Call Admission and End-to-end delay Allocation for Fair Queueing Networks

Y. P. Chu  W. C. Chiang  C. L. Lee  Y. H. Huang
Department of applied math. of National Chung Hsing University
Department of applied math. of National Chung Hsing University
Department of Infor-mation Management of HsiuPing Institute
Department of Infor-mation Management of HsiuPing Institute
No. 250, KuoKuang Rd., Taichung City, Taiwan 401
No. 250, KuoKuang Rd., Taichung City, Taiwan 401
No. 11, Gungye Rd., No. 11, Dali City, Taichung, Taiwan 412
No. 11, Gungye Rd., No. 11, Dali City, Taichung, Taiwan 412
ypchu@amath.nchu.edu.tw  wzjiang@vghtc.gov.tw  clee@mail.hit.edu.tw  ehhwang@mail.hit.edu.tw

Abstract

Many multimedia applications, including audio and video, require quality of service (QOS) guarantees from network. A typical user is only concerned with the QOS requirements on end-to-end basis and does not care about the local switching node QOS. In this paper, we propose a general framework to map the end-to-end QOS requirement into the local switching node QOS requirements. Most of recent research efforts only focus on worst-case end-to-end delay bound but pay no attention to the problem of distributing the end-to-end delay bound to local switch node. Therefore, we focus on the local switching node QOS requirement and design a novel QOS requirement allocation scheme to get better performance. Using the number of maximum supportable connections as the performance index, we derive an optimal delay allocation (OPT) policy. In addition, we also design an analysis model to evaluate the proposed allocation scheme and equal allocation (EQ) scheme, which apply to the switching nodes along the flow’s path with the Rate-controlled scheduling architecture, including a traffic shaper and a fair queueing scheduler.

1. Introduction

The newly applications of high-speed integrated service networks is to provide quality of service guarantees, such as packet loss ratio, end-to-end delay, and throughput, to follows requiring different classes of services. Hence, the network needs to reserve resources, for example, buffer and bandwidth, to promise the performance requirements for each application. There are two categories of guaranteed-service models in order to support guaranteed-service applications [4,5]. One is the deterministic guaranteed service. This model can promise the worst-case performance bound for all packets in the same connection and get better quality of service. The other is the statistical guaranteed service. This model only promises probabilistic performance bound but gets the higher utilization from network. No matter what the service is deterministic or statistical, the improvement of network utilization is an important consideration.

Recent research efforts only focus on
worst-case delay bound but pay no attention to
the problem of distributing the end-to-end delay
bound to the local switching node
[2,3,7-9,13,14,16-18]. How to map the
end-to-end QOS requirements into the local
switching node QOS requirements is the one of
the most important considerations for
maximizing network utilization. In order to
improve the network utilization, when the new
connections is admitted and enter to the network
after admission control, they equally allocate the
excess delay and reserve the same bandwidth at
each switch along the flow’s path. In [5], the
author has proposed a scheme that first
computes the aggregated local worst-case delay
bound, and calculates the difference between the
aggregated value and the application-required
delay. Then, the excess value is equally assigned
to the local switching node along the flow’s path.
However, it cannot improve the network
utilization efficiently. Our approach is based on
the mapping of QOS requirement into local
resources to be reserved at each scheduler. In
this paper, we addressed the following issues:

1. How to allocate the end-to-end
delay to the local delay?

2. How many connections for this type
of application can be allowed into
the network under the different
allocation policies?

3. Which factor will affect the
allocation policy?

We consider the Fair queueing packet
scheduling policy [3-5,12]. We use the
Rated-controlled service architecture to prevent
the traffic distortion [11,15,19,20]. The
Rated-controlled service architecture consists of
a traffic shaper and a scheduler. The traffic
shaper can shape the incoming traffic at each
switching node as they enter the network by
controlling the output time of packets in a traffic
shaper. Here, we use fair queueing as the
scheduler because the fair queueing is easy to
implement and offers a diverse set of delay
bounds to connections. In addition, routing of
connections is not addressed. We assume that
the routing decision has already been made. This
allows us to focus on the problem of resource
allocation without having to address the
combined problem of route selection and
resource allocation. In section 2, we define our
system architecture and show how to compute
the end-to-end delay bound under this
architecture. Section 3 will propose the
admission control procedure and analyze the
allocation policy. Section 4 will give some
important numerical results. Section 5 is our
conclusion and future direction.

2. System Architecture

![Fig. 1 The end-to-end delay computation model](image)

Fig. 1 The end-to-end delay computation model

![Fig. 2 The architecture of the switching node](image)

Fig. 2 The architecture of the switching node

In this section, we described our system
architecture and the computation of the
end-to-end delay bound. We first add a policing function for every connection at the entrance of the network to shape the incoming traffic. We define \( I_n(s,s+t) \) be the number of bits that arrive in the interval \([s,s+t]\) for connection \( n \).

Let \( I_n(s,s+t) = 0 \) for \( t \leq 0 \), otherwise \( I_n(s,s+t) > 0 \) for \( t > 0 \), and \( \sigma_n, \rho_n, t \geq 0 \), such that

\[
I_n(s,s+t) \leq \sigma_n + \rho_n t = \bar{I}_n(t)
\]

We call \( \bar{I}_n(t) \) the input traffic envelope. Where \( \sigma_n \) is the bucket size and \( \rho_n \) is the average arrive rate [11,15]. We assume all packets have the same length and equal to \( l \). Therefore, we use the leaky bucket model to characterize the input traffic. Suppose one application wants to enter the network, and this application has the input traffic envelope \( \bar{I}_n(t) \). Consider the end-to-end delay as the application’s QOS requirement. For a single connection \( n \) that passes through \( K \) switching nodes as in Fig. 1, we assume that before the packets enter the scheduler of node \( m \), they pass through the traffic shaper \( A_n \) as the fig.2 shown. Moreover, we assume the propagation delay for each output link is zero. Let \( D_s(A_n^w, A_n^{w+1}) \) denote the delay that a packet from connection \( n \) experiences between the time it exits shaper \( A_n^w \) and \( A_n^{w+1} \), for the fig.1 described. Ref. [10] has shown that

\[
D_s(A_n^w, A_n^{w+1}) \leq D_s^w + D_s(A_n^w \| A_n^{w+1}), \quad \cdots \cdots \quad (1)
\]

Where \( D_s^w \) is the upper bound of the scheduling delay at node \( m \), and \( D_s(A_n^w \| A_n^{w+1}) \) denote the upper bound of the delay in the shaper \( A_n^{w+1} \) where the input of shaper \( A_n^w \) is the output of the shaper \( A_n^w \). From the fig. 1, we find that the end-to-end delay \( D_n \) of the connection \( n \) with the input traffic envelope \( \bar{I}_n(t) \) is

\[
D_n = D_s(I_n \| A_n^1) + \sum_{i=0}^{K-1} D_s(A_n^w, A_n^{w+1}) + D_s^w.
\]

From equation (1), we have

\[
D_s \leq D_s(I_n \| A_n^1) + \sum_{i=0}^{K-1} D_s^w + \sum_{i=0}^{K-1} D_s(A_n^w \| A_n^{w+1}) + D_s^w.
\]

We will compute the end-to-end delay \( D_s \).

First, we consider \( D_s(A_n^w \| A_n^{w+1}) \). Note that if traffic envelope \( \bar{A}(t) \leq \bar{B}(t) \), then \( D(A \| B) = 0 \). Therefore, we use the same parameter for all traffic shaper at each switching node, that is, \( \bar{A}_n^w(t) = \bar{A}_n^{w+1}(t) = \bar{A}_n(t) \), for each \( m=1,2,..,K-1 \), then

\[
\sum_{i=0}^{K-1} D_s(A_n^w \| A_n^{w+1}) = 0
\]

Reference [17] has shown that use the identical traffic shaper \( A_n \) to replace the various traffic shapers for each switching node will have the same worst-case end-to-end delay bound. Hence, the worst case end-to-end delay bound \( D_n \) is given by

\[
D_n \leq D_s(I_n \| A_n^1) + \sum_{i=0}^{K-1} D_s^w
\]

Secondly, we consider \( D_s(I_n \| A_n^1) \). We assume \( \bar{I}_n(t) = \sigma_n + l + \rho_n t \). However, how to assign the parameter of traffic shaper \( A_n \)? The choice of this parameter will affect the
end-to-end delay bound. In the following, we will discuss this problem.

We first let $\overline{A}_n(t) = \sigma_n + h_nt$. For the traffic of the sources entering the first shaper, it must satisfy that the average output rate is great or equal to average input rate, $h_n \geq \rho_n$. Otherwise, it will result in an infinite delay at the first shaper. For the stability condition, we have $h_n \leq r_n$, at each switching node $m$, where $r_n$ is the available output capacity at switching node $m$. Therefore, for the envelope of the traffic shaper, we have the following inequality.

$$\rho_n \leq h_n \leq r_n$$

Note that the service rate $g_n^w$ for the connection $n$ at switching node $m$ must satisfy following equation.

$$h_n \leq g_n^w \leq r_n$$

Hence, we can define $h_n = \rho_n$, and the above equation will become

$$\rho_n \leq g_n^w \leq r_n.$$ \hfill (2)

Since this assignment can have the maximum flexibility for the allocation of the service rate $g_n^w$ at each local switching node, even it may not have the optimal (minimum) end-to-end delay bound. Because the evaluation for the different delay allocation policies is the purpose of this paper, we believe this restriction will not affect our results of evaluation. Now we define $\overline{A}_n(t) = \sigma_n + h_nt$. And then

$$D_n(\mathcal{I}_n \| \mathcal{A}_n^w) = \frac{\overline{L}_n(t) - \overline{A}_n(t)}{\rho_n} = \frac{l}{\rho_n}.$$ 

Thirdly, we consider $\sum_n D_n^w$. It is the scheduling delay bound at each switching node. Since the input of the scheduler is the output of the traffic shaper $A_n$. If we let the service rate of switching node $m$ for the connection $n$ is $g_n^w$, and the scheduling algorithm at each switching node is fair queueing, then the worst scheduling delay bound for connection $n$ at the switching node $m$ is [2,9,13,19]

$$D_n^w = \frac{\sigma_n}{g_n^w} + \frac{l}{C_n^w}.$$ \hfill (3)

Where $C_n^w$ is the capacity of the output link for switching node $m$, and $l$ is the packet length. Finally we can get the end-to-end delay bound $D_n$ for connection $n$ is

$$D_n = \frac{l}{\rho_n} + \sum_m \left( \frac{\sigma_n}{g_n^w} + \frac{l}{C_n^w} \right)$$ \hfill (4)

In next section, we propose an admission control procedure that we used by this result and analyze two allocation policies for the end-to-end delay requirement.

3. The analysis model

Take the service rate constraint in eq. (2) into eq. (3), we define

$$\max d_n^w = \frac{\sigma_n}{\rho_n} + \frac{l}{C_n^w},$$ \hfill (5)

and

$$\min d_n^w = \frac{\sigma_n}{r_n^w} + \frac{l}{C_n^w}.$$ \hfill (6)
Where $\max d_n^\alpha$ is the maximum value of delay bound from connection $n$ at node $m$, and $\min d_n^\alpha$ is the minimum value of delay bound from connection $n$ at node $m$.

Consider a single source-destination route, where there are $K$ nodes. If a connection $n$ wants to enter the network with the traffic envelope being $\tilde{I}_n(t) = \sigma_n + l + \rho_n t$, $t \geq 0$ and the end-to-end delay requirement is $D$. We compute the maximum value and minimum value of end-to-end delay $D^\max_n$ and $D^\min_n$ based on eqs. (4), (5), and (6) with

$$D^\min_n = \frac{l}{\rho_n} + \sum_{m=1}^{K} \max d_n^\alpha$$

and

$$D^\max_n = \frac{l}{\rho_n} + \sum_{m=1}^{K} \min d_n^\alpha$$

That is, under the condition of the input traffic envelope being $\tilde{I}_n(t) = \sigma_n + l + \rho_n t$, $t \geq 0$, if the end-to-end delay requirement of the new connection $n$ is belong to the interval $[D^\min_n, D^\max_n]$, it implies the network can support the need for this connection. The connection will be accepted and the excess delay should be allocated properly for network utilization. If the end-to-end delay requirement $D$ is larger than the maximum value of end-to-end delay $D^\max_n$, that means the delay requirement of the new connection is low, the network also can accept the connection and must allocate the local delay with the value $\frac{\sigma_n}{\rho_n} + \frac{l}{C}$ for each switching node $m$ to ensure the schedulable region under fair queueing scheduler. However, even there were many admission control methods proposed [1,6,12], we adopt our admission control procedure as follows:

(1) If $d < D^\min_n$, then the connection is rejected,

(2) If $d > D^\max_n$, then the connection can be accepted and we allocate the local delay with the value of $\frac{\sigma_n}{\rho_n} + \frac{l}{C}$ for each switching node $m$.

(3) If $D^\min_n \leq d \leq D^\max_n$, then the connection can also be accepted and we can make the proper allocation policy for the end-to-end delay requirement.

In our analysis model, we used the fair queueing scheduler as our scheduling algorithm, unlike FCFS scheduler, it service packets with separate queues for different connections even in cross traffic model. That is, cross traffic model is only a combination of several tandem models. Therefore, it is enough to show the differences for different allocation policies under tandem model. In the following, we analyze two different allocation policies: equal allocation (EQ) and optimal allocation (OPT) policy under a tandem network model.

### 3.1 EQ policy

Assume that the maximum value of end-to-end delay upper bound from node 1 to node $K$ is $D^\max_1$, the minimum value of end-to-end delay lower bound is $D^\min_1$. The end-to-end delay requirement for the new class application is $D$, $D^\min_1 \leq D \leq D^\max_1$. First, we consider an EQ policy that assigns an equal amount of the extra end-to-end delay
requirement for a connection to each switching node. In the following we will calculate the number of allowable connections, $N_{eq}$, for this new class under the EQ policy. With this value, we can evaluate the allocation policy and derive the delay allocated to node $m$. Let $\Delta d^m$ be the amount of the extra delay at node $m$, then

$$\Delta d^m = \frac{d - D_{min}}{K},$$

Since the amount of the extra delay is allocated to each switching node equally. $\Delta d^m = \Delta d$, for $m = 1, 2, \cdots, K$. Let $N^m$ be the number of supportable connections for node $m$, and $\text{min} d^m$ be the minimum value of delay bound for connection $n$ at node $m$. Consider the worst case, then

$$\text{min} d^m + \Delta d = N^m \left( \frac{\sigma_n}{r_n} + \frac{l}{C_n} \right), \quad \text{----------------(7)}$$

We have

$$N^m = \frac{\text{min} d^m + \Delta d}{\frac{\sigma_n}{r_n} + \frac{l}{C_n}}, \quad \forall m = 1, 2, \cdots, K, \quad \text{----------------(8)}$$

Let

$$N_f = \text{min} \{ N^1, N^2, \cdots, N^K \}. \quad \text{----------------- (8)}$$

For the stability condition, the sum of all arriving rate must be no more than the capacity of all the links. Therefore,

$$\sum_{i=1}^{n} \rho_i + N_i \rho_n < C^n, \quad \forall m = 1, 2, \cdots, K,$$

Where $N_e$ is the maximum number of new connection under stability condition. And then,

$$N_e < \frac{C^n - \sum_{i=1}^{n} \rho_i}{\rho_n}, \quad \text{------------------- (9)}$$

Combine eqs (8), (9), then we have

$$N_{eq} = \text{min} \{ N_f, N_e \}. \quad \text{------------------ (10)}$$

Finally, Replacing $N^m$ of eq.(7) with $N_{eq}$, we can derive the delay that is allocated to node $m$.

### 3.2 OPT policy

Next, we consider the optimal allocation policy. We want to determine the value of the allocated delay that maximize the number of connections. Let $N$ be the maximal number of connections for the new class. And let $\Delta d^m$ be the amount of extra delay at node $m$ in OPT policy, then

$$\sum_{n=1}^{K} \Delta d^n = D - D_{min},$$

Where $D$ is the end-to-end delay requirement of new application, and $D_{min}$ is the minimum value of end-to-end delay.

For each switching node $m$ considering the worst case,
\[ \min d^* + \Delta d^* = N \left( \frac{\sigma_s}{r^p} + \frac{l}{C^*} \right), \quad \ldots \quad (11) \]

Then,

\[ N = \min \frac{d^* + \Delta d^*}{\frac{\sigma_s}{r^p} + \frac{l}{C^*}}, \]

We let

\[ N_s = \left| \frac{\min d^* + \Delta d^*}{\frac{\sigma_s}{r^p} + \frac{l}{C^*}} \right|, \quad \ldots \quad (12) \]

For the stability condition, the sum of all arriving rate must be no more than the capacity of all the links. Therefore,

\[ \sum_{i=1}^{m} \rho_i + N_s \rho_s < C^*, \quad \forall m = 1, 2, K, K, \]

Where \( N_s \) is the maximum number of new connection under stability condition. And then,

\[ N_s < \frac{C^* - \sum_{i=1}^{m} \rho_i}{\rho_s}, \]

\[ N_s = \left| \frac{C^* - \sum_{i=1}^{m} \rho_i}{\rho_s} \right|, \quad \ldots \quad (13) \]

Combine eqs (12), (13), then we have

\[ N_{opt} = \min \{ N_s, N_s \}, \quad \ldots \quad (14) \]

Finally, Replacing \( N \) of eq.(10) with \( N_{opt} \), we can derive the delay that is allocated to node \( m \).

4 Numerical results

In this section, we present some numerical examples to compare the performance of the EQ and OPT delay allocation polices. Two performance measure indexes are adopted to decide the efficiency and optimality of the allocation policies. One is the network utilization, in terms of the number of supportable connections. Moreover, another index is the relative gain (RG) to evaluate the optimality of the allocation policy.

Fig. 3 shows this model consisting of a single source-destination pair of nodes in an ATM network. Each packet be assigned to the same size \( l = 53 \) bytes. Suppose the network has three nodes and only a single route. The capacity for the links of all nodes is 300 packets per second, and the remained capacity for the link of node 1 and node 3 are 150 packets per second. The input traffic follows the leaky-bucket constrained with burst size \( \sigma = 10 \) packets and average rate \( \rho = 5 \) packets per second. The value of \( \sigma \) must be great than \( l \) to ensure that at least one packet can be filtered by the traffic shaper.

With the remained capacity for the link of node 2 being 120 packets per second, we calculate the number of supportable connections for EQ and OPT delay QOS allocation policies by eq. (10) and (14) respectively. Fig. 4 shows...
the variation in the network utilization over end-to-end delay requirement. The end-to-end delay requirement $D$ for new connections is ranged from 1 to 6.6. From the results we find that the OPT allocation policy performs better than the EQ policy because the number of connections admitted by the former is far more than the later in each delay bound in the end-to-end delay requirement. It implies that the OPT policy has higher network utilization than EQ policy. This is because that OPT policy is based on each node’s situation to allocation local delay, so it can prompt more network performance than EQ policy which allocate local delay equally.

![Fig. 4 Delay requirement vs. no. of connection](image)

Fig. 4 Delay requirement vs. no. of connection

![Fig. 5 No. of admitted connections vs. the end-to-end delay requirement with the different remained capacity of node 2](image)

Fig. 5 No. of admitted connections vs. the end-to-end delay requirement with the different remained capacity of node 2

connections over the end-to-end delay requirement in the network model with the remained capacity of node 2 varied. By the way, we can observe the variation of the network utilization over the end-to-end delay requirement when the bottleneck link capacity changed. We find that, the OPT policy performs better than the EQ policy under any network conditions.

![Fig. 6 RG vs. bottleneck ratio](image)

Fig. 6 RG vs. bottleneck ratio

We compute the relative gain (RG) for OPT policy relative to EQ policy with different bottleneck ratio. The RG is defined as follow:

$$RG = \frac{N_{\text{opt}} - N_{\text{eq}}}{N_{\text{eq}}}.$$
network resources efficiently.

5. Conclusion

In this paper, we applied the rate-controlled service architecture with the fair queueing packet-scheduling algorithm to evaluate the effects of the allocation policies. With the number of maximum allowable connections as the performance index, a solution for an optimal delay allocation was derived. The results of this analysis have shown the relationship of the delay to the number of maximum allowable connections. In addition, from the numerical results, we have found that the bottleneck ratio will influence the performance of the allocation policy.

Of course, the local allocation of end-to-end QOS requirement must be combined with an admission control algorithm. When the traffic model parameters, link bandwidth in each node along the path connection pass through, and propagation delay are known, the network can compute the local QOS requirement in each node according to the analysis results in this paper.

The performance index may be changed depending on the various requirements of the network and end-users. In the future, we will determine the diverse performance index to design and evaluate a method for local QOS allocation.

References


