VBR traffic in ATM networks. *IEEE/ACM Transactions on Networking*, 8(2):212–223, 2000.

[4] A. Erramilli, O. Narayan, A. Neidhart, and I. Saniee. Performance impacts of multi-scaling in wide area TCP traffic. In *IEEE INFOCOM 00*, Tel Aviv, Israel, 2000.

[5] A. Feldmann, A. C. Gilbert, and W. Willinger. Data networks as cascades: Investigating the multifractal nature of internet WAN traffic. In *SIGCOMM*, pages 42–55, 1998.

[6] B. Flury. *A first course in Multivariate Statistics*. Springer Verlag, 1997.

[7] Y. F. Hu, G. Maral, and E. F. (Editors). *Service Efficient Network Interconnection via Satellite-EU COST action 253*. John Wiley and Sons, 2002.

[8] Z. Jiang, Y. Li, and V. C. Leung. A predictive demand assignment multiple access protocol for broadband satellite networks supporting internet applications. In *IEEE ICC*, New York, NY, June 2002.

[9] I. Norros. On the use of Fractional Brownian Motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13(6):953–962, August 1995.

[10] S. A. N. Ostring and H. Sirisena. The influence of long-range dependence on traffic prediction. In *Proceedings of ICC 2001*, 2001.

[11] A. Sang and S. Li. A predictability analysis of network traffic. In *Proceedings of IEEE INFOCOM 2000*, 2000.