Journal of Physics: Conference Series **490** (2014) 012110 doi:10.1088/1742-6596/490/1/012110

Optimal capacity and buffer size estimation under Generalized Markov Fluids Models and QoS parameters

José Bavio and Beatriz Marrón

Universdidad Nacional del Sur, Av. Alem 1253, Bahía Blanca, Buenos Aires, Argentina

E-mail: jmbavio@yahoo.com.ar - beatriz.marron@uns.edu.ar

Abstract.

Quality of service (QoS) for internet traffic management requires good traffic models and good estimation of sharing network resource.

A link of a network processes all traffic and it is designed with certain capacity C and buffer size B.

A Generalized Markov Fluid model (GMFM), introduced by Marrón (2011), is assumed for the sources because describes in a versatile way the traffic, allows estimation based on traffic traces, and also consistent effective bandwidth estimation can be done.

QoS, interpreted as buffer overflow probability, can be estimated for GMFM through the effective bandwidth estimation and solving the optimization problem presented in Courcoubetis (2002), the so call inf-sup formulas.

In this work we implement a code to solve the inf-sup problem and other optimization related with it, that allow us to do traffic engineering in links of data networks to calculate both, minimum capacity required when QoS and buffer size are given or minimum buffer size required when QoS and capacity are given.

1. Introducción

Modeling different digital sources is a wide topic. Markov models have been studied for many kinds of sources like ON/OFF sources or video sources but they have some limitations when the data rate transfers can take too many value. In this work a Generalized Markov Fluid model, introduced in [1] is used to describe the traffic of each source.

For this model, a sources is a data network assumes the state Z_s at time s, where Z is a continuous time, homogeneous and irreducible Markov chain, with finite state space $K = \{1, \cdots, k\}$, invariant distribution π and infinitesimal generator $\mathbb{Q}^Z$. Let us consider $f_1, f_2, \cdots, f_k$, k laws of probability with known and disjoint support. When the chain Z reaches states i, at time s, the speed with which the sources transfer data it is drawn, independently of the chain Z, by the law f_i. This is, the random variable $Y_s|Z_s = i$, is distributed according the probability law f_i for $i = 1, \cdots, k$.

This new model can be interpreted as follows: the state of the chain indicate some type of activity in the transfer data such as email, chat, conversation, video conferences, etc., and the rate for each state is drawn according to a probability law, within a range of reasonable values for such activity.

The process Y takes the drawn value as long as the chain Z remains in that state, if the modulating chain change state, a new value for Y is drawn. Let us note that the process Y_s is observable and, since the supports of the k laws of probability are known and disjoint, so is the process Z_s.

The Markov flow modulated by the chain Z_s that represent the work load received from the source that delivers information with speed Y_s is

$$X_t = \int_0^t Y_s \, ds. \tag{1}$$

Given an expected QoS, interpreted as the probability of buffer overflow, the actual resources that should be reserved lie between the mean rate and the peak rate of the connection. These resources are generally referred to as the Effective Bandwidth (EB) of the traffic sources, was proposed by Kelly in [2] and is defined as follows.

Let X_t be a process with stationary increments, representing the amount of work arriving from a source in the interval $[0, t]$, then the EB of the source is

$$\alpha(s,t) = \frac{1}{st} \log E e^{sX_t} \qquad 0 < s, t < \infty, \tag{2}$$

where the parameters s and t characterize a link's operating point and depends on the context of the stream. Specifically, the space parameter s indicate the degree of multiplexing and the time parameter t corresponds to the most probable duration of the buffer busy period prior to overflow.

2. Effective bandwidth

We can compute the EB for the GMFM by Kesidis,Walrand, Chang formula [3] by the following theorem,

Theorem 1 *Let $\{X_t\}_{t\geq 0}$ be a GMFM modulated by a continuous time, homogeneous and irreducible Markov chain Z with invariant distribution π and infinitesimal generator $\mathbb{Q}^Z$. Let us consider the random variables Y_i with density function f_i, mean μ_i and variance σ_i^2 for $i = 1, \cdots, k$ and the k dimensional diagonal matrix $\mathbb{H}$, whose nonzero elements are the first moments μ_i of each distribution, then*

$$\alpha(s,t) = \frac{1}{st} \log \left\{ \pi \exp \left[\left(\mathbb{Q}^Z + s\mathbb{H} \right) t \right] \mathbf{1} \right. \quad , \tag{3}$$

where $\mathbf{1}$ is a column vector with all entries equal to 1.

The importance of this theorem is that it provides an expression for the EB that depends on the infinitesimal generator of the modulating chain, its invariant distribution and a matrix containing information of the transfer rate, and all these elements can be estimated with traffic traces. In [1] a consistent and asymptotically Gaussian estimator is obtained from traffic traces.

3. Operational point

Each node or link of the network has two major design parameters, capacity and buffer size, and one QoS parameter buffer overflow probability (BOP).

The EB is relationally with the probability of overflow by the called $\inf - \sup$ formula $\lim_{N\to\infty} \frac{1}{N} \log P(Q_N > B) = -\gamma$, with

$$-\gamma = \inf_{t\geq 0} \sup_{s\geq 0} \left\{ (b + ct)s - st\alpha(s,t) \right\}, \tag{4}$$

where Q_N is the stationary workload in tail, c is the capacity of the node, b is the buffer size, N is the number of sources in the link and $\alpha(s,t)$ is its EB of the link [4]. The values t^* and s^* in which the $\inf - \sup$ is reach is called the link operating point.

The parametric approach it is possible with a model for the traffic, and our aim is to estimate γ, t^* and s^* using an estimator of $\alpha(s,t)$.

In [1] is proved that if $\alpha^n(s,t)$ is an estimate of $\alpha(s,t)$, then replacing $\alpha^n(s,t)$ in (4) we obtain (s_n^*, t_n^*) that are consistent estimators of (s^*, t^*).

Finally we can estimate γ with γ_n using α_n and solving the $\inf - \sup$ equation.

4. Algorithm and Code

In the paper [5] there is an algorithm to solve (4). First call $F(s,t) = (b+ct)s - st\alpha(s,t)$. The algorithm involves two optimization procedures:

(i) find for a fixed time, the maximum $F^*(t) = \max_s F(s,t)$,

(ii) find the minimum of $M = \min_t F^*(t)$.

The first step can be numerically solved in an efficient manner by taking into account that the logarithmic moment generating function $st\alpha(s,t)$ is convex in s, whereas $s(b+ct)$ is linear in s. Due to this, $F_t(s) = F(s,t)$ is an unimodal function of s and the maximizer is unique. To find the maximum one can start from an initial "uncertainly" interval $[s_a, s_b]$ that contains the maximum and decrease it using a golden section search until its length is less than some small value ϵ.

Unlike the function $F_s(t)$, there is no general property for $F^*(t)$ that we can exploit in order to perform minimization $\min_t F^*(t)$ efficiently. For this reason, the minimization is solved by linearly searching the values of t in the interval $[0, \rho]$. The value of ρ is determined empirically and depends on the buffer size. This value indicates the time scale at buffer overflow occurs.

4.1. Code

We present here an example of effective bandwidth estimation solving the $\inf - \sup$ problem.

First of all we generate a trace from a GMFM constructing a Markov chain of 13 states where each state correspond to one of the transfer speed range:

```
transf_range=[0 64 128 256 512 1024 2048 3072 4096 5120 6144 7168 8292 10240];
```

this is, when the chain is in state 1, the transfer rate of the trace lies between $[0; 64]$ bits per second, in state 2 between $[64; 128]$, and so on.

The higher transfer rate available in the transmission channel, in our case the state 13, is expected to be the usual state and is also usual it jump from one state to neighboring states. With these considerations infinitesimal generator is designed to simulate the model. The speed is choose randomly within it range with a truncated gaussian centered.

As we know when the trace is within a range, the modulating chain is in certain state, to estimate the infinitesimal generator we count the jumps, from one state to another in a in increasing window time. We also estimate the average rate of each state by averaging the states that the trace visits in the corresponding range in an increasing window time.

With the infinitesimal generator estimate (Q_{estim}) and the average rate estimated ($gammas$) we can calculate the effective bandwidth from 3:

```
pi_est=expm(Q_estim*100);
alpha=@(s,t) 1/(s*t)*log(pi_est(1,:)*expm((Q_estim+diag(gammas)/1000*s)*t)*ones(13,1));
```

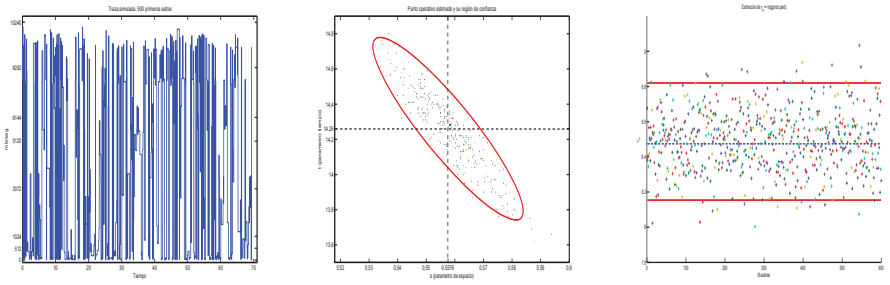
Now we can solve the $\inf - \sup$ problem: For fix t_0 you find the maximum s_{star}, and evaluate the function F.

```
F=@(s,t) s*(B+C*t)-log(pi(1,:)*expm((Q+h*s)*t)*ones(13,1)); %The last part
%coincides with the alpha
g=(sqrt(5)-1)/2; epsilon=0.01;  %Golden Ratio and Error threshold
%The following while structure performs the golden search
while s_min-s_max>epsilon
gamma=g*(s_max-s_min); sl=s_max-gamma; sr=gamma+s_min;
Fsl=subs(F,[s,t],[sl,tfijo]); Fsr=subs(F,[s,t],[sr,tfijo]);
    if Fsl>Fsr ;    s_max=sr;
    elseif Fsl<Fsr ; s_min=sl;
    elseif Fsl==Fsr
        s_min=sl; s_max=sr;     end;    end; s_star=(s_min+s_max)/2
```

The value s_{star} is the maximum in the variable s when $t = t_0$. This gives three numbers $[s_{\text{star}}, t_0, F(s_{\text{star}}, t_0)]$.

We have to repeat the procedure until a minimum in is found. This may be a long loop, but always possible to reach. Finally we have found the link operating point, the BOP and the parameters of s^* and t^*.

The above results can be extended to the design of a network link that requires certain quality of service. The minimum buffer size of a link can be calculated for a given capacity of the link, the traffic and the maximum loss probability desired. In a similar way, having the same information as before, but defined the desired buffer size, it is possible to calculate minimum link capacity required to ensure certain probability of loss. In Figure 1 we show a trace, the estimation of the operational point and the estimation of the probability of loss.



(a) (b) (c)

Figure 1. (a) GMFM trace. (b)Operational point estimation. c) Probability of loss estimation.

5. Conclusions and Remarks

In this work we want to give an insight of the traffic engineering process given QoS parameters, and presenting an application to Generalized Markov Fluids exploiting the availability of "good" effective bandwidth estimators for this source class. Knowledge on this subject allows better (QoS) networks design and resource utilization, two important features in telecommunications.

Code presented in section 4, is developed in Matlab environment an, due to space requirements, it is an extract to illustrate some key parts of the algorithm. For a fully compilable version contact the authors at *jmbavio@yahoo.com.ar*.

References

[1] Marrón B 2012 *Estadística de Procesos Estocásticos aplicados a Redes de Datos y Telecomunicación* Ph.D. thesis Departamento de Matemática, Universidad Nacional del Sur
[2] Kelly F 1996 Notes on effective bandwidths
[3] Kesidis G, Walrand J and shang Chang C 1993 Effective bandwidths for multiclass markov fluids and other atm sources
[4] Courcoubetis C and Weber R 1995 *Journal of Applied Probability* **33** 886–903
[5] Courcoubetis C and Siris V A 2002 *Perform. Eval.* **48** 5–23