Optimal Service Schedule for CBR Traffic in a High Speed Network

N.Sai Shankar A.P.Shivaprasad
Department of Electrical Communication Engineering.
Indian Institute of Science.
Bangalore - 560012. INDIA
Phone : +91-80-3092656, Fax : +91-80-3340563
Email : {nsai,aps}@ece.iisc.ernet.in

Abstract

This paper studies optimal service schedule of CBR traffic when the server rate is modelled according to a Markov process. The CBR traffic is mixed with traffic of another session(interfering traffic) which might be either CBRs or VBRs(Variable Bit Rate) or both. The CBR traffic can suffer a maximum delay of T. The scheduler must decide which session must be scheduled for service based on future server rate, estimated from the currently available information. The objective of this paper is to find an optimal time $t_s < T$ for switching the service from interfering traffic to CBR traffic. This switching of the scheduler from one traffic to another traffic takes into account the stochastic variation of the server rate. Three cases of variation of the source rate are considered; viz. (1) when the server rate is a constant, (2) when the server rate jumps randomly between two values and (3) when the variation of server rate follows a diffusion process. In all the three cases, this paper shows that there exists an optimal switching time t_s after which the low priority CBR traffic can be scheduled for service. The maximum delay and the cell loss rate are determined.

1 Introduction

The quality of real-time voice and video is most affected by the data loss and delay jitter. Therefore, the combination of maximum allowable delay T and the fraction of data not delivered within this delay is a measure of network quality of service(QOS) for a real-time session. This paper studies the optimal service schedule of the constant bit rate(CBR) session, when the Multiplexer processing rate(server rate) varies in accordance to a Markov process. We will use server rate instead of multiplexer processing rate throughout this paper. The scheduler must decide which session must be scheduled for service based on future server rate, which is estimated using currently available information. The two sessions considered in this paper are (1) CBR session and (2) interfering traffic session. The interfering traffic session might be combinations of CBRs and VBRs. These sessions arrive at the multiplexer which has a buffer of size B. The buffer size is assumed to be infinite in this paper.

The fluid traffic model is adopted as opposed to the point process model, so that the system model considered is suited for Asynchronous Transfer Mode (ATM) networks. The fraction of CBR traffic that is not delivered within time T is estimated. The bit-rate of CBR traffic is denoted by c and the server rate is modelled as a constant, Markov-modulated and diffusion process. The interfering traffic satisfies the average rate and burstiness constraint defined in [1] as the (σ,ρ)-constraint, which is the constraint of the traffic shaped by leaky bucket regulation as in [2]. The multiplexer buffer size B is assumed to be infinite so that there is no loss of incoming traffic. Various priority schemes [6]- [13] have been proposed for use as a scheduling method at a switching node in an ATM network. The simplest priority scheme is the static priority scheme. In this scheme, the delay sensitive class is always scheduled for service. While providing relatively low delays for delay sensitive traffic, this scheme causes relatively high losses for the loss sensitive class of traffic in case of finite buffer. If the large portion of the network traffic is high priority traffic then the quality of service(QOS) of

low priority traffic is severely degraded. [3] and [4] discuss the longest packet delay under (σ,ρ)- constraints. [5] considers the loss rate of CBR traffic suitable for real-time audio and video. In [5] the optimal schedule of traffic streams was not considered and interfering traffic stream had preemptive priority over CBR stream. In our work, we assign service to both interfering and CBR session and the loss of CBR is minimized by scheduling the server to the CBR traffic session in an optimal way. It is shown that the maximum delay suffered by the CBR traffic is $\leq\ t_s$ and there is no other time t less than t_s.

The optimal schedule of CBR traffic for service is analyzed by stochastic dynamic programming. Three cases of varying source rates are considered; in the first case, the server rate is constant. In the second case, the server rate jumps between a high and a low level. The jump is represented by a two state Continuous Time Markov Chain(CTMC). In the third model, the server rate varies continuously following a diffusion process. In both the cases, the scheduler must make the decision of scheduling either the CBR stream or the interfering traffic stream based on its estimate of future server rate obtained from current rate.

2 System Model

The system considered in this paper comprises of a $N \times N$ ATM switch. The switch consists of input as well as output buffers. Each input port has a buffer where the cells arrive. The input side of the switch is called the source side and the output side of the switch is called the server side. In our paper, the output buffer is called the multiplexer buffer because different input ports in the switch may wish to send data to a particular output port under consideration. This buffer is fed by CBR traffic and the interfering traffic. The CBR traffic is fed at a rate of c bits per second and the interfering traffic is fed at the rate $b(t)$ with the (σ,ρ)-constraint explained in [1] . The server rate is modelled by Markov-modulated fluid sink *i.e.*, where the instantaneous rate at which the fluid is released is governed by the the state of the CTMC. Such a server is characterized by state-space(o) and generator(O) of the modulating Markov chain, and the vector of fluid release rates $\mu(=\{\mu(i), i \in O\})$. The state of server captures the variations of cross traffic, link failures, etc. Note that the case of constant service rate is a special case.

2.1 Interfering Traffic Model

Let $b(t)$ denote the amount of interfering data that has arrived at the output buffer up to time t. The (σ,ρ)-constraint is defined by $b(t') - b(t) \leq \rho(t' - t) + \sigma, \forall t < t'$. Interfering data is fed into a bucket of size σ at a constant rate ρ. For stability of the server queue, we assume that $c+\rho \leq \mu,\ \forall t$, where μ is the server rate. The interfering data is fed from bucket to the server buffer, and this feeding can take place at an arbitrary rate as long as the bucket is not empty, including zero rate and bulk feeding. Bulk feeding makes $b(t)$ discontinuous. In our paper, we assume that $b(t)$ is right continuous. When the bucket is empty the data cannot be fed into the multiplexer at a rate higher than ρ. When the bucket is full, if the data fed into the multiplexer is less than ρ, the bucket overflows; the overflowing data is not fed into the multiplexer at all and is assumed to be lost. Thus the uncertainty of the interfering traffic entering the multiplexer within the (σ,ρ)-constraint can be viewed as the uncertainty of the momentary rate at which the interfering traffic is fed from bucket into the multiplexer. This uncertainty in the interfering traffic may result in the CBR traffic being denied access to the server because of its lower priority.

3 Problem Formulation

As explained before, the source serves the incoming traffic at a rate determined by the underlying stochastic process. The CBR traffic that is present in the multiplexer buffer must be sent out before time T in order

to be reproduced without delay and loss at the receiving end. If the CBR traffic does not abide by the delay QOS, it is considered as lost at the destination. So the scheduler must decide the time at which to switch the scheduling of service from interfering traffic to CBR traffic before time T. Assuming that interfering traffic is always scheduled first for service, because interfering traffic is assigned higher priority than the CBR traffic, our work explores the optimal switching time, $t_s < T$, after which the CBR traffic will be scheduled for service. Let $I(t)$ be the amount of interfering traffic served by the server till time t. The total amount of interfering traffic sent out increases as t increases but saturates for a large t, because the scheduler has to schedule the CBR traffic to service before time T. By definition $I(t)$ is monotonically increasing function. Assume that the network is earning revenue when it schedules interfering traffic for service. The function depicting the revenue earned by the network in serving the interfering traffic is given by

$$G[\mu, I(t)] = \frac{aI(t)\mu}{1 + bI(t)} \tag{1}$$

where μ is the server rate and a, b are constants. $G[\mu, I(t)]$ can be thought of as the net returns the network obtains because of sending out interfering traffic. The saturation of the above function indicates that there is no use in scheduling the interfering traffic session beyond a certain point before T. After the function saturates the scheduler can schedule CBR session for service. Although there are many saturating functions, the above function was chosen so that the saturation takes place for large $I(t)$ keeping in view of the higher priority assigned to the interfering traffic session. Different saturating functions can be got by changing the values of the constants a and b. So the above function ensures that enough interfering traffic is served before the scheduler schedules CBR session for service. Any other equivalent function which has the same characteristic can also be chosen as $G[\mu, I(t)]$. When both a and b are equal to 1, then this function can also be interpreted as the system utilization function of the typical $M/M/1$ queue as given in [14]. For $G[\cdot]$ to describe the real system, it must be positive and its values must belong to $(0, 1)$ for $x \in (0, \infty)$. Further we also assume that the function $G[\cdot]$ is differentiable twice. This function $G[\mu, I(t)]$ ensures that there exists an optimal switching time before T. The above function not only looks for the amount of interfering traffic sent out but it also takes the server rate into consideration in determining the switching time. Let $u(t)$ denote the fraction of time the server is allocated by the scheduler to CBR traffic for service at time t. This $u(t)$ must satisfy the constraint

$$0 \leq u(t) \leq 1, \ \forall t \tag{2}$$

The net revenue earned by the system in serving the interfering traffic sent out increases as the scheduler allocates it for service. So the amount of the interfering traffic sent out increases and is given by

$$\frac{dI(t)}{dt} = [1 - u(t)]G[I(t), \mu] \tag{3}$$

starting from $I(0) = I_0$. The expected total CBR fluid sent out in time $[0, T]$, is given by

$$\phi = E\left[\int_0^T u(t)G(\mu, I(t))dt\right] \tag{4}$$

where $E[\cdot]$ indicates the average with respect to the stochasticity caused by the server. Eqn. (4) is equivalent to average amount of CBR fluid that has been sent out successfully. This also indicates the amount of CBR fluid sent out following the schedule $u(t)$. If the source rate is constant or Markov modulated or follows a diffusion process over time, the optimal schedule in determining the switching time t_s is arrived by solving partial differential equations obtained by stochastic dynamic programming. The source rate μ is determined by the model adopted *viz.*

(1)constant source rate,

(2)Markov modulated source rate or
(3)when the source rate is modelled according to a diffusion process.
The optimal schedule has a clear switching from serving interfering traffic to CBR traffic and is given by

$$u^*(t) = \begin{cases} 0 & if \quad t < t_s \quad (5a) \\ 1 & if \quad t > t_s \quad (5b) \end{cases}$$

where the switching time t_s is determined by

$$T - t_s = \frac{[1 + bI(t)]^2}{a\mu} \tag{6}$$

as a function of Interfering traffic, $I(t)$, sent out by the source and the source rate being μ. Eqn.(6) is the decision rule of the scheduler. According to Eqn.(6) the scheduler continues to allocate interfering traffic for service as long as server rate is higher. Even if there is a continuous arrival of interfering traffic the scheduler schedules CBR session for service after t_s thus removing that ambiguity of offering preemptive service to the interfering traffic. Here μ_1 refers to high level of source rate and μ_2 refers to low level of source rate. If the source rate is very low then t_s is very low indicating that the scheduler will try to schedule the CBR traffic most of the time. The scheduler schedules the CBR session based on current source rate. The scheduler thus tracks the local environment instead of using a fixed service schedule to switch from interfering traffic to CBR traffic. The scheduler must estimate the future source rate to optimally allocate service for both the sessions. If the source rate varies continuously then taking the beheaviour adapted to current rate might be problematical.

3.1 Derivation of the function $G(I(t), \mu)$

As said in [14], we are interested in the time dependent behaviour of the simple M/M/1 queue. Consider a queue governed by the following equation.

$$Q(t) = A(t) - D(t) + Q(0) \tag{7}$$

$A(t)$, $D(t)$ and $Q(t)$ are the cumulative number of arrivals, cumulative number of departures and the number of entities in the system. $Q(0)$ is the number of entities at the beginning of the system. As we are interested primarily in the average quantities, we assume that the time behaviors of $A(t)$, $D(t)$ and $Q(t)$ have been measured on several occasions, e.g., during successive days of normal use of the system, and that the averages $\hat{A}(t)$, $\hat{D}(t)$ and $\hat{Q}(t)$ are thus known. We assume further that the averages $\hat{A}(t)$, $\hat{D}(t)$ and $\hat{Q}(t)$ are continuous functions of time, differentiable piecewise in $(0, T)$. Then it follows from Eqn.(7)

$$\frac{d\hat{Q}(t)}{dt} = \frac{d\hat{A}(t)}{dt} - \frac{d\hat{D}(t)}{dt} \tag{8}$$

with the initial condition $Q(0) = Q_0$. To simplify the notation, let $\lambda(t)$ denote the average number of arrivals in the system and is equal to $\frac{d\hat{A}(t)}{dt}$. Similarly $x(t)$ denote the number of entities in the system and is equal to $x(t) = \hat{Q}(t)$. It is assumed that smooth functions are assumed for averages.

If the system is not empty we approximate the $\frac{d\hat{D}(t)}{dt}$ by the function $\mu G[x(t)]$, where μ is the server capacity. So Eqn.(8) becomes

$$\frac{d\hat{x}(t)}{dt} = \lambda(t) - \mu G[x(t)] \tag{9}$$

Let us now consider the normalized form of Eqn(9), which is obtained by assuming the server rate, ($\mu = 1$), to be 1 and then using time transformation. So Eqn.(9) becomes

$$\dot{x}(t) = -\rho(t) + G[x(t)] \tag{10}$$

where $\rho(t) = \lambda(t/\mu)/\mu$. For a constant intensity input, ρ_0, the system will reach the steady state and so Eqn.(9) becomes

$$0 = -G[x(t)] + \rho_0 \tag{11}$$

So for the stationary state we have the queue length given by $\hat{x} = \frac{\rho}{1-\rho}$, from which we have $G[x(t)] = \frac{x(t)}{1+x(t)}$. Now to make the function more general we add the constants a and b at the numerator and denominator respectively. Now instead of the queue length, $x(t)$, we assume the total number of Interfering fluid that has been sent out till time t.

4 Example

Substituting the leaky bucket equation for $I(t)$ in Eqn.(6), we get

$$T - t_s = (1 + \sigma + \rho t_s)^2 \tag{12}$$

The above equation reduces to a quadratic equation given by

$$\rho^2 t_s^2 + [2\rho(1+\sigma) - 1]t_s + (1+\sigma)^2 - T = 0 \tag{13}$$

Consider a example in which the interfering traffic is modelled by the (σ, ρ) constraint. Assuming the server rate to be 2.5Gb/s, we calcu lated the switching time t_s and the values for various values of (σ, ρ) is given below.

μ = 2.5 Gbps, c = 64 Kbps, ρ = 3.00 Gbps, T(deadline) = 6ms	
σ	t_s in milliseconds
1 Gb	0.374575
50 Mb	1.39045
10 Mb	1.42981
5 Mb	1.43472
1 Mb	1.43866
μ = 2.5 Gbps, c = 64 Kbps, ρ = 2.50 Gbps, $T(deadline) = 6ms$	
σ	t_s in milliseconds
1 Gb	0.56155281
50 Mb	1.73035
10 Mb	1.7791
5 Mb	1.7852
1 Mb	1.79007
μ = 2.5 Gbps, c = 64 Kbps, ρ = 2.00 Gbps, $T(deadline) = 6ms$	
σ	t_s in milliseconds
1 Gb	0.746833
50 Mb	2.28599
10 Mb	2.34378
5 Mb	2.35739
1 Mb	2.62095

μ = 2.5 Gbps, c = 64 Kbps, ρ = 1:25 Gbps, $T(deadline) = 6ms$	
σ	t_s in milliseconds
1 Gb	1.4641
50 Mb	4.32719
10 Mb	4.44318
5 Mb	4.45766
1 Mb	4.46924

The above example clearly indicates that there exists a optimal switching time $t_s(< T)$ and our algorithm ensures that all the CBR traffic are transmitted before T.

References

[1] R.L.Cruz, Calculus for network delay-Part I:Network elements in isolation, *IEEE Trans. Inform. Theory*, 37, (1), (1991) 114-131.

[2] J.S.Turner, New directions in communications (or, which way to the information age?) *IEEE Commun. Mag.*, Oct. (1986).

[3] A.Parekh and R.Gallager, A generalised processor sharing approach to flow control in integrated service networks: The single- node case, *IEEE/ACM Trans. Networking*, 1, (3), (1993), 344-357.

[4] D.C.Lee, Effects of leaky bucket parameters on the average queueing delay: Worst case analysis, in *Proc. IEEE INFOCOM 1994*, 482-489, 1992.

[5] D.C.Lee, Worst-case fraction of CBR teletraffic unpunctual due to statistical multiplexing, *IEEE/ACM Trans. Networking*, 4, (1), (1996), 98-105.

[6] T.M.Chen, J.Walrand and D.G.Messerschmitt, Dynamic priority protocols for packet voice, *IEEE J. Select. Areas. Commun.*, 7, (5), (1989).

[7] H.Goldberg, Analysis of the earliest due date scheduling rule in queueing systems, *Math. Oper. Res.*, 2, (1977), 143-154.

[8] Y.Lim and J.Kobza, Analysis of a delay-dependant priority discipline in a multiclass traffic packet switching node, *Proc. IEEE INFOCOM*, (1988), 9A.4.1-9A.4.1.10.

[9] L.Kleinrock, *Queueing System, Vol II: Computer Application*, (Wiley Interscience, New York, 1976).

[10] L.Kleinrock, *Queueing System, Vol I: Theory*, (Wiley Interscience, New York, 1975).

[11] L.Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, (McGraw -Hill, New York, 1964).

[12] L.Kleinrock, A Delay Dependant Queue Discipline, *Naval Res. Logistics Q.*, II (1964) 329-341.

[13] L.Kleinrock and R.P.Finklestien, Time Dependent Priority Queues, *Operations Res.*, 15 (1976) 104-116.

[14] J.Filipiak, *Modelling and Control of Dynamic Flows in Communication Networks* (Springer-Verlag, 1988).

[15] S.Karlin and H.D.Taylor, *A Second Course in Stochastic Process* (Academic Press, New York).

[16] W.Feller, *Introduction to Probability Theory and Its Applications, Volume II* (Wiley, New York, 1966).

[17] M.Sniedovich, *Dynamic Programming* (MARCEL DEKKER, INC., New York, 1992).

Session 7-B

IP over ATM 4 (Performance)