<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Statistical connection admission control framework based on achievable capacity estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Zhu, H; Li, VOK; Ma, Z; Zhao, M</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>IEEE International Conference on Communications, 2006, v. 2, p. 748-753</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2006</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/45938">http://hdl.handle.net/10722/45938</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>©2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.; This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
Statistical Connection Admission Control Framework based on Achievable Capacity Estimation*

Huling Zhu¹, Victor O. K. Li¹
1Dept. of Electrical and Electronic Eng.
University of Hong Kong, Hong Kong, China
[hlzhu, vli]@eee.hku.hk

Zhengxin Ma²
2Dept. of Electronic Eng.
Tsinghua University, China
mazz@mail.tsinghua.edu.cn

Miao Zhao³
3Dept. of Electrical and Computer Eng.
Stony Brook University, NY 11794, USA
mzhao@ic.sunysb.edu

Abstract—Traditional traffic descriptor-based and measurement-based admission control schemes are typically combined with a node by node resource reservation scheme, rendering them unscaleable. Although some Endpoint Admission Control schemes can resolve this problem, they impose significant signaling overhead. To cope with these two problems, this paper proposes a statistical connection admission control framework which can easily and efficiently estimate the network resource for a pair of ingress-egress nodes and make admission decision based on this estimated result. In this framework, the network is considered as a “black box.” For a certain ingress-egress node pair, the egress node measures the QoS constraint violation ratio and feeds this information back to the ingress node periodically. With this information and the measured statistical characteristics of the existing aggregated traffic, the ingress node estimates the achievable capacity between the ingress-egress node pair, and makes the admission decision for a new traffic connection request. The signaling overhead of this framework is very small. Simulation results show the effective throughput is relatively high.

Keywords- QoS; admission control; statistical estimation

I. INTRODUCTION

Real-time applications have stringent requirements on Quality of Service (QoS) metrics, such as packet loss probability and end-to-end delay. Connection Admission Control (CAC) is one of the primary mechanisms to provide QoS guarantees. A new traffic connection can be accepted if and only if the network can accommodate the new connection while still maintaining levels of QoS promised to the existing accepted connections.

Traditional connection admission control can be roughly classified into traffic descriptor-based and measurement-based approaches [1][2]. Algorithms on these approaches put primary focus on extracting statistical characteristics of individual or aggregated traffic flow and derive equations to estimate the impact of the newly arrived connection on the performance of the existing traffic connections. Generally, these algorithms assume that a source node has accurate information of the network resource provided for the aggregated traffic between the source node and a destination node, such as the link capacity and the buffer size. Thus, the implementation of these algorithms often involves a resource reservation mechanism. However, in some QoS service architecture without resource reservation mechanisms, for example Differentiated Services architecture, it is infeasible to accurately obtain the capacity provided to a source-destination pair. Moreover, typically a simple dumb-bell network model is assumed for the analyses of these CAC algorithms. In this model, a single link is assumed between a pair of source-destination edge nodes called the ingress and egress edge nodes, respectively. In attempting to extend the single link model to the multiple-link environment, state coordination among nodes is required. For example, in some schemes, the end-to-end QoS constraint, e.g. the end-to-end delay, must be decomposed to an individual constraint on each link, which imposes a complex optimization problem of decomposition [3][4]. These factors impose scalability restriction to these kinds of admission control schemes.

To cope with this issue, some researchers proposed Endpoint Admission Control [5]-[9]. The common idea of Endpoint Admission Control [5]-[7] is that the admission decision is made by the network endpoint (the egress edge node or the host) and is based on injecting “probing” packets into the network at setup to measure the level of network services, e.g. packet loss probability. However, this kind of endpoint admission control has its flaws. The set-up delay is substantial, thus limiting its appeal to delay-sensitive applications. Furthermore, under sufficiently high loads, the opportunity of simultaneous probing by many sources may lead to a situation known as thrashing, that is, even though the number of admitted flows is small, the cumulative level of probe packets prevents further admissions [6]. References [8] and [9] propose an egress admission control framework, and introduce the measurement-based arrival envelope and the measurement-based service envelope concept to perform admission control at the egress node. This approach does not rely on resource reservation and on probing packets. However, it still has some deficiencies. The peak rate and rate variance of the arrival envelope are measured based on the time-stamp in each packet. Thus, packet loss will degrade the accuracy of the measured envelope. In addition, in this approach, the egress node makes the admission decision and it needs to get the reservation request from a new connection request. That is to say, each new connection must send a reservation request to the egress node. This kind of signaling overhead may be very large especially when traffic load is heavy.

To deal with the above scalability and overhead problems and to provide statistical QoS guarantees for traffic with packet loss ratio requirement, a novel statistical CAC framework, denoted as SCAC in this paper, is proposed. In order to design an effective CAC framework, we study the function of CAC.
In a CAC process, when a new request arrives, the edge node/host (at the source or the destination side) makes the admission decision based on the achieved impact of the new connection request on the transmission performance. The estimation is generally based on some obtained information, such as traffic statistical characteristics and packet loss ratio. Thus, there are three key factors for an effective CAC process: the node to perform the admission decision, effective information metrics to estimate the impact of a new request, and a simple method to obtain the information at low cost. In our proposed SCAC framework, the ingress edge node is in charge of the admission decision. The information metrics for making admission decision include the statistical characteristics of the existing traffic connections measured at the ingress node and the packet loss ratio measured at egress. The packet loss ratio is measured and fed back from the egress node to the ingress node periodically. With the measured information, an estimation procedure is proposed to estimate the achievable capacity utilized by a given pair of ingress and egress nodes. Based on the estimated achievable capacity, the packet loss probability associated with admitting a new connection request is predicted. In this admission control framework, only the packet loss ratio is required to be collected from the egress node periodically. Compared with current end-point admission control approaches, the signaling overhead of our proposed approach is small.

The rest of this paper is organized as follows. Section II describes the proposed admission control framework. Section III presents an implementation scheme for the admission control framework. Section IV presents the simulation results to analyze the performance of SCAC. Conclusion and future work are given in Section V.

II. ADMISSION CONTROL FRAMEWORK

A. Objective

Our objective is to develop an admission control framework which has the following properties:

- Scalability: if a network consists of edge nodes and core nodes, there is no requirement for per-flow signaling or state management at core nodes;
- Flexibility: the framework should be adaptive to different service models and traffic sources;
- Low overhead: the overhead of collecting information for making admission decision should be light;
- High utilization: the utilization of network resources should be effective and high.

To achieve our objective, we use a “black box” network model as shown in Fig. 1. Since the path from an ingress node to an egress node will consist of multiple physical network links, which may be shared by traffic corresponding to other ingress-egress pairs, we call all such traffic from other ingress-egress pairs transit traffic. In this model, the transit traffic characteristics and the network information, such as the packet service disciplines and routing methods, are hidden from the connection admission decision between the given ingress-egress pair.

B. Admission Control Framework

If the utilized network resource occupied by a pair of edge nodes can be accurately estimated, we can establish an efficient CAC framework based on the existing measurement-based admission control (MBAC) approaches. The main obstacle is how to easily and efficiently estimate the utilized network resources. We select the ingress node to perform admission decision and introduce a mechanism to estimate the network resources utilized by a pair of ingress-egress nodes. In this paper, we call this estimated network resource for a pair of edge nodes the achievable capacity of this nodes pair. The estimation is performed by the ingress node, based on the performance information fed back from the egress node, as shown in Fig. 1. With the estimated achievable capacity and current MBAC models, the admission decision can be made.

According to the above analyses, our admission control framework includes three parts: information collection, estimation of achievable capacity for the ingress-egress node pair, and admission decision.

1) Information collection: We characterize the statistical QoS service provided for each application by the network in terms of probability. For example, suppose the QoS metric of an application is end-to-end packet delay, then if the end-to-end delay of a packet is larger than $D$, we say that a violation event has occurred. The occurrence probability of the violation event is called the violation probability. For a QoS requirement on packet loss, a packet loss is looked at as a violation event and the packet loss probability is the violation probability. The statistical QoS service is to guarantee the violation probability to be smaller than a given threshold denoted as the QoS violation probability constraint. In our SCAC framework, the collection of violation probability is necessary to provide statistical QoS guarantee. So, the egress node measures the fraction of violation events (called the violation ratio) for this pair of nodes and feeds this information back to the ingress node. The ingress node takes this ratio as the violation probability, and with the measured stochastic characteristics of the existing aggregated traffic, makes admission decisions.

2) Estimation of achievable capacity (for an ingress-egress pair): For a given QoS violation probability constraint, [10]-[17] have derived mathematical models to calculate the required capacity to guarantee the constraint. One example of such capacity is the equivalent capacity, which is the capacity to guarantee that the constraint for an aggregated traffic with stationary arrival rate is not violated [10]-[14]. According to the measured statistical characteristics of accepted aggregated
traffic between a pair of ingress-egress nodes and the violation ratio fed back from the egress node, the ingress node can estimate the capacity achieved by the aggregated traffic according to these existing mathematical models. To distinguish this capacity from the traditional concept of equivalent capacity, we call it achievable capacity.

3) Admission decision: According to the statistical characteristics of a new connection request and the estimated achievable capacity, the violation probability is estimated. If the violation probability is smaller than the given threshold, the new connection is accepted. Otherwise, it is rejected.

Fig. 2 shows the proposed admission control framework.

III. IMPLEMENTATION OF SCAC

In this section, we provide an implementation method of our proposed SCAC framework. Considering that real-time traffics have stringent requirements on end-to-end packet delay and packet loss probability, this paper selects packet loss as the QoS metric. With this metric, the QoS violation probability constraint for an ingress-egress pair is that the end-to-end packet loss probability, this paper selects packet loss as the trajectory rate ) () τ τ + − .

\[ P_{loss} = P(\text{end-to-end packet loss}) \leq \varepsilon_{loss} \quad (1) \]

A. Information Collection

We define three time scales to measure packet loss ratio and aggregated traffic flow behavior. The first is the sampling period, the second is the collection window for packet loss ratio measurement, and the third is the measurement window. These quantities are more formally defined as follows.

1) Sampling Period \( \tau \): To capture the stochastic characteristics of the existing aggregated traffic flow, we hope to measure the aggregated traffic rate at an arbitrary time \( \tau \). However, in practical systems, it is impossible to measure the instantaneous rate. For this reason, we divide time into slots of length \( \tau \), and the aggregated traffic rate \( R(i) \) at time slot \( i \) denotes the time average of the traffic coming into the network from ingress node in the time interval \([i\tau, (i+1)\tau)\).

\[ R(i) = A(i)/\tau, \quad i = 0, 1, 2, \ldots \quad (2) \]

where \( A(i) \) is the amount of traffic in bits transmitted between the given ingress-egress nodes pair in the interval \([i\tau, (i+1)\tau)\).

2) Collection Window \( T_{loss} \): The egress node counts the number of lost packets and calculates the packet loss ratio every \( N \) sampling periods, i.e. \( T_{loss} = N \cdot \tau \). The \( n \)th collection window is the interval \([n \cdot N \cdot \tau, (n+1) \cdot N \cdot \tau)\). In this interval, the measured packet loss ratio is denoted as \( \varepsilon(n) \). The information of packet loss ratio is fed back to the ingress node with period of \( T_{loss} \).

3) Measurement Window \( T_M \): A measurement window is taken to be \( M \) sampling periods, i.e. \( T_M = M \cdot \tau \), where \( M \) is chosen to be large enough to reflect the statistical behavior of the aggregated traffic flow. The measurement window is smoothed slot by slot. At time slot \( i \), the measurement window is \([ (i - M + 1) \cdot \tau, i \cdot \tau]\).

B. Estimation of Achievable Capacity

At the ingress node, the achievable capacity estimation is triggered by the reception of the packet loss ratio information. Thus, this estimation is performed with the period \( T_{loss} \). Assume that at time slot \( n \), the ingress node receives measured packet loss ratio \( \varepsilon(n) \) from the egress. In current measurement window, the measured traffic is characterized as a series of aggregated traffic rates \( \{R(i)\}_{i=n-M+1, \ldots, n} \). For the existing traffic connections with the measured characteristics, the achievable capacity is defined as the bandwidth that can guarantee the packet loss ratio to be smaller than \( \varepsilon_{loss} \). Then, the estimation problem becomes how to calculate the achievable capacity \( C_{achv}(n) \) with the parameters \( \{R(i)\}_{i=n-M+1, \ldots, n} \) and \( \varepsilon(n) \).

\[ C_{achv}(n) = G(\{R(i)\}_{i=n-M+1, \ldots, n}, \varepsilon(n)) \quad (3) \]

Obviously, to estimate \( C_{achv}(n) \), we should describe the stochastic characteristics of the aggregated traffic. Addie et al. [18] has shown that if the Central Limit Theorem applies to traffic processes, as more traffic is aggregated, due to the sharing of a link by more and more traffic streams, the traffic becomes more Gaussian, or, more formally, weakly converges to a Gaussian process. This relies on the assumption that the traffic arriving in a given time interval has finite variance, a fact which seems to be confirmed by measurements. It is appropriate to say that if sufficient traffic is aggregated, a
Gaussian model behaves much the same as the aggregated traffic [10]. Thus, in this paper, we describe the aggregated traffic behavior with a Gaussian model. When the packet loss ratio information arrives at the ingress node at time slot \( n \), the ingress node calculates the mean and variance of aggregated traffic rate for the current measurement window, as follows:

\[
\bar{R}(n) = \frac{1}{M} \sum_{i=n-M+1}^{n} R(i)
\]

\[
\sigma^2_R(n) = \frac{1}{M-1} \sum_{i=n-M+1}^{n} (R(i) - \bar{R})^2
\]

Based on the Gaussian model, we have the following estimation methods according to the large deviation approximation.

\[
C_{achv}(n) = \min \{ C | P(R > C) \leq \varepsilon(n) \}
\]

\[
C_{achv}(n) = \bar{R}(n) + \sigma_R(n) \cdot \sqrt{-2 \ln(\varepsilon(n)) - \ln(2\pi)}
\]

Here, we only use the Gaussian model to illustrate the estimation of the achievable bandwidth. In fact, for different traffic sources with different QoS requirements, different estimation models can be implemented in our SCAC framework. Thus, this admission control framework can be adapted to different service models and traffic sources.

C. Admission Decision Criteria

A new connection request will inform the ingress node the mean and variance of its traffic flow, denoted as \( \bar{r} \) and \( \sigma^2_r \). Assume that when a connection request arrives, the mean and variance of the aggregated traffic in the current measurement window are \( \bar{R} \) and \( \sigma^2_R \), and the estimated achievable capacity is \( C_{achv}(n) \). According to (8) and (9), the ingress node estimates the packet loss probability \( \varepsilon_{est} \), as follows:

\[
C_{achv}(n) = (\bar{R} + \bar{r}) + \sqrt{\sigma^2_R + \sigma^2_r} \cdot \sqrt{-2 \ln(\varepsilon_{est}) - \ln(2\pi)}
\]

\[
\varepsilon_{est} = \frac{1}{\sqrt{2\pi}} \exp\left[ -\frac{(C_{achv}(n) - (\bar{R} + \bar{r}))^2}{2(\sigma^2_R + \sigma^2_r)} \right]
\]

If \( \varepsilon_{est} < \varepsilon_{loss} \), accept the new connection. Otherwise, reject it.

IV. SIMULATIONS AND PERFORMANCE ANALYSES

A. Topology

In this section, we evaluate the performance of the proposed admission control framework via a set of simulations under the network topology as shown in Fig. 3.

There are two kinds of traffic connections: the admission-controlled connections are the traffic governed by the admission control at the ingress node; and the transit connections traffic which characterizes traffic due to other ingress-egress pairs. Both kinds of traffic aggregate at the ingress node and share the same buffer. Although it is a simplified scenario, it is representative of more complex network scenarios when a variety of transit traffic models are used. In the simulations, packets of each connection are labeled with sequence number. In each collection window, the egress node counts the number of lost packets for each connection according to the sequence number of the last received packet. In all simulations, the capacity of the link between the ingress and egress node, denoted as \( C \), is 10Mbits/s, the sampling period is 0.01s, and the collection window and measurement window take the value of 1s. In our simulations, we use three types of traffic sources as shown in Table 1. Source 1 is an On-Off model with exponential on and off times. Source 2 is a Poisson model with exponential packet inter-arrival times. Source 3 is a Bursty Source whose packet inter-arrival time distribution is the Pareto distribution with the scale parameter \( a \) and the shape parameter \( c \). The transit traffic connection has no QoS requirement and is generated from these three types of sources for different simulations. The admission-controlled traffic is generated by multiple of independent traffic sources each of which is a Type 1 (On-Off) source. The packet loss ratio of admission-controlled traffic connections is constrained to be less than 0.01, i.e., \( \varepsilon_{loss} = 0.01 \).

The inter-arrival time of these traffic sources follows an exponential distribution with mean value \( \bar{t}_i \) varying with the traffic load simulated. Assume in one simulation, the transit traffic is Type \( i \) \((i = 1, 2, 3)\). The traffic load \( \rho \) between the ingress-egress pair is defined as (10)

\[
\rho = \frac{\bar{t}_i \cdot \bar{r}_i / \bar{v}_i + \bar{d}_i / \bar{v}_i}{C} \quad (i = 1, 2, 3)
\]

where \( \bar{v}_i \), \( \bar{r}_i \) and \( \bar{d}_i \) are average inter-arrival time, average rate and average life time of connections of Source Type \( i \), respectively.

B. Performance Analyses

1) Link Utilization: We consider two performance metrics.

- The utilization of the link between ingress and egress nodes \( G_{agg} \)

\[
G_{agg} = \text{Link Throughput} / C
\]

<table>
<thead>
<tr>
<th>Type 1: On-Off</th>
<th>Type 2: Poisson</th>
<th>Type 3: Bursty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burst Rate (bits/s) = 9.6K</td>
<td>Packet Size (bits) = EXP(480)</td>
<td>Packet Size (bits) = EXP(480)</td>
</tr>
<tr>
<td>On Time (s) = exponential(0.05)</td>
<td>Packet Interval (s) = exponential(( \tau_{exp} ))</td>
<td>Packet Interval (s) = Pareto(( a, c ), c = 1.25)</td>
</tr>
<tr>
<td>Off Time (s) = exponential(0.05)</td>
<td>Off Time (s) = 20</td>
<td>Life Time (s) = simulation duration</td>
</tr>
</tbody>
</table>

Figure 3. Simulation Topology

TABLE I. TRAFFIC SOURCES

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.
**The satisfaction ratio of aggregated admission-controlled traffic $\alpha_e**

\[
\alpha_e = \begin{cases} 
\frac{N_{\text{satisfied}}}{N_{\text{accepted}}}, & N_{\text{accepted}} > 0 \\
1, & N_{\text{accepted}} = 0 
\end{cases} 
\]  \tag{12}

where $N_{\text{satisfied}}$ is the number of connections whose packet loss ratio is smaller than $\epsilon_{\text{loss}}$ and $N_{\text{accepted}}$ is the number of accepted connections.

Here we compare these performance metrics of the proposed framework with those of the MBAC based on the Gaussian model and the peak rate based admission control scheme (PCAC). In these comparisons, the transit is characterized by multiple Type 1 traffic sources (On-Off). The traffic connection inter-arrival time of the admission-controlled traffic connection and the transit traffic connection follow the exponential distribution with means of 0.032s, 0.024s, 0.02s, 0.016s, 0.012s and 0.0096s, respectively, corresponding to traffic load $\rho = 0.6, 0.8, 0.98, 1.2, 1.6$ and 2.

We also show the impact of the buffer size on these performance metrics with buffer size of 10 packets and 100 packets, respectively. The comparison results are shown in Fig. 4 and 5. It can be seen that the buffer size has almost no effect on the performance of PCAC. With the increase of the buffer size, the satisfaction ratio of MBAC and SCAC are improved. The link utilization of our proposed SCAC scheme is much higher than the PCAC scheme. The link utilization of SCAC is close to that of the MBAC scheme. When the buffer size is 100 packets, the satisfaction ratio of SCAC is even better than that of MBAC. The reason is that when the buffer size is small, the Gaussian model may not be fit the aggregated traffic very well. The admission decision of our proposed admission control scheme is not only based on the measured characteristics of the existing aggregated traffic, but also on the packet loss information fed back from the egress node. Thus, it can achieve better performance.

2) Impact of Traffic Burstiness

This part analyzes the impact of the burstiness of transit traffic on the performance of SCAC. In these simulations, the buffer size is 100 packets. The transit traffic connection is assumed to be generated from Type 2 (Poisson) and Type 3 (Bursty source). The admission-controlled connection still follows the On-Off traffic source and the connection inter-arrival time follows the exponential distribution with means of 0.032s, 0.024s, 0.02s, 0.016s, 0.012s, and 0.0096s. Under different simulations, the transit traffic parameter varies with different traffic loads of 0.6, 0.8, 0.96, 1.2, 1.6, and 2.0. For the Poisson traffic, the mean of the packet inter-arrival time of the transit traffic assumes values of 0.00016, 0.00012, 0.0001, 0.00008, 0.00006 and 0.000048, respectively. For the bursty traffic, the packet inter-arrival time of the transit traffic follows the Pareto distribution with shape parameter $c = 1.25$, and scale parameter $a$ assuming values of 0.000032, 0.000024, 0.00002, 0.000016, 0.000012 and 0.0000096, respectively. Assuming the rate of the transit traffic is $r_t(n)$ at time slot $n$ and $R_t = \{r_t(n)\}_{n=1,2,\ldots}$, we use (13) to describe the burstiness of the transit traffic.

\[
J_{\text{poisson}} < J_{\text{bursty_pareto}} \tag{14}
\]

For the source Type 2 and Type 3, expression (14) holds.

We consider three performance metrics: the utilization of the link between the ingress and egress nodes, the satisfaction ratio of the admission-controlled traffic, and the packet loss ratio estimation error.

Assuming the estimated packet loss ratio is $\epsilon_{\text{est}}(n)$ of the $n$th estimation period and the corresponding measured packet loss ratio is $\epsilon(n)$, the estimation error is defined as (15)

\[
E_{\text{err}} = \frac{1}{K} \sum_{i=1}^{K} (\epsilon_{\text{est}}(i) - \epsilon(i))^2 \tag{15}
\]

$K$ is equal to the value of simulation duration divided by estimation period $T_{\text{loss}}$. 
the network must be minimized. For better performance, the fluctuation of the aggregated traffic flow in the network must be minimized.

V. CONCLUSION

Traditional traffic descriptor-based and measurement-based admission control schemes are not scalable since node by node resource reservation schemes are required for their implementations. Endpoint Admission Control schemes are proposed to resolve this problem. However, these schemes impose significant signaling overhead. To deal with these two problems, this paper proposes a statistical connection admission control framework which is based on the feedback control mechanism. In this framework, an achievable capacity estimation mechanism for a pair of ingress-egress nodes is introduced. For an ingress-egress pair, the achievable capacity is defined as the network resource utilized by this node pair under a measured packet loss ratio. The egress node measures the QoS constraint violation ratio and feeds this information back to the ingress node periodically. With this information and the measured statistical characteristics of the existing aggregated traffic, the ingress node estimates the achievable capacity between the ingress-egress node pair, and makes the admission decision for a new connection request. Since only the violation probability is required to be fed back, the signaling overhead of this framework is very small. Based on simulations, we show the estimation error of the proposed admission scheme based on this framework is very small and the satisfaction ratio is relatively high. The fluctuation of aggregated traffic in a network has great impact on the performance of SCAC. Thus, in our future research, we shall focus on minimizing the fluctuation of traffic flow in the network.

REFERENCES