Delay Guarantee by Adaptive Weighted Fair Queue in Real Time VoIP Environments

Reu-Ching Chen and Chen-Sung Chang

Department of Information Management, Nan-Kai College

No. 568, Jhongjheng Rd., Caotun Township, Nantou, Country 542, Taiwan (R.O.C.)
E-mail: che1627@ms18.hinet.net

Abstract

This paper provides an optimal delay discipline for real time latency guarantees in VoIP environments. System resources are properly divided for delay optimization due to various traffic loads. We investigate the optimal delay guarantees of weighted fair queueing. In our approach, network resource is allocated according to the factor of weight-queue-length product. The numerical results demonstrate that the optimal delay can be achieved under the proposed bandwidth allocation scheme. The proposed scheme has the advantages both on delay optimization and easy implementation. The policy depicted here meets the requirements of minimum delay under various traffic types and this will be beneficial on solving the stringent problem of delay guarantees in voice communications encountered in modern Internet networks. Our contribution is focused on minimizing the system delay due to heterogeneous input traffics and the total system delay is significantly improved.

Key-words: Weighted Fair Queue, Adaptive, VoIP, Discipline, Delay Guarantee.

1 Introduction

Achieving an adequate resource division is essential in modern multimedia communication applications. (Note that the terms “resource” and “bandwidth” are used interchangeably in the remainder of this paper). Providing an application with either too much or too little bandwidth is clearly unsatisfactory, **t**he former is uneconomic from a cost point of view, where the later lead to application delays, reduced throughput, and loss rate issues from a system performance viewpoint.

In multimedia communications, different types of traffic require different quality of service (QoS). For instance, data traffic requires accurate transmissions, but can tolerate delays provided that they fall below a certain threshold. In another aspect, voice traffic is highly delay-sensitive, but can tolerate a limited loss rate provided that the specified QoS requirement is satisfied [2], [4]**.** QoS guarantees include various policies to satisfy the particular requirements of various classes of user. (Note that the terms “user” and “customer” are used interchangeably in the remainder of this paper).

Bandwidth allocation schemes generally assign users of a higher class a greater priority in order to prevent users of a lower class from seizing an unreasonable bandwidth when contention occurs. Contention schedulers generally favor higher priority

customers when the available system resources are insufficient to satisfy the combined requirements of all the network customers. In other words, priority-based allocation schemes tend to sacrifice users of lower priority in the event of network congestion.

Several algorithms are proposed using leaky-bucket mechanism for bandwidth allocations, e.g., packet-by-packet, delay-earliest-due-date are found in the literatures [6]. In [9] a novel scheme called Endpoint Admission Control (EAC) based on the analysis of the probing flow delay variation statistic is proposed to perform admission control. In [12], a proportional, integral and differential (PID) controller (using feedback control theory) is used to allocate bandwidth in a weighted manner, where a control packet is propagated through the ring node to reach the fairness criterion.

Unlike using the feedback control theory, in our study, an adaptive weighted fair queue method is proposed for bandwidth allocations in achieving delay optimization. In our approach, system bandwidth is properly allocated to fulfill the minimum delay requirements.

The system bandwidth is divided fairly among different classes of user in accordance with a weight-queue-length factor. The major contributions of the proposed approach include its ability to ensure that the total system delay is optimized.

The remainder of this paper is organized as follows: Section 2 describes the proposed system model, while Section 3 develops a mathematic analysis of the proposed approach based on queuing theory. Section 4 presents the numerical results and evaluates the performance of the proposed scheme. Finally, conclusion is provided in Section 5.

2 Model Description

Fig. 1 shows the speech powers for voice A and voice B respectively, also the generated packets corresponding to voice A and B are depicted. It is noted that voice A generates more packets than voice B since voice A has higher speech power than voice B, therefore, the former corresponds to higher traffic rate than the later. The generated packets from the voice sources are then fed forward to the input queue as shown in Fig.2, where higher amount of packets per unit time corresponds to higher traffic rate.

For the purpose of introducing the optimal bandwidth allocation, the input traffic types are labeled from 1 to *N* sessions as Fig. 2 shows. Where, λ_i, w_i and q_i indicate the arrival rate, weight factor and queue length for session i respectively. (Note that input stream i type and input session i are used interchangeably in the remainder of this paper). Without loss of generality, in this paper, the value of the total system service rate μ is assumed to be unity for the sake of simplicity. In steady state, we can allocate bandwidths for the *N* different input streams with the minimum total system delay.

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.

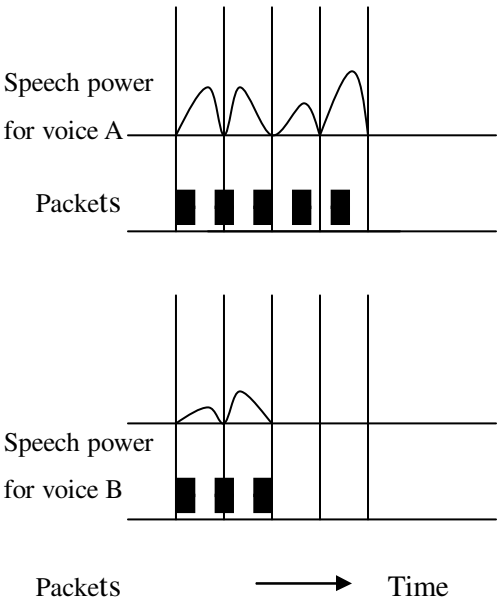

Fig. 1 Packet generating diagrams

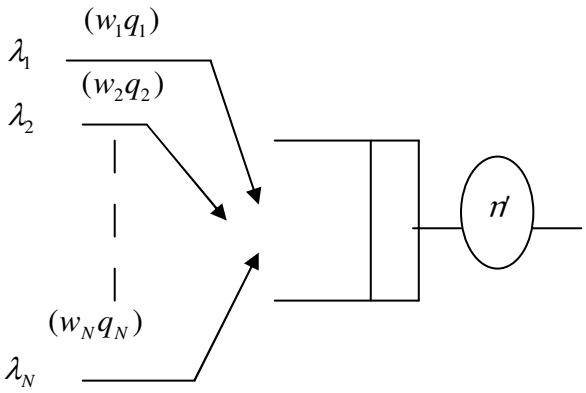

Fig. 2 Real Time Queueing Model

3 Mathematical Analysis

An N input streams with single node queueing model is adopted for analysis in this paper. In our queueing model, the service rate is variant for each input stream and its allocating rate is depending on the value of weight-queue-length product.

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. In this paper, the M/M/1 single server system is adopted for analysis. Without loss of generality, the system is assumed to be stationary and each input stream is the Poisson process.

Assume the traffic stream of arrival is Poisson and observation point is selected at the departure point (PASTA [1]), then the system can be thought of as an embedded Markov Chain. (state transition solution of Markov Chain can be found in [1] and is not presented here owing to space limitation). We prove a lemma concerning the optimal delay by weighted fair queue policy in the following.

Lemma : For a single server multiplexer with N queue sessions, let q_i , w_i and d_i denote the queue length, weight factor and delay respectively, then for a specific M/M/1 queue, the optimal total system delay exists under the constraints:

$$(\alpha_i - \alpha_j)\lambda = \frac{(q_i w_i - q_j w_j)}{M} \text{ -----(1)}$$

where $0 < i, j \le N$ and w_i is a positive real number, $M = \sum q_i w_i$,

Before proof, the following notations for parameters are adopted for convenience:

K_i: the partial fraction of service rate provided to stream i.

α_i : the partial fraction of input stream i.

ρ : the system utility, its value is less than 1.

N: the total number of input streams.

T_i : the delay for session i.

μ : the total system service rate.

proof>

The average system delay for input stream λ_i of

M/M/1 queueing model is [1]:

$$T_i = \frac{1}{(K_i\mu - \alpha_i\lambda)} \text{ ----------(2)}$$

then, the total system delay is

$$\sum_{i=1}^{N} T_i = \frac{1}{\mu}\sum_{i=1}^{N}\left[\frac{1}{K_i - \rho\alpha_i}\right] \text{ ----------(3)}$$

with the auxiliary equation $g(\alpha_1,...,\alpha_N) = \alpha_1 + ... + \alpha_N - 1$ ----(4)

let $f(\alpha_1,...\alpha_N) = \sum_{i=1}^{N}\frac{1}{K_i - \rho\alpha_i}$ and $F(\alpha_1,...,\alpha_N) = f(\alpha_1,...,\alpha_N) + rg(\alpha_1,...,\alpha_N)$.

Since total service rate is normalized to be unity, then, from Lagrange multiplier methods 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

$$\frac{\rho}{(K_i - \alpha_i\rho)^2} = r$$

=constant.----------(5)

by replacing $K_i = \frac{q_i w_i}{\sum_{j=1}^{N} q_j w_j}$ to equation (5)

we obtain

$$(\alpha_i - \alpha_j)\lambda = \frac{(q_i w_i - q_j w_j)}{M} \text{ --------(6)}$$

where M= $\sum_{j=1}^{N} q_j w_j$, its value can be set to be constant from the fact that service rate is divided according to fraction $K_i = \frac{q_i w_i}{\sum_{j=1}^{N} q_j w_j}$

From normalization characteristic, we have

$$\sum_{j=1}^{N}\alpha_j = 1 \text{ ----------(7)}$$

Conjunction equations (8) and (9), the weight for a specific queue j is given by

$$w_j = \frac{1}{q_j}[q_1 w_1 - M\lambda(\alpha_1 - \alpha_j)]$$

where 1 < j < N+1 and the initial given value is w_1, its value can be selected arbitrarily for convenience.

4. Numerical Results and Discussions

Fig. 3 depicts the comparisons of the total system delay for three different policies, i.e., average, proportional and optimal methods; where four different arrival rates 0.1, 0.2, 0.3 and 0.4 are taken as the input streams for the single server system. Hence in Fig. 3, the service rate equal to 0.25 for each input session in average policy; on the other hand, in proportional policy, service rates equal to 0.1, 0.2, 0.3 and 0.4 correspond to sessions 1, 2, 3 and 4 respectively. It is noted the optimal method proposed in this paper has the least value of delay than the other two methods. Hence, delay is an increasing function of traffic load as desired.

Fig. 4 depicts the relations between weight values and the arrival rates under the constraint of optimal delay, where w1, w2, w3 and w4 correspond to the arrival rates 0.1, 0.2, 0.3 and 0.4 respectively. It is noted, under the constraint of optimal delay guarantees, more weight value is added to the input stream with lower arrival rate. Hence, the value of the added weight for the input stream is a decreasing function of traffic load. This is reasonable since in our study, weight-queue-length is adopted as the factor for bandwidth dividing, therefore, high traffic load induces higher value of queue length, this will significantly decrease the weight required for optimal delay.

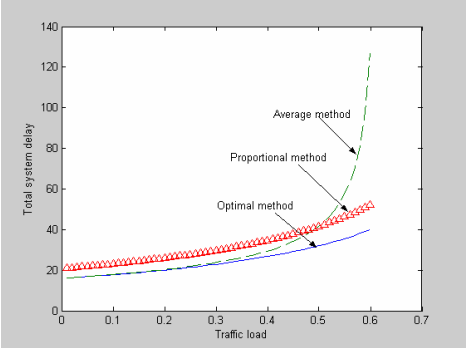

Fig. 3 Delay comparisons for different methods

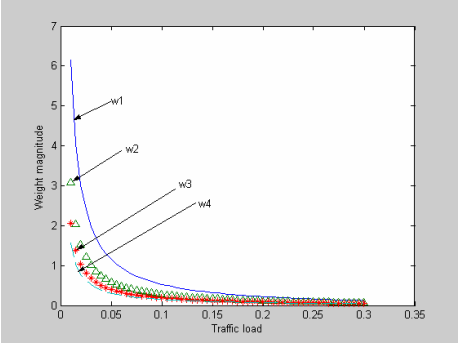

Fig. 4 Weight additions for different arrival rates with the constraint of optimal system delay.

5. Conclusions

In this paper, the optimal bandwidth allocations for delay guarantees in VoIP environments have been investigated. The problem of dividing bandwidth in achieving the optimal system delay using the Lagrange Multiplier method is provided. Our solution for VoIP bandwidth allocation is satisfied for the optimal delay.

We present a simple and effective scheme for real time bandwidth allocations in VoIP environments. The QoS requirements for delay sensitive networks are fulfilled by enhanced weight to each queue. 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 resource allocations with delay guarantees, especially, in connection-oriented architectures of real time environments.

The advantage of our scheme is focusing on providing optimal latency guarantees. The approach presented in this paper is simple and easy in implementations. It can be applied to various environments in high speed real time networks.

References

[1] Leonard Kleinrock, " Queueing system Volume I,II ," 1974.

[2] R. L. Cruz and H. Liu, "end-to-end queueing delay in ATM networks," High speed networks, vol.3, no4,1994.

[3] C.Li, R. Bettati and W.Zhao, " Response Time Analysis for Distributed Real-Time Systems with Bursty Job Arrivals," Proceedings of IEEE ICPP, 1998.

[4] Daniel Minoli, Delivering Voice Over IP Networks , 1998.

[5] C. Li, R. B., Wei Zhao, "New Delay Analysis in High Speed Networks," Department of Computer Science Texas A & M University, June 1999.

[6] V. Sivaraman and F.Chiussi, "Statistical Analysis of Delay Bound Violations At an Earliest Deadline First (EDF) Scheduler," In performance '99, Istanbul, Turkey, 1999.

[7] J. Walrand, P. Varaiya, "High Performance Communication Networks," 2000.

[8] K.H.Yum,E.J.Kim,andC.R.Das,"QoS Provisioning in Clusters: An Investigation of Router and NIC Design," in Proc. of ISCA, pp. 120--129,

June 2001.

[9] G. Bianchi, F. Borgonovo, A. Capone, L. Fratta, and C. Petrioli, “Endpoint Admission Control with Delay Variation Measurements for QoS in IP Networks,” ACM Computer Communication Review, Vol. 32, No. 2, pp. 61-19, April 2002.

[10] E. J. Kim, K. H. Yum, C. Das, M. Yousif, and J. Duato. “Performance enhancement techniques for infiniband architecture. In International Symposium on High Performance Computer Architecture,” Feb. 2003

[11] Frank Olaf Sem-Jacobsen, S. A. Reinemo, T. Skeie and O. Lysne, “Achieving Flow Level QoS in Cut-through Networks through Admission Control and DiffServ,” Vol. 3, pp. 1084-1090, Jun. 2004.

[12] L. Tan, Y. Yang, C. Lin and N. Xiong, “ PID-RPR: A High Performance Bandwidth Allocation Approach for RPR Networks,” IEICE Trans. Commun., Vol. E88-B, No. 7 July 2005.