

Optimal Bandwidth Allocations for VoIP Latency Guarantees

REU-CHING CHEN, WEI-TSONG LEE* AND JIM-MIN LIN

Department of Information Engineering

Feng Chia University

Taichung, 407 Taiwan

E-mail: che1627@ms18.hinet.net

**Department of Electrical Engineering*

Tamkang University

Taipei, 251 Taiwan

E-mail: wtlee@fcu.edu.tw

In recent years, a tremendous number of multimedia applications have come into widespread use in Internet environments. In multimedia communications, data, video and voice are transmitted through high speed networks. Their common transmission characteristics must satisfy the requirements of quality of service (QoS). For instance, low delay, bounded blocking rate and high throughput are fundamental factors influencing QoS. There exist many scheduling disciplines designed to guarantee QoS. In QoS guarantees, delay and loss are the main factors determining system performance. In voice transmission over the Internet (VoIP), latency is the most sensitive issue. In this paper, we focus on optimal system delay calculations based on proper bandwidth allocations. We address the problems related to the total system delay required for optimization and present a blocking probability estimation based on the large deviation principle (by LDP). More bandwidth can decrease the value of the delay, but more system resources will be exhausted. Less bandwidth may cause traffic to become congested. Schemes improve the performance when congestion occurs is not intelligent. Therefore searching for an efficient scheme to achieve optimal bandwidth allocation based on the minimum delay before congestion occurs is an important research goal. We use the Lagrange multiplier methods to achieve the optimal delay control in this paper. The system resources are dynamically adjusted according to different bandwidth requirements. Our contribution to the optimal delay is bounded as a function of the ratio of the input stream rate to the service rate, and a quick calculation for estimating the blocking probability accompanied by effective bandwidth calculations under some constraints is also provided. It is proved that a unique optimal solution exists. The technology presented here can be used by resource managers to design gateways for Internet environments.

Keywords: effective bandwidth, weight-fair queue (WFQ), large deviation principle (LDP), Markov-chain, first-in-first-out (FIFO)

1. INTRODUCTION

Real time communication for multimedia applications used on the Internet is expected to become a necessary feature in the future because multimedia, such as audio and video require much tighter performance control. Performance guarantees, along with other service guarantees, will be important for defining various grades of service. We

Received December 9, 2003; accepted February 24, 2004.
Communicated by Chu-Sing Yang.

believe that the domains of application of these services will grow rapidly. In multimedia environments, streams are transmitted from a server to a receiver. The streaming path consists of the components of servers (i.e., routers), the network and the receivers. The continuous media stream is retrieved from the server's storage unit for transmission. The media stream is then packetized and sent out through the network, arriving at the receiver's network. At the receiver end, the stream is decoded and played back. After the media is picked up from the receiver's streaming buffer, it is discarded to free up buffer space for new incoming media streams.

Multimedia communication has rather strict delay requirements, especially in real time environments. Real time service can create real time channels on demand and guarantees the satisfactory performance. These guarantees often offer sufficient bandwidth, and the delay is bounded within an acceptable range. Too more bandwidth will waste system resources, while less bandwidth will cause service to be degraded. Especially in high speed networks, adequate bandwidth allocation is an extremely challenging and important problem (i.e., in asynchronous transfer mode (ATM) networks). One interesting problem is the development of searching schemes to support QoS for a variety of applications with diverse traffic characteristics. The QoS guarantees could be in the form of a delay bound, loss probability etc. QoS guarantees applied in high speed communication networks can be divided into two types: guaranteed-service and best-effort service guarantees. Guaranteed-service guarantees provide a fixed quality of service to traffics. In contrast, best-effort guarantees do not guarantee a fixed level of service. For instance, on the Internet, there are no service guarantees, but ATM networks offer service guarantees to incoming traffic (i.e., a constant bit rate (CBR) and a variable bit rate (VBR) used in the ATM backbone).

The delay-sensitive characteristic of voice transmission is a main factor influencing the system's QoS. In this paper, we concentrate on the performance of delay guarantees and optimal delay policy for packet voice transmissions. In VoIP environments, voice is transmitted by means of packet switching; if enough bandwidth is allocated, voice at the source end can be transmitted to the output port. Adequate bandwidth allocation is necessary to obtain the best delay, and the resulting end-to-end delay between the source and destination can be kept within a tolerable range. If the latency duration is greater than the threshold value, the voice transmitted to the destination side will not be correctly translated, and the QoS will not be guaranteed.

At here, we provide different bandwidth for different traffic rates. How to divide the bandwidth properly and at the same time manage resources efficiently is a critical problem. Usually, dynamic scheduling is the most efficient way to solve this problem. We propose an efficient dynamic bandwidth allocation scheme for delay optimization. A continuous Markov Chain with four states is adopted in the performance analysis.

To achieve efficient utilization of system resources, the network attempts to transfer packets with maximum throughput based on some performance constraints (here, a proper bandwidth allocation is performed to achieve the optimal delay). Large throughput induces the network to be congested. Congestion occurs when traffic transmitted without control (i.e., core router). Various control schemes have been proposed to prevent congestion from occurring. For instance, two different congestion control schemes are adopted in TCP and ATM networks. In the TCP context, the source increases its rate slowly, and when congestion is detected, the destination informs the source of packet overflow, then the source decreases its rate to avoid congestions. In

overflow, then the source decreases its rate to avoid congestions. In ATM congestion control, the network allows intermediate nodes to inform the source if the input rate is overflow at that node. In this scheme, each intermediate node is capable of restricting a data rate that causes overflow. Alternatively, the control problem can be solved using feedback control, where queue length is used as the control factor for input streams. From control theory, the system performance can be stabilized by the position-integral-differential (PID) controller [5] to achieve steady state response. The control scheme is beyond the scope of this paper. Here, we adopt queuing theory in our analysis.

Since more and more voice traffic is being diverted from public switching telephone networks (PSTN) to the Internet (this phenomenon is expected to continue in the future), VoIP has become an important application in real time multimedia systems. For instance, IP telephony may be built on top of TCP or UDP, depending on whether they are loss-sensitive or time-sensitive, respectively. The factors that influence voice QoS are end-to-end speech quality and end-to-end delay, in which speech quality is affected by encoding and decoding operations. Voice transmission is the main issue in VoIP applications. For VoIP applications, at least the following components should be provided: First is an encoder, which samples the voice signal and creates the packet stream for transmissions. The encoder G.711 generates an 8-bit sample for one channel, and a traditional 64K-bit per second stream is created. Other sampling technologies are G.726 (whose rate varies from 16K to 40K bits per second) and G.723.1 (whose rate varies between 5.3K and 6.4K). Following the sampling stage, a packetizer is used to encapsulate the speech samples (including talkspurt and silence) into packets for transmissions. The corresponding components, including a playback buffer at the receiving end, a de-packetizer and a decoder are provided at the receiving end for playback of the original signal. Each of the above components will induce delay and loss. A high overall delay may impede interactive communication, such as a phone call. To overcome this problem, many schemes have been developed to share system resources efficiently for voice calls [6, 7]. Since the quality of VoIP does not match the quality of circuit switching, many activities are implemented in order to guarantee high quality voice service. In QoS control, the best-effort guarantee is the basic service policy on Internet, which serves users well enough but is not adequate for real time environments. It is imperative for an implementation of VoIP to remain cognizant of quality. Basically, QoS features provide predictable service by means of (1) adequate bandwidth allocation, (2) priority scheduling, (3) congestion control, (4) loss bounding, and (5) delay tolerance. We will focus here on adequate bandwidth allocation for optimal latency guarantees. Analytical and numerical models for studying voice performance will be developed. The system model used here is approximated by means of Continuous Time Markov Chains (CTMC) for the purpose of analysis. When we analyze system performance, the more states we consider, the better the accuracy we obtain, but the more complex is in the analysis. Voice models of two states have been adopted in analysis [13]. A Detailed voice model involving four or even more states can be found in [14]. We adapt a four-state Markov Chain for system performance analysis. As for bandwidth divisions, there are several algorithms for bandwidth allocations in the literature. They include the packet-by-packet version of generalized processor sharing (PGPS) [18], the virtual clock, self-clocked fair queuing (SFQ) [19], and delay-earliest-due-date (delay EDD) [20]. These algorithms are all capable of providing deterministic upper bounds on the end to end delay by leaky-

bucket. Our policy is defined as being work-conserving, which implies that if one or more of the queues do not use the fraction of the capacity allocated to them, the excess amount is distributed among the remaining queues. Our results presented here can be extended to end to end delay guarantees for specific purposes.

The rest of this paper is organized as follows. In section 2, we describe the system model. In section 3, an optimal latency using Lagrange methods is calculated, and section 4 presents numerical results and a discussion. Conclusions are drawn in section 5.

2. SYSTEM MODEL DESCRIPTIONS

In general, a continuous sequence of voice packets from a talker (coded by a vocoder) includes two parts: one is a talkspurt period and the other part is a listener period. When pulse code modulation (PCM) techniques are used, the talking periods correspond to more packets and the silent periods correspond to small packets during a conversation cycle. To simplify the system model, we adopt the symmetric speaker-listener pair in our analysis. Generally, four states exist in voice conversations, including two talkspurts (a long talkspurt and a short talkspurt) and two silences (a long silence and a short silence) as shown in Figs. 1 (a), (b), and (c). Within a conversation cycle, a long talkspurt will follow a short silence alternatively in a talking period. Similarly, each long silence will follow a short talkspurt during the listener interval. The model proposed by Peter O'Reilly [2] for describing voice activities includes three states of a Markov Chain, in which two kinds of silent states accompanied by one long talkspurt is introduced for voice analysis. We expand the voice process to four states for detailed analysis. Four states denoted as LT, SL, LS, and ST are presented. These four states correspond to four voice types and each type of voice signal is sampled and packetized before being transmitted. Fig. 1 (c) shows that LT will generate the longest packets and SS the shortest packets for transmissions. This indicates LT will require more bandwidth than the other three types. Therefore, each state will require a bandwidth for conversations; different bandwidths will be provided to meet different requirements. We add a short talkspurt for completeness of describing voice state transactions, and a more accurate estimation of system performance can be obtained.

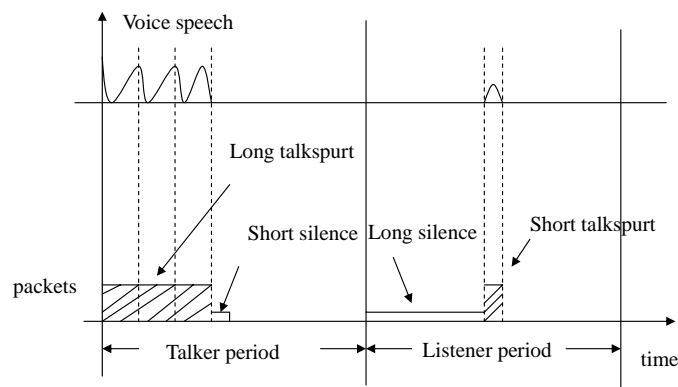


Fig. 1 (a). Four states for a talkspurt and silence pair.

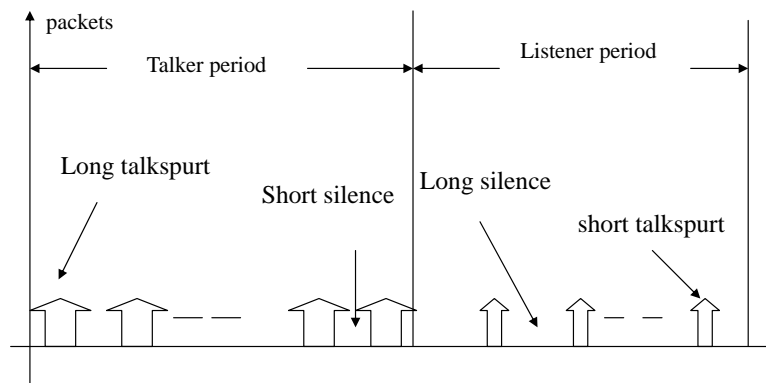


Fig. 1 (b). Packet sizes corresponding to four states.

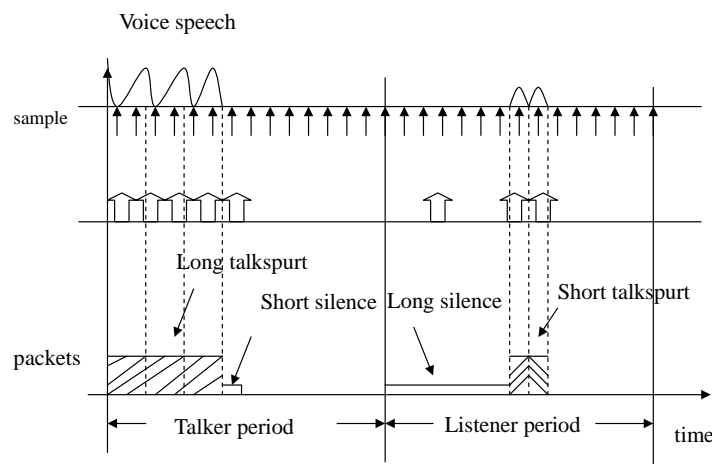


Fig. 1 (c). Packet size comparison for four states.

In Fig. 2, a four-state transition diagram is depicted for one phone call. Here states LT, LS, ST, and SS correspond to multiple input streams consisting of a long talkspurt, long silence, short talkspurt, and short silence, respectively. LT and SS constitute one pair of states, and LS and ST constitute another pair. We assume that the possible transitions between the two pairs is from LT to LS or conversely. States transitions between ST and SS are possible; nevertheless, these probabilities are very small and can be neglected. Indexes from 1 to 4 are used to indicate the four states, LT, LS, ST, and SS for the sake of convenience.

In spite of using MMPP (Markov Modulated Poisson Process) for performance analysis, we adopt a birth-death Markov chain as shown in Fig. 2. The states in Fig. 2 correspond to those in Fig. 1 (c), where the service distribution for each state is assumed to be exponential for convenience, and where μ_1 , μ_2 , μ_3 , and μ_4 indicate the service rate of state LT, SS, ST, and LS respectively. The operation of state transitions in a conversation cycle can be described as follows: Assume person A is talking with person

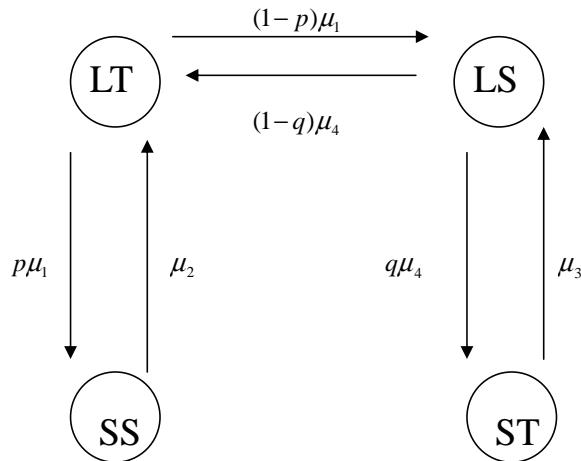


Fig. 2. Four transition states of a Markov chain.

B ; then, A starts talking in the long talkspurt state. At the same time, B is in the long silence state. Let p and q indicate the probability that transfer from a long talkspurt to a short silence state and from a long silence to a short talkspurt state separately. Then, $1 - p$ is the probability that A will transfer from a long talkspurt state to a long silence state, and $1 - q$ is the probability that A will transfer from a long silence to a short talkspurt state. When A is in a short silence state, it can only go back to a long talkspurt state (the short talking period followed by a short silence period is possible, nevertheless this probability is assumed to be zero in this paper), similar processes are operated for short talkspurt period.

3. SYSTEM ANALYSIS

3.1 Markov Model Descriptions

Based on the state transitions described in the last section, let π_1 , π_2 , π_3 , and π_4 indicate the stationary states corresponding to LT, SS, ST, and LS respectively. Then the steady state solution can be easily obtained based on the assumption that the system is ergodic. From Markov property, the metric of balance equations [1] are as follows:

$$\begin{bmatrix} \mu_1 & -\mu_2 & 0 & -(1-q)\mu_4 \\ -p\mu_1 & \mu_2 & 0 & 0 \\ 0 & 0 & \mu_3 & -q\mu_4 \\ -(1-p)\mu_1 & 0 & -\mu_3 & \mu_4 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{bmatrix} = 0 \quad (1)$$

with constraint

$$\sum_{i=1}^4 \pi_i = 1, \quad (2)$$

where π_i is the stationary probability of state i .

From (1) and (2), the solution can be expressed in vector form as follows:

$$\phi = eA,$$

$$\text{where } \phi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{bmatrix}, e = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, A = \begin{bmatrix} 1 + \mu_1 & 1 - \mu_2 & 1 & 1 - (1 - q)\mu_4 \\ 1 - p\mu_1 & 1 + \mu_2 & 1 & 1 \\ 1 & 1 & 1 + \mu_3 & 1 - q\mu_4 \\ 1 - (1 - p)\mu_1 & 1 & 1 - \mu_3 & 1 + \mu_4 \end{bmatrix}^{-1}.$$

Since each state indicates one type of input stream with some bandwidth requirements, then there are four types of arrivals (two types constitute one pair) in our model. Each type corresponds to a different bandwidth requirement. Each transition state occupies some bandwidth resources provided by the system. We adopt a simple M/M/1 conservative queuing system with FCFS discipline in our analysis. From the above, let N be the traffic source of input queue that fits the constraint $\sum_{i=1}^N x_i \leq \mu$, where μ is the total system bandwidth capacity and x_i is the bandwidth provided for input i by the system. Since we are considering a conservative system, the single server (i.e., the router in one node of the Internet network) will not be idle when there are packets waiting for transmission.

3.2 Optimal System Delay Calculations

We analyze the system performance from the global point of view. As Fig. 3 (a) shows, there are four types of input traffic requiring bandwidth, i.e., LT, SS, LS, and ST. It should be noted that each type of traffic may be extended to contain different packet lengths. For instance, in Fig. 3 (b), LT-type traffic may include long, medium, and short packets of LT, and the same is true for the other three traffic types, SS, LS, and ST. Long packets require more bandwidth; therefore N equals twelve in Fig. 3 (c). Multiple input streams arriving randomly in the common queue are treated as individual virtual queues served based on the FIFO policy. We will focus here on the appropriate bandwidth allocations, based on the minimum total system delay. For each time slot, the input streams selected for optimal bandwidth allocations are forwarded to the networks at the transmission rate provided by the optimal scheme. Packets are then served according to the first-in-first-out (FIFO scheme) non-preemptive discipline. For the purpose of introducing the optimal bandwidth allocation, the input traffic types are labeled with 1 to N sessions as Fig. 3 (c) depicts for the optimal bandwidth allocation. Let λ_i be the arrival rate of session i , its value equals $\alpha_i \lambda$ (input stream i type and input session i will be used alternatively in the following sections), let μ be the total system service rate whose value is taken to be unity for the sake of simplicity, and let α_i and K_i be the non-negative fractions of the arrival rate and service rate for the session i stream, where $0 \leq \alpha_i \leq 1$ and $0 \leq K_i \leq 1$. In steady state, we can allocate N different input streams with the minimum total system delay.

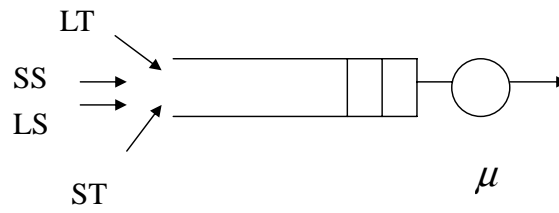


Fig. 3 (a). Aggregated of input traffics.

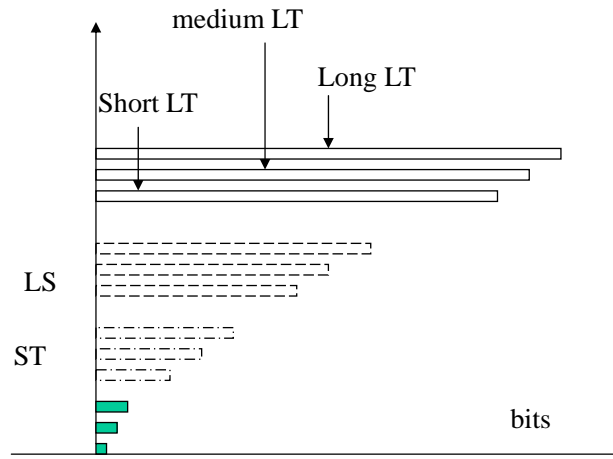


Fig. 3 (b). Bit rates of different states.

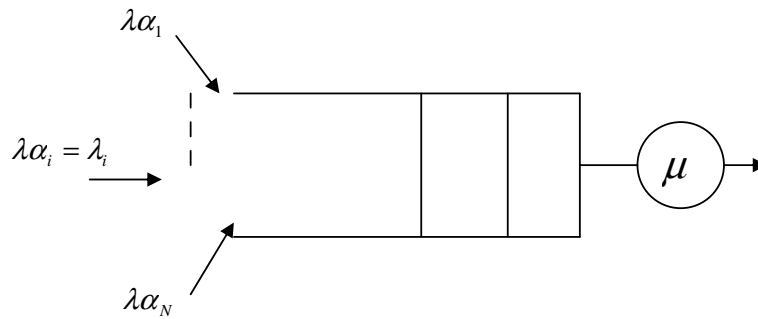


Fig. 3 (c). Equivalent queuing model.

Based on the additive property of the Markov process, the N input identical independent distribution (i.i.d.) stream can be thought of as being equivalent to one virtual infinite buffer for each queue. Under steady state and assuming that each input stream is the Poisson process, we have the following lemma for the optimal total system delay with the multiple input M/M/1 queuing model:

Lemma 1 For an N virtual queue M/M/1 type single server system with a unity service rate, the optimal system total delay is characterized by the constraint that the individual arrival rate of each input traffic stream ($\rho\alpha_i$ in Eq. (4)) is less than the ratio of the allocated bandwidth (K_i in Eq. (4) below) and their ratio is a constant.

In the proof, the following notations for parameters are adopted for the sake of convenience:

- K_i : the partial fraction of service rate provided to stream i ;
- α_i : the partial fraction of arrival rate for stream i ;
- ρ : the system utility, its value is less than 1;
- N : the total number of input streams;
- T_i : the delay for virtual queue i ;
- μ : the total system service rate.

Proof: The average system delay for input stream λ_i of the M/M/1 queuing model is [1]:

$$T_i = \frac{\left(\frac{1}{K_i\mu}\right)}{\left(1 - \frac{\alpha_i\lambda}{K_i\mu}\right)}. \quad (3)$$

Then the total system delay is

$$\sum_{i=1}^N T_i = \frac{1}{\mu} \sum_{i=1}^N \left[\frac{1}{K_i - \rho\alpha_i} \right]. \quad (4)$$

Based on the Lagrange multiplier methods, let

$$f(\alpha_1, \dots, \alpha_N) = \sum \frac{1}{K_i - \rho\alpha_i}$$

with the auxiliary equation $g(\alpha_1, \dots, \alpha_N) = \alpha_1 + \dots + \alpha_N - 1$. (5)

Set $F(\alpha_1, \dots, \alpha_N) = f(\alpha_1, \dots, \alpha_N) + rg(\alpha_1, \dots, \alpha_N)$

For optimal solution, we have

$\frac{\partial F}{\partial \alpha_i} = 0$, where i is a positive integer with $0 < i < N + 1$. This implies that

$$\frac{\rho}{(K_i - \alpha_i\rho)^2} = -r = \text{constant and that}$$

$$(K_i - K_j) = (\alpha_i - \alpha_j)\rho \quad (6)$$

conjunction of Eq. (6) with constraints $\sum_{i=1}^N K_i = 1$, $\sum_{j=1}^N \alpha_j = 1$ we have

$$K_i - \rho\alpha_i = \frac{1-\lambda}{N} \quad (7)$$

in which $1 \leq i \leq N$. □

3.3 Blocking Estimation Performed using Methods Based on the Large Deviation Principle

In applications, Large Deviation methods can be used to estimate the speed of convergence of the system delay. Based on the previous assumption of an i.i.d. Poisson process to each input flow of the stream is assigned a fraction of the total bandwidth with respect to the optimal system delay. Then, system bandwidth resources are properly allocated to each input to achieve the optimal delay guarantee under the condition that the buffer size is large enough. Let B be the system buffer size; the invariant probability that the occupancy rate of the buffer exceeds B is [17]

$$P(W > B) \approx \exp(-\Delta_i B), \quad (8)$$

in which Δ_i is a function of the input traffic type with a specific arrival rate and W is the buffer size. Let r be the empirical arrival rate and let c be the service rate; then, the value of c can be obtained for a specified decay rate Δ_i . Since that the traffic is stationary and ergodic; then, the number of arrivals in the time interval $[0, t]$ is $A_i(t)$ for a specific i -type traffic. Under the assumption that the Gartner-Ellis [17] theorem is satisfied, the log moment generating function of i type in one queue is

$$s_i(\Delta_i) := \lim_{t \rightarrow \infty} \frac{1}{t} \log E(e^{A_i(t)\Delta_i}).$$

From LDP (the large deviation principle) [15], we have

$$\Delta_i = \inf_{r > c} \frac{S_i(r)}{r - c}, \text{ where } S_i(a) \text{ is the dual Legendre transform of } s_i,$$

$$S_i(r) := \sup_{\theta \in R} [\theta r - s_i(\theta)] \text{ where } R \text{ is a positive real number.}$$

Assume that the Poisson arrival rate is λ_i ; then, based on the above condition, the effective bandwidth for i type traffic in one queue can be expressed as

$$K_i(\Delta_i) = \frac{s_i(\Delta_i)}{\Delta_i} = \frac{\lim_{t \rightarrow \infty} \frac{1}{t} \log [e^{\lambda_i t (e^{\Delta_i} - 1)}]}{\Delta_i} = \frac{\lambda_i (e^{\Delta_i} - 1)}{\Delta_i}, \quad (9)$$

where s_i and S_i are convex functions. From the important additive property of effective bandwidth, the total effective bandwidth of the single server system can be calculated as the sum of individual random variables; then, under the assumption of a unity service rate, the total effective bandwidth is $K_i = 1 = \sum_{i=1}^N K_i$. This means the effective bandwidth of N independent arrival sequences equals the sum of individual effective bandwidth under the requirement that each individual independent arrival is allocated with its effective bandwidth.

Let Eq. (9) to be the effective bandwidth for one type of each queue, then $\lambda_i(1 + \Delta_i + \frac{\Delta_i^2}{2} + \dots - 1) = K_i \Delta_i$, where K_i is the fractional value for optimal bandwidth allocation. For $\Delta_i \ll 1$, we obtain

$$\Delta_i = 2\left(\frac{K_i}{\lambda_i} - 1\right). \quad (10)$$

The blocking probability can be easily estimated using Eq. (10).

In a critical case, if the service rate is very close to the value of arrival rate for each type of input stream, a simple estimation of blocking probability delay can be obtained as follows:

Lemma 2 For N input sessions of an $M/M/1/B$ queue, where B is large enough, if each session is allocated with an effective bandwidth as its service rate, and if the service rate is greater than the arrival rate within a bounded range denoted by the equation $K_i = (1 + \epsilon_i)\lambda_i$, where K_i is the effective bandwidth allocated for session i and ϵ_i is a small positive real number less than 1, then

<1> the blocking probability of session i is approximately $\exp(-2B\epsilon_i)$.

<2> the optimal blocking probability is bounded.

Proof: <1> From Eq. (10) and the given constraint, we have

$$\Delta_i = 2\left(\frac{K_i}{\lambda_i} - 1\right) = 2\epsilon_i.$$

This completes the proof.

<2> From $\Delta_i = 2\left(\frac{K_i}{\lambda_i} - 1\right) = 2\frac{K_i - \lambda\alpha_i}{\lambda\alpha_i}$ and Eq. (6),

by setting $BT = \sum_{i=1}^N \frac{1}{\exp(B\Delta_i)}$ and taking Lagrange multiplier we obtain

$$2B \frac{\exp(B\Delta_i)}{\exp(2B\Delta_i)} \frac{K_i}{\lambda \alpha_i^2} = \text{constant, which implies that}$$

$$\exp(-B\Delta_i) \frac{K_i}{\alpha_i^2} = \text{constant.} \quad (11)$$

Eq. (11) is the approximation constraint for optimal blocking probability.

4. NUMERICAL RESULTS AND DISCUSSIONS

In Fig. 4, the system total delay, including three different arrival rates, 0.2, 0.3 and 0.5, is plotted, (other values can be selected with the constraint that the total arrival summation is 1), where three kinds of policies called the optimal, average and weight-fair queue (abbreviated as WFQ) are also plotted for the purpose of comparison. Since the service rate is normalized to be one unit, the scale of delay will also be scaled too, under real condition, total delay is scaled by different service rates. For instance, if the service rate is one hundred jobs per second, then the delay in Fig. 4 should be divided by 100. As for the optimal and weight-fair queue policies, different bandwidth requirements of customers are provided by the different fractions of the total system resources; a smaller fraction corresponds to less allocated bandwidth, while a large fraction corresponds to more allocated bandwidth. The average policy is achieved by providing equal bandwidth for different bandwidth requirements. In our continuous Markov chain model, the buffer size is assumed to be large enough for the optimal total system delay. It is obvious that optimal delay policy is always better than a non-optimal one. This means that our policy is better than any non-optimal scheduling as expected. As for the total system delay, the critical point for the average policy occurs when the denominator of Eq. (4) is zero. As Fig. 4 shows, the average policy is better than WFQ for a light load. The delay curve of

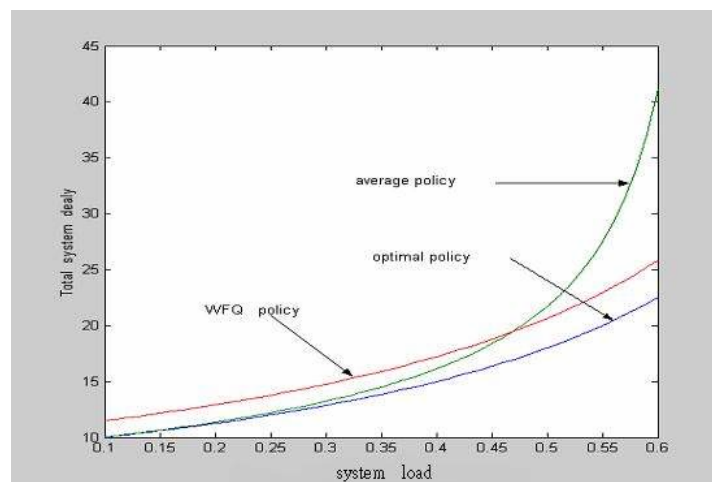


Fig. 4. Comparison of different policies.

the optimal policy is below the other two policies as expected. It can be easily proved that the optimal policy is equivalent to the weight-fair queue policy when the utility factor (the ratio of the arrival rate to the service rate) is equal to one. Under normalization of the service rate, the system delay is an increasing function of the arrival rate; a high arrival rate results in a long system delay. From Eq. (8), the blocking probability of each queue can be kept constant for various input rates by adequately regulating the queue buffer size. Fig. 5 (a) shows the blocking probability versus the system load with a buffer size of 50, Fig. 5 (b) plots a buffer size equal to 10. Fig. 5 (a) depicts the LDP approximate estimation. The blocking probability obtained using the WFQ policy will be better than that obtained using the optimal delay policy or the FIFO policy. The buffer size is assumed to be large enough for blocking estimation in this paper. If the buffer is not

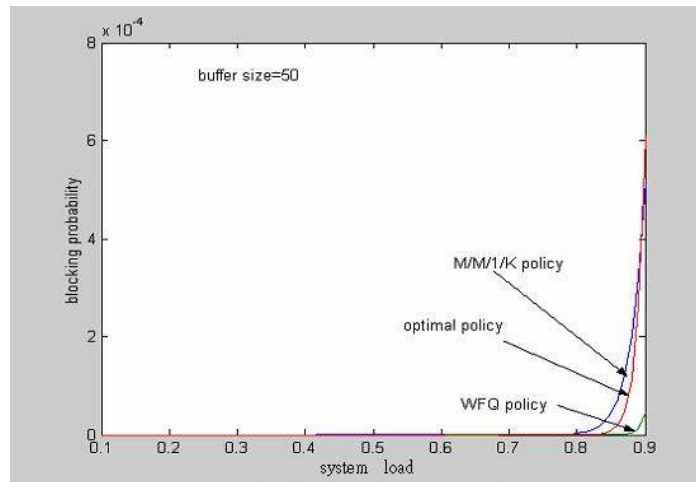


Fig. 5 (a). Buffer size = 50.

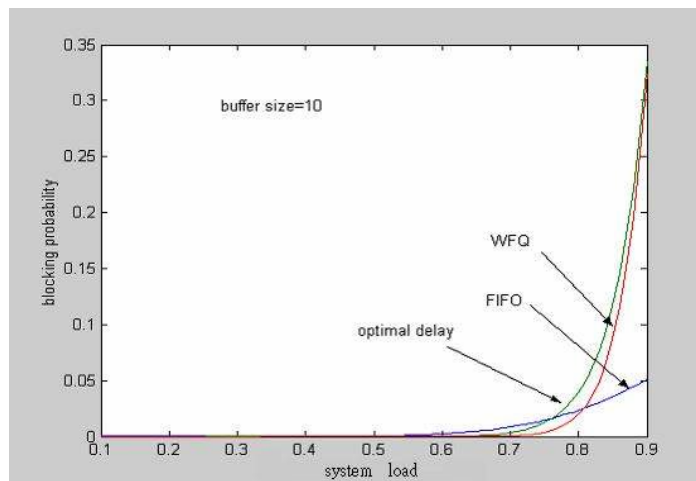


Fig. 5 (b). Buffer size = 10.

large enough, the estimation method will fail, i.e., the FIFO policy will be preferred to avoid blocking, as Fig. 5 (b) shows. The blocking probability in both Figs. 5 (a) and 5 (b) is an increasing function of the arrival rate, as expected. Lemma 2 indicates that the blocking probability is bounded by the value of ϵ_i if the allocated bandwidth is greater than the arrival rate within a specific small range of values. Reasonably, under a fixed load, the blocking probability is a monotonically decreasing function of the buffer size. This means that a stricter requirement of blocking probability will require more buffers since a large buffer size will benefit the performance of blocking probability. In Fig. 5 (b), the curve of the optimal delay policy is very close to that of the weight-fair queue policy and weight-fair queue is trivial in our optimal delay constraints. From Eq. (8), It is desirable that less effective bandwidth be required if more buffer capacity is provided under the constant blocking probability bounds.

5. CONCLUSIONS

In this paper, optimal bandwidth allocations for delay guarantees in VoIP environments have been investigated. The problem of dividing bandwidth so as to achieve the optimal system delay has been using the Lagrange Multiplier method. Our solution for VoIP bandwidth allocation is satisfied for the optimal delay, moreover, to calculate the optimal blocking probability, each arrival rate is assigned a value equal to the effective bandwidth. We have provided a simple estimation of the blocking probability obtained by using the large deviation principle. The blocking probability converges and is bounded by functions of the arrival rate and buffer size. Our policy for bandwidth allocations with different resource requirements of input traffic is robust and powerful. In Internet applications, the scheme depicted here can be very useful for controlling network node (i.e., router) so as to transmit information with delay guarantees. Especially, in connection-oriented architectures of real time environments, the advantage of our scheme is that it provides optimal latency guarantees and simple estimations of the blocking probability. The approach presented in this paper is simple and easy to implement. It can be applied to various environments in high speed real time networks. Future works will include estimating the confidence interval of limiting buffer size to satisfy both delay and blocking guarantees. The optimal policy for Non-Poisson traffic control will be our next interests.

REFERENCES

1. L. Kleinrock, *Queueing System Vol. 2: Computer Applications*, Wiley, New York, 1974.
2. P. O'Reilly and S. Ghani, "Data performance in burst switching when the voice silence periods have a hyperexponential distribution," *IEEE Transactions on Communications*, Vol. C-35, 1987, pp. 537-542.
3. H. Zhang and D. Ferrari, "Rate-controlled static priority queuing," in *Proceedings of IEEE INFOCOM '93*, 1993, pp. 227-236.
4. L. Zhang, "Virtual clock: a new traffic control algorithm for packet switching networks," in *Proceedings of ACM SIGCOM '90*, 1990, pp. 19-29.
5. K. Zhou, J. C. Doyle, and K. Glover, *Robust Optimal Control*, Prentice Hall, New

- Jersey, 1996.
6. H. Zhang, "Providing end-to-end performance guarantees using non-work-conserving disciplines," *Computer Communications: Special Issue on System Support for Multimedia Computing*, Vol. 18, 1995, pp. 769-781,
 7. R. L. Cruz and H. Liu, "End-to-end queuing delay in ATM networks," *High Speed Networks*, Vol. 3, 1994, pp. 413-427.
 8. J. G. Kim and M. Krunz, "Bandwidth allocation in wireless networks with guarantees packet-loss performance," *IEEE Transactions on Networking*, Vol. 8, 2000, pp. 337-349.
 9. C. Li, R. Bettati, and W. Zhao, "Response time analysis for distributed real-time systems with bursty job arrivals," in *Proceedings of IEEE International Conference on Parallel Processing*, 1998, pp. 432-440.
 10. E. W. Knightly, "Enforceable quality of service guarantees for bursty traffic system," in *Proceedings of the IEEE INFOCOM '97*, 1997, pp. 635-642.
 11. C. Z. Li, R. Bettati, and W. Zhao, "New delay analysis in high speed networks," Department of Computer Science, Texas A & M University, 1999.
 12. M. Vojnovic, J. Y. Le Boudec, and C. Boutremans. "Global fairness of additive-increase and multiplicative-decrease with heterogenous round-trip times," in *Proceedings of IEEE INFOCOM 2000*, 2000, pp. 1303-1312.
 13. D. Minoli, *Delivering Voice Over IP Networks*, 1998.
 14. M. May, T. Bonald, and J. C. Bolot, "Analytic evaluation of RED performance," in *Proceedings of INFOCOM 2000*, 2000, pp. 1415-1424.
 15. J. Gartner, "On large deviations for invariant measure," *Theory Probability, Applications*, Vol. 22, 1977, pp. 24-39.
 16. S. M. Ross, *Stochastic Process*, Wiley, 1996.
 17. J. Walrand and P. Varaiya, *High Performance Communication Networks*, Morgan Kaufmann, 2000.
 18. A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the multiple-node case," *IEEE/ACM Transactions on Networking*, Vol. 2, 1994, pp. 137-150.
 19. S. Golestani, "A self-clocked fair queuing scheme for broadband applications," in *Proceedings of IEEE INFOCOM '94*, 1994, pp. 636-646.
 20. V. Sivaraman and F. Chiussi, "Statistical analysis of delay bound violations at an earliest deadline first (EDF) scheduler," *Performance Evaluation*, Vol. 36, 1999, pp. 457-470.



Reu-Ching Chen (陳瑞卿) received the B.S.E.E. from Taiwan Technology University, R.O.C. in 1982 and the M.S. degree in Applied Mathematics from Tunghai University in 1996. He is currently pursuing for a Ph.D. degree in Information Science at Feng Chia University, Taichung, Taiwan, R.O.C. His research interests include stochastic process and traffic engineering.



Wei-Tsong Lee (李維聰) received the B.S.E.E., the M.S. and the Ph.D. degrees in Computer Science from National Cheng Kung University, Taiwan, R.O.C., in 1984, 1986 and 1995, respectively. He is currently the Associate Professor of Department of Electrical Engineering of Tungku University, Taiwan, R.O.C.. His current interests include high speed network, cable modem, and stochastic ordering.



Jim-Min Lin (林志敏) was born on March 5, 1963 in Taipei, Taiwan, R.O.C. He received the B.S. degree in Engineering Science and the M.S. and the Ph.D. degrees in Electrical Engineering, all from National Cheng Kung University, Tainan, Taiwan, R.O.C., in 1985, 1987, and 1992 respectively. Since February 1993, he has been an Associate Professor at the Department of Information Engineering, Feng Chia University, Taichung, Taiwan, R.O.C. His research interests include operating system, software integration/reuse, embedded systems, and software agent technology.