A Scalable Call Admission Control Algorithm

Zafar Ali, Member, IEEE, Waseem Sheikh, Member, IEEE, Edwin K. P. Chong, Fellow, IEEE, and Arif Ghafoor, Fellow, IEEE

Abstract—In this paper we propose a scalable algorithm for connection admission control (CAC). The algorithm applies to a Multiprotocol Label Switching (MPLS) ATM switch with a FIFO buffer. The switch carries data from statistically independent variable bit rate (VBR) sources that asynchronously alternate between ON and OFF states with exponentially distributed periods. The sources may be heterogeneous both in terms of their statistical characteristics (peak cell rate, sustained cell rate, and burst size attributes) as well as their Quality of Service (QoS) requirements.

The performance of the proposed CAC scheme is evaluated using known performance bounds and simulation results. For the purpose of comparison, we also present scalability analyses for some of the previously proposed CAC schemes. Our results show that the proposed CAC scheme consistently performs better and operates the link close to the highest possible utilization level. Furthermore, the scheme scales well with increasing amount of resources (link capacity and buffer size) and accommodates intelligently the mix of traffic offered by sources of diversified burstiness characteristics.

Index Terms—Traffic Management, Multiprotocol Label Switching (MPLS), Call Admission Control (CAC).

I. INTRODUCTION

In today’s backbone networks the demand for equipment and applications which support simultaneously traditional ATM services and optimized IP transport using MPLS is continuously growing. The use of MPLS has also resulted in new services such as IP Virtual Private Networks (VPNs) to both IP+ATM (Cisco’s trademark for equipment that simultaneously support IP transport using MPLS and ATM services) networks and router networks. MPLS is used to integrate IP over ATM inside the Autonomous System (AS). This integration is done using IP+ATM in which the ATM switches continue to operate according to the ATM Forum and ITU-T standards while running MPLS in parallel [1]. Our CAC scheme provides a scalable admission control mechanism for such switches.

The role of admission control algorithm in a network is to decide whether or not to admit a new traffic flow such that all the previously admitted flows continue to receive their requested QoS requirements. A large number of admission control algorithms exist in literature. A comprehensive survey of these algorithms was provided in [2]. In [2] the authors classify various admission control algorithms and perform a large number of experiments to evaluate the accuracy and efficiency of these algorithms. Our proposed scheme comes under the category of additive effective bandwidths, in which each flow independently reserves a bandwidth for itself known as the effective bandwidth. One of the important considerations in finding the effective bandwidth is the underlying traffic model. Most of the traffic in today’s networks such as web traffic and compressed video traffic is long range dependent (LRD) traffic. In [3] the authors present a measurement based effective bandwidth estimation for LRD traffic.

A traffic contract for variable bit rate (VBR) video connection is made by signaling the Cell Loss Ratio (CLR) requirement, Peak Cell Rate (PCR), Sustained Cell Rate (SCR), and maximum burst size attributes of the traffic [4]. These parameters are used by the network to compute the Equivalent Cell Rate (ECR) for the connection and to realize a Connection Admission Control (CAC) policy based on the ECR value [4], [5], [6], [7].

While the notion of effective capacity yields a simple CAC criterion, its effectiveness is dictated by how well the ECR value represents the resource requirement of the source under question. A good CAC algorithm strives to achieve a proper balance of link efficiency and QoS guarantee. A desirable objective, which is often ignored in evaluating performance of a CAC algorithm, is its ability to achieve high link utilization under different traffic characteristics, numerous resource setups, and diverse QoS requirements.

In this paper we propose a CAC algorithm that scales well with increasing amount of resources (link capacity and buffer size) and accommodates intelligently the mix of traffic offered by sources of diverse burstiness characteristics. The major contributions of this paper include a scalable CAC algorithm for an MPLS ATM switch with a FIFO buffer and heterogeneous sources, its performance analysis using analytical bound and simulation results, and benchmark comparison of the performance of some of the previously proposed CAC algorithms with our proposed algorithm. Our results show that the proposed algorithm renders a superior CAC scheme that guarantees QoS while offering a high link utilization.

The paper is organized as follows. Section II describes some approximations that are used in designing various CAC algorithms. This section also outlines a simple approximation that forms the basis for the proposed CAC scheme. Section III describes our proposed CAC algorithm. In Section IV, we study scalability aspects of the proposed algorithm and evaluate its performance. For this purpose, the section defines some scalability metrics and outlines some of the computationally attractive CAC algorithms that are used in commercial MPLS.
ATM switches. Section V concludes the paper with some comments. Table I summarizes the notations used in the following discussion. In this table, all the parameters with subscript $i$ refer to the source $i$. The meaning of these parameters becomes clear as we proceed through the paper. The CAC algorithms considered in this paper rely on using the Mean Burst Size.

### II. ASYMPTOTIC APPROXIMATIONS FOR CAC

To design a CAC algorithm that can run the link at a high utilization level while meeting the QoS requirements, we need a good measure to reflect the state of congestion at an ATM node. Once such a measure is found, one could make a CAC decision based on whether or not admission of the source in question drives the congestion to an unacceptable value. Therefore, the accuracy of the congestion measure is important for achieving high utilization of network resources and for guaranteeing the performance for the admitted connections.

The steady state distribution of the queue length, $\mathbb{P}(\{Q > x\})$, has been considered as a prime measure of network congestion at a switching node [8], [9], [10], [11], [12], [13]. However, even for traffic models such as the Batch Markov Arrival Processes (BMAP) or the Markov Modulated Fluid Processes (MMFP), for which exact analytical techniques are available, one quickly runs into computational infeasibility problems as the number of multiplexed traffic sources increases towards more practical limits [8], [9], [13]. On the other hand, fortunately, the presence of large numbers of sources in a practical system allows us to study the asymptotic behavior of $\mathbb{P}(\{Q > x\})$ to design models that are reasonably accurate in describing systems with heavy amounts of multiplexing. In the following we discuss some of the asymptotic approximations that have been used in the design of various CAC algorithms. The approximation that forms the basis for the proposed CAC is also presented.

#### A. Effective Capacity Approximation

The effective capacity approximation is based on large deviation theory to study the asymptotic behavior of $\mathbb{P}(\{Q > x\})$ [5], [6]. In such approximations, the following asymptotic relationship has been used with considerable generality:

$$\lim_{x \to \infty} \frac{\ln(\mathbb{P}(\{Q > x\}))}{x} = -\eta,$$  \hspace{1cm} (1)

where $\eta$ is a positive constant typically referred to as the asymptotic decay rate. The following is a more general form of the effective capacity approximation:

$$\mathbb{P}(\{Q > x\}) \sim e^{-\eta x},$$  \hspace{1cm} (2)

where $f(x) \sim g(x)$ means that $\frac{f(x)}{g(x)} \to 1$ as $x \to \infty$.

The effective capacity approximation is appealing because the key parameter of interest, $\eta$, is typically easy to obtain, both exactly as well as in an approximate form. Furthermore, one of the main reasons why Equation (2) has been popular in the literature is that it provides an easy way of allocating capacity independent of the number of sources being multiplexed [5], [14], [15], [4], [7], [6], [16], [17]. However, since the effective capacities add, statistical multiplexing gains are not exploited in this scheme.

In the following subsection, we discuss a more general exponential approximation that also takes into account the effect of statistical multiplexing.

#### B. Generalized Exponential Approximation

A stronger form of asymptotics has been developed for diverse classes of queuing models. These asymptotics show that the effective capacity approximation given by Equation (2) can be significantly strengthened to obtain the following generalized exponential approximation (e.g., see [8], [9]):

$$\mathbb{P}(\{Q > x\}) \sim \alpha e^{-\eta x},$$  \hspace{1cm} (3)

where $\alpha$ is a positive constant called the asymptotic constant. Clearly, the effective capacity approximation Equation (2) is a special case of Equation (3) where $\alpha = 1$.

In the generalized exponential approximation, the effect of statistical multiplexing is captured by the asymptotic constant [8]. However, unlike the asymptotic decay rate $\eta$, the exact value of the asymptotic constant $\alpha$ is difficult to compute [9], [8], [13]. For only a few classes of queuing models, $\alpha$ can be computed exactly through numerical algorithms. However, as the number of traffic sources increases, the computational complexity and the numerical error of these algorithms increase very rapidly; hence, it quickly becomes infeasible to compute the value of the asymptotic constant $\alpha$ with any degree of accuracy. In fact, in most practical

### Table 1: Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>Link Rate</td>
<td>Cells/Sec.</td>
</tr>
<tr>
<td>$B$</td>
<td>Buffer Size</td>
<td>Cells</td>
</tr>
<tr>
<td>$P_C$</td>
<td>Peak Cell Rate (PCR)</td>
<td>Cells/Sec.</td>
</tr>
<tr>
<td>$S_C$</td>
<td>Sustained Cell Rate (SCR)</td>
<td>Cells/Sec.</td>
</tr>
<tr>
<td>$M$</td>
<td>Mean Burst Size (MBS)</td>
<td>Cells</td>
</tr>
<tr>
<td>$L$</td>
<td>Cell Loss Ratio (CLR)</td>
<td>-</td>
</tr>
<tr>
<td>$B_i$</td>
<td>Link-to-Peak Rate Ratio</td>
<td>-</td>
</tr>
<tr>
<td>$E_C$</td>
<td>Effective Cell Rate (ECR) value</td>
<td>Cells/Sec.</td>
</tr>
<tr>
<td>$E_P$</td>
<td>ECR-to-PCR Ratio</td>
<td>-</td>
</tr>
<tr>
<td>$S$</td>
<td>Source Activity Factor</td>
<td>-</td>
</tr>
<tr>
<td>$B_L$</td>
<td>Burst Loss Ratio (BLR)</td>
<td>-</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Asymptotic Decay Rate</td>
<td>-</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Asymptotic Constant</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 1. Burst and cell scale components and different asymptotic approximations for the steady state distribution of queue length.
cases, computing the exact value of $\alpha$ is almost as difficult as computing the exact tail probability $P\{Q > x\}$ [13].

In the following, we discuss a simple approximation for the asymptotic constant that forms the basis of the proposed CAC scheme.

We start our discussion with a heuristic argument. Consider a typical steady state distribution of queue length shown in Fig. 1. The data plotted in this figure is adapted from [18] where a T1 link multiplexes 80 voice sources with PCR = 32 kbps, MBS = 11 kbits, $\approx$ 27 ATM cells and voice activity factor, $\theta = \frac{\text{SCR}}{\text{PCR}} = 0.35$. The figure demonstrates how the delay distribution can be decomposed into the sum of two independent components when a superposition of VBR sources is offered to an MPLS ATM switch. These components are the cell component quantifying cell-level congestion due to the arrival of simultaneous cells from independent streams, and the burst component quantifying burst-level congestion due to burst blocking when the sum of PCR values of sources in the ON state exceeds the link capacity.

Fig. 1 also plots the true asymptote along with the one obtained by applying the effective capacity approximation (for $\alpha = 1$) and the asymptote that corresponds to the burst level congestion. As can be seen from the Fig. 1, exemplifying a general trend, and as argued in [8], [9], [13], the effective capacity approximation is too crude to give a reasonable estimate of the true distribution. As can be seen from the figure, the exact steady state distribution of queue length can be approximated by the maximum of the cell and burst level distributions [18]. This is because the burst level component dictates the queue length distribution at large buffer values, and hence the asymptote corresponding to the burst level congestion gives a reasonable approximation at large buffer values. One can therefore approximate the asymptotic constant $\alpha$ by the probability of burst blocking [8], [9]. That is, the generalized exponential approximation given in Equation (3) can be further approximated as:

$$P\{Q > x\} \approx \beta e^{-\eta x}, \quad \eta \geq 0,$$

where $\beta$ denotes the probability of burst blocking and the symbol $\approx$ represents approximate equality for large values of $x$. In addition to the burst level congestion argument, the heuristic is also supported by the fact that it is known to be exact for the GI/M/s queue as well as the GI/GI/1 queue under a heavy traffic assumption [9].

The approximation given in Equation (4) forms the basis of the proposed CAC. Note that probability of burst blocking, $\beta$, is relatively easier to compute than a more accurate value for the asymptotic constant, $\alpha$. This is also the strength of the simple approximation given in Equation (4).

### III. The Proposed CAC

Consider the multiplexing of traffic generated by a set of $N$ independent sources that are heterogeneous in their traffic characteristics onto a channel of capacity $C$ equipped with a buffer of size $B$. Each source $i$ is assumed to alternate asynchronously between ON and OFF states with exponentially distributed periods with mean lengths $\frac{1}{\lambda_i}$ and $\frac{1}{\mu_i}$, respectively.

Let the ATM traffic descriptor for the source $i$ be defined in terms of parameters $P_i, S_i, M_i$, and $L$, denoting PCR, SCR, MBS attributes, and CLR requirement of the source, respectively. Further assume that $\eta_0$ denotes the dominant (largest negative) eigenvalue of the system. We can write, for large values of $x$, the approximation given in Equation (4) presented in Section II in terms of $\eta_0$ as follows [13]:

$$\mathcal{L} \approx \beta e^{\eta_0 x}, \quad \eta_0 \leq 0.$$  \hspace{1cm} (5)

As before, in the above expression we approximate the CLR objective for the finite buffer system of buffer size $B$ by the steady-state probability that the queue length exceeds $B$ value in an infinite buffer system, i.e., $\mathcal{L} \approx P\{Q > B\}$. Furthermore, as $\beta$ represents probability of blocking in a bufferless system, we have $\mathcal{L} \leq \beta$, which guarantees that $\eta_0 \leq 0$ in Equation (5).

Let $C^N$ denote the aggregate capacity usage of the set of $N$ heterogeneous sources considered here. Kosten [19] provides a simple relation between the dominant eigenvalue $\eta_0$ and the aggregate capacity usage of the multiplexed stream:

$$C^N \approx \frac{1}{2\eta_0} \sum_{i=1}^{N} \left( P_i \eta_0 + \mu_i + \lambda_i - \sqrt{(P_i \eta_0 + \mu_i - \lambda_i)^2 + 4\lambda_i \mu_i} \right)$$ \hspace{1cm} (6)

By combining Equations (5) and (6) and performing some parameter conversion, we can rewrite the above equation in terms of parameters contained in the ATM traffic descriptor as follows:

$$C^N \approx \sum_{i=1}^{N} P_i \cdot \frac{\xi_i - \frac{\sqrt{\xi_i^2 + 4\xi_i}}{2(\xi_i - \frac{\sqrt{\xi_i^2 + 4\xi_i}}{2})}}{M_i}$$ \hspace{1cm} (7)

where, $\xi_i = \frac{S_i}{P_i}$, $\eta_i = \frac{\theta_i}{1 - \theta_i}$, and $\xi_i = 1 + q_i + M_i \cdot \ln\left(\frac{C^N}{M_i}\right)$. Note that $M_i$ indicates the mean burst size, which satisfies the following relationships:

$$\mu_i = \frac{P_i}{M_i}, \quad \lambda_i = q_i \cdot \frac{P_i}{M_i}, \quad \lambda_i + \mu_i = \frac{P_i}{M_i}.$$  \hspace{1cm} (8)

Note that in Equation (7), the term inside the summation is the effective capacity of a single source in isolation. This suggests that if a CAC algorithm works based on Equation (7), we can assume that the effective capacity of a single connection is given by the following expression:

$$\xi_i \approx \frac{P_i \cdot \left(\xi_i - \frac{\sqrt{\xi_i^2 + 4\xi_i}}{2(\xi_i - \frac{\sqrt{\xi_i^2 + 4\xi_i}}{2})}\right)}{2(\xi_i - \frac{\sqrt{\xi_i^2 + 4\xi_i}}{2})}.$$  \hspace{1cm} (9)

Note that Equation (8) has the same structure as the one obtained by Gibbens and Hunt [5]. However, unlike the Gibbens and Hunt formula, which does not contain a term for the link capacity, the proposed expression (Equation (8)) takes the effect of statistical multiplexing through the blocking parameter $\beta_i$. Computation of an ECR value using Equation (8) requires us to compute the probability of burst blocking for source $i$, $\beta_i$. In the following section, we therefore first discuss burst blocking in a heterogeneous setup.
A. Burst Blocking in a Heterogeneous Setup

Since publication of the famous Erlang-loss formula for blocking in telephone networks of the early days (homogeneous environment), blocking in telecommunication networks has been a subject of extensive research. Later advancements in telecommunication systems, such as support of multiple rate connections in ISDN, slot aggregation technique for wireless and wired networks, and emergence of packet switching technology, catalyzed the study of blocking in an environment where sources may differ in their bandwidth requirements as well as statistical characteristics [20]. Reference [21] provides an excellent overview of the prior work on this subject.

Kaufman [22] and Roberts [23] provide the well-known recursions for the computation of blocking in a heterogeneous environment. Although their formulation yields an exact solution, it is not viable for computation in real-time. Elwalid and Mitra [21], on the other hand, present an approximate solution, it is not viable for computation in real-time. Elwalid recursions for the computation of blocking in a heterogeneous environment. Although their formulation yields an exact solution, it is not viable for computation in real-time. Elwalid and Mitra [21], on the other hand, present an approximate solution, it is not viable for computation in real-time.

The proposal is based on the Uniform Asymptotic Approximation (UAA). We find the approximations quite accurate for all practical purposes. Furthermore, our numerical investigation with the UAA leads us to its real-time adaptation discussed in the following.

B. A Real-time Adaptation of the UAA Approximation

Reference [21] specializes the UAA approximation to the overloaded, critical, and under-loaded regimes, and presents an expression for blocking probability for the individual cases. The type of scenarios a CAC deals with, fall into the category of the critical regime. In the following we present a real-time adaptation of the results presented in [21] for the computation of burst blocking probability for a source $i$, $\beta_i$, in a heterogeneous environment. By real-time adaptation we mean that the computational complexity of our CAC scheme is polynomial in the order of the number of traffic sources admitted in the switch.

Let $\mathcal{A}$ represent the set of admitted sources and $n_i$ the number of occurrences of the source $i$ in the set $\mathcal{A}$. The load offered to the system can be represented as $\sum_{i \in \mathcal{A}} n_i P_i \theta_i$. Per the UAA method, we can associate a constant $\alpha$ to a critically loaded system such that,

$$\sum_{i \in \mathcal{A}} n_i P_i \theta_i = C - \alpha \sqrt{C}$$

$$\Rightarrow \alpha = \frac{1}{\sqrt{C}} \left( C - \sum_{i \in \mathcal{A}} n_i P_i \theta_i \right) \quad (9)$$

We are interested in the burst blocking probability for source $i$, $\beta_i$. For this purpose, first define $\sigma > 0$ such that,

$$\sigma^2 = \frac{2}{\sum_{i \in \mathcal{A}} P_i \theta_i (1 - \theta_i)} \quad \epsilon_i \quad (10)$$

For a system that satisfies the loading condition specified by Equation (9), a real-time adaptation of the result presented in [21] is given as follows:

$$\beta_i = \frac{P_i}{\sigma \sqrt{\pi C}} e^{-\frac{n_i}{\sigma}} \quad (11)$$

Note that a CAC controlled system satisfies the loading condition stated in Equation (9). We, therefore, use Equation (11) to evaluate $\beta_i$ to compute an ECR value based on Equation (8). Section IV addresses the issue of correctness of Equation (11).

C. Summary of the Proposed CAC

This completes our presentation of the essential ingredient of our proposed CAC methodology. The proposed CAC is summarized in an algorithmic form in Algorithm 1.

Let $\mathcal{A}$ be the set of admitted sources. Also suppose $\hat{\mathcal{A}} = \mathcal{A} \cup \{ n \}$. In other words, set $\hat{\mathcal{A}}$ represent the projected set of admitted sources, if we admit the source $n$ under CAC test.

**Algorithm 1. The Proposed CAC Algorithm**

Let $\mathcal{A}$ be the set of admitted sources. Also suppose $\hat{\mathcal{A}} = \mathcal{A} \cup \{ n \}$. In other words, set $\hat{\mathcal{A}}$ represent the projected set of admitted sources, if we admit the source $n$ under CAC test.

if $\sum_{i \in \mathcal{A}} \epsilon_i > C$

Available capacity is less than the ECR value, reject the request.

else

Enough room is available, admit the connection.

end

end

Computational Complexity of the Proposed CAC Algorithm: $\beta_i$ and $\epsilon_i$ can each be computed in $O(1)$ time. Therefore, the computational complexity of the proposed CAC algorithm is $O(n)$, where $n = \| \hat{\mathcal{A}} \|$. The effectiveness of the CAC procedure summarized in Algorithm 1 is dictated by how well the ECR value represents the resource requirement of the source under question. In the following, we evaluate the performance of the proposed CAC scheme.

IV. Scalability Analysis and Performance Evaluation

The scalability of a CAC algorithm is a measure of its ability to boost up link utilization with increasing amounts of resources (link capacity and buffer size) and with different mixes of traffic generated by sources of diverse burstiness characteristics. That is, a scalable CAC keeps the link running at a high utilization level while meeting the QoS requirements under various resource setups and with sources of different characteristics. This section studies scalability aspects of the proposed algorithm presented in Section III and evaluates its performance with respect to some of the previously proposed schemes.

A. Performance Benchmarks

In this section, we present summary of some of the previously proposed CAC schemes. We compare our proposed scheme with these CAC approaches.
Fig. 2. Effect of link rate to PCR on various CAC algorithms: CLR = $10^{-6}$, buffer size = 500 cells (a) VBR.c with $\theta = 0.01$ (b) VBR.b (c) VBR.a (d) VBR.d with $\theta = 0.75$.

1) Gibbens-Hunt Approach: Gibbens and Hunt [5] provide the following expression for the ECR value of a VBR source:

$$\xi_i = \frac{P_i \cdot Z + \mu_i + \lambda_i - \sqrt{(P_i \cdot Z + \mu_i - \lambda_i)^2 + 4\lambda_i\mu_i}}{2 \cdot Z}$$

(12)

where, $Z = \frac{\ln(L)}{B}$.

Notice that because the Gibbens-Hunt Equation (12) is based on the effective capacity approximation discussed in Section II-A, the ECR value given by Equation (12) is independent of the link capacity, $C$.

2) Generalized Gibbens-Hunt Method: For General Markovian Sources Elwalid and Mitra [6] extend the Gibbens-Hunt results for more general Markovian sources. In their formulation a source assumes a large number of states as oppose to just ON and OFF. Even for ON-OFF fluid sources, the results shed new light on the origins of Equation (12). Specifically, consider an aggregate source for the individual sources over state space $\Omega$. Let $G$ be the irreducible generator of the aggregate source. This aggregate source generates fluid at a constant rate $\lambda_\omega$ when in state $\omega$ ($\omega \in \Omega$). Let $\lambda = \{\lambda_\omega | \omega \in \Omega\}$. The aggregate source is characterized by the set given by $(G, \lambda)$. The effective capacity of the source is the maximal real eigenvalue of the matrix $(\Lambda - \frac{B}{\ln(L)} \cdot G)$.

Here $\Lambda$ is defined as $\Lambda = \text{diag}(\lambda)$.

The generalized Gibbens-Hunt method is also based on the effective capacity approximation discussed in Section II-A. Therefore, the ECR value given by the maximal real eigenvalue of the matrix $(\Lambda - \frac{B}{\ln(L)} \cdot G)$ is independent of the channel capacity.

3) Gibbens-Hunt with Gaussian Approximations: Guerin, Ahmadi, and Naghshineh [7] independently obtained the Gibbens-Hunt formula and found that the effective capacity approximation does not take the effect of statistical multiplexing into account. Consequently, they propose a heuristic where the computation of the equivalent capacity of an individual connection is based on the effective capacity approximation, while the computation of the aggregate capacity of multiplexed connections is based on a Gaussian approximation. In other words, the CAC scheme uses Gaussian approximation to complement effective capacity approximation discussed in Section II-A to take effect of statistical multiplexing into account. Specifically, the ECR value of an individual source
Fig. 3. Admissible regions showing the effect of link rate to PCR on various CAC algorithms in a heterogeneous environment: CLR = $10^{-6}$, buffer size = 500 cells.

i, based on the effective capacity approximation, is given by:

$$E_{i}^{EC} = P_i \cdot \frac{h_i - B + \sqrt{(h_i - B)^2 + 4h_i \theta_i B}}{2h_i},$$

where, $h_i = \ln(\frac{L}{2}) \cdot M_i (1 - \theta_i) P_i$. On the other hand, reference [7] computes the ECR of the aggregated traffic of $n$ sources using the Gaussian approximations as $\sum_{i=1}^{n} E_{i}^{g} = \hat{S} + \alpha g \hat{\sigma}$. In this equation, $\alpha g = \sqrt{-2 \ln(L) - \ln(2\pi)}$, $\hat{S} = \sum_{i=1}^{n} S_i$, and $\hat{\sigma}^2 = \sum_{i=1}^{n} S_i \cdot (P_i - S_i)$. Note that the expression is only applicable to a homogeneous CLR objective, $L$.

The reference [7] uses the minimum of the effective capacity and Gaussian approximations to compute the ECR of the aggregated traffic. That is, $\sum_{i=1}^{n} E_{i} = \min (\hat{S} + \alpha g \hat{\sigma}, \sum_{i=1}^{n} E_{i}^{EC})$. Furthermore, a new connection is admitted only if the aggregate ECR value of the already admitted calls and the new connection does not exceed the link capacity.

To study scalability aspects of the various CAC schemes, we need a set of parameters that can be used to quantify scalability of a CAC scheme.

### B. Metrics of Scalability

In this section, we define some metrics that help us study scalability aspects of CAC algorithms.

1) **Link-to-Peak Rate Ratio:** The Link-to-peak rate ratio is defined as the ratio of the link rate to the PCR requirement of a connection. That is, if $C$ represents the capacity of the link and $P_i$ denotes the PCR value of source $i$, the link-to-peak rate ratio is given by $R_i = \frac{C}{P_i}$.

Clearly, the link-to-peak rate ratio represents the relative size of the link with respect to the source. The statistical multiplexing gain, of course, increases with increasing link-to-peak rate ratios. Hence, a desirable attribute of a CAC algorithm is that it should be able to scale up the link utilization with increasing values of the link-to-peak rate ratio.

2) **Source Activity Factor:** Within the framework of ATM, burstiness of a source is characterized by its source activity factor, which is defined as the ratio of the average to peak bandwidth requirement of the source. That is, if $S_i$ and $P_i$ respectively denote the SCR and PCR requirement of source $i$, its source activity factor is defined as $\theta_i = \frac{S_i}{P_i}$. A CAC algorithm is expected to yield consistent performance while dealing with sources of diverse activity levels.
3) Buffer Size to MBS Ratio: Buffers are used to absorb bursts of traffic when the instantaneous rate of arrival from the admitted connections exceeds the link rate. Of course, large buffers can help achieve better link utilization but more buffering means more end-to-end delay and delay variation. Typically, end-to-end delay and delay variation requirements dictate the buffer size. Furthermore, such a selection is made at the time of network configuration. It is therefore desirable for a CAC algorithm to scale well to numerous buffer settings tailored to diverse networking scenarios and requirements.

Let $M_i$ denote the MBS attribute of source $i$, and $B$ represent the buffer size available at the switching facility. The buffer size to MBS ratio, $\frac{B}{M_i}$, is a parameter that can be used to quantify the relative size of the buffer at a switching facility.

As expected, the achievable link utilization increases with increasing buffer sizes at the switch, in a typical setup. A scalable CAC is one that can consistently yield high link utilization with various settings of buffer to MBS ratio.

We make use of the metrics of scalability discussed above to define performance evaluation criteria.

C. Performance Evaluation Criteria

The following set of questions establishes the evaluation criteria considered in this paper:

- How close is the link utilization achieved by the CAC to the maximum achievable link utilization?
- How well does the algorithm scale with increasing sizes of the link and switch buffers, and how well does it accommodate sources of diverse burstiness characteristics?
- Does the algorithm, under different loading scenarios, guarantee the QoS requirements for all connections?

The achievable link utilization for a CAC algorithm is defined here as the ratio of the sum of the SCR values of the admitted sources to the link capacity. That is, if $A$ represents the largest set of admitted sources, the link utilization $\eta$ is given by $\eta = \sum_{i \in A} S_i \frac{1}{C}$. Clearly, to study the scalability and performance of a CAC scheme, a characterization of the highest possible link utilization surface is required. Buffet and Duffield [24] provide an upper bound on the cell loss ratio, in other words a lower bound on the link utilization for homogeneous sources. Hence, in evaluating the performance of the proposed CAC, we also consider the case of homogeneous sources.
along with the heterogeneous sources is two fold. First, for homogeneous sources, the highest possible utilization surface is well characterized by the Buffet-Duffield bound [24], which can be used to provide a performance reference. Second, the effect of various parameters on the admissible region is better represented in a homogeneous setup than in its heterogeneous counterpart. An outline of Buffet-Duffield bound is given below.

**D. Buffet-Duffield Bound**

Buffet and Duffield [24] propose a lower bound on the achievable link utilization for homogeneous sources. Unfortunately, the formulation is prohibitively expensive for real-time realization. Nonetheless, the bound can be used as a performance target for other viable CAC algorithms. Specifically, any CAC scheme is guaranteed to satisfy the QoS requirement as long as it does not take the link utilization below what is permitted by the bound on utilization. We also perform simulation studies to evaluate the tightness of the bound.

The type of buffered facility and sources considered in this paper are the same as considered by Buffet-Duffield. Their results provide an upper bound on the CLR for a given loading of the system. Specifically, if \( n \) denotes the number of sources filling-up the buffer, the CLR of the system is bounded according to the following expression:

\[
L_i \leq \lambda_i \mu_i \chi_i \nu_i \\
\lambda_i = \frac{\omega_i}{\theta_i (1 - \omega_i)} \\
\omega_i = \frac{C}{n \cdot P_i \cdot \left( 1 + \frac{n \cdot S_i - C}{M_i \cdot C(1 - \theta_i)} \right)} \\
\mu_i = \frac{\omega_i (n \cdot P_i - C)}{C(1 - \omega_i)} \\
\chi_i = \frac{n \cdot P_i (1 - \theta_i)}{n \cdot P_i - C} \\
\nu_i = \frac{\theta_i (M_i \cdot P_i - C)}{C(1 - \theta_i)}
\]

**E. Simulation Environment**

In addition to the Buffet-Duffield bound, we also run simulations for different resource environments and source types considered in this section. For the purpose of traffic
characterization, several measurement studies have been conducted. Reference [25] summarizes some of these studies in the form of Table 2. In this table, VBR.a, VBR.b, and VBR.c are Bellcore traffic models whereas VBR.d represents TCP/IP Internet traffic. We use all these traffic models for the purpose of performance evaluation. However, only a selected subset of the results is presented to illustrate relevant trends.

This completes the description of the essential ingredients of our numerical investigation. The following subsections shed some light on scalability aspects of various CAC algorithms, including our proposed scheme.

**F. Effect of Link-to-Peak Rate Ratio**

In this section, we investigate the effect of the link-to-peak rate ratio on the various CAC methods. We first discuss the results for a homogeneous environment as they demonstrate the effect of changing link size more transparently than in a heterogeneous setup.

Fig. 2 illustrates how different CAC algorithms scale with the increasing size of the link. The lowest possible link utilization surfaces characterized by the Buffet-Duffield bound and simulation results are also presented. Results are plotted for four different source activity factors.

The Buffet-Duffield bound and simulation results indicate an increasing trend in the link utilization with the increasing size of the link with respect to the sources. However, the Gibbens-Hunt method fails to account for the statistical multiplexing gain that comes with increasing value of link-to-peak ratio. The same is true of the generalized Gibbens-Hunt approach discussed in Section IV-A2. Furthermore, both algorithms run the link at an unacceptably low utilization level while dealing with the more bursty sources. (See, e.g., Fig. 2(a)-(c). These results are consistent with our discussion on limitations of the effective capacity approximations. At high source activity factors (e.g., $\theta = 0.75$ or better), all algorithms perform equally well.

Fig. 2 also illustrates that the advantage of using a Gaussian approximation to the effective capacity approximation.
However, the advantage can only be enjoyed at large link-to-peak rate ratios. For example, in Fig. 2(a), the Gaussian approximation kicks in for a link-to-peak rate ratio of 40 or greater. It can also be seen that even this improvement is not very significant. In this example, a deviation of as high as 23% from the Buffet-Duffield bound is observed.

The performance achieved by the proposed CAC is also plotted in Fig. 2. It can easily be seen that the proposed CAC algorithm outperforms the other CAC methods and runs the link very close to the safety region dictated by the Buffet-Duffield bound. It can also be seen from the figure that the proposed CAC yields consistent results for highly bursty as well as smooth sources.

Fig. 3 complements the above discussion for a heterogeneous environment where VBR.a and VBR.b type of sources share the same facility. The figure plots admissible regions for various CAC algorithms at different link-to-peak rate ratios. The figure also shows simulation results. In a heterogeneous environment, the link-to-peak rate ratio is defined as the ratio of the link rate to the average PCR value of the sources, i.e., in this example, \( R = \frac{C}{\sum_{i=1}^{2} \frac{C_i}{P_i}} \). Again, it can be observed from the results that the Gibbens-Hunt approach and its generalization does not scale up the performance with the increasing size of the link. Gibbens-Hunt with Gaussian approximation performs better but the proposed CAC consistently outperforms the rest.

G. Effect of Source Activity

In this section we study the effect of the changing source activity factor on the performance of various CAC algorithms. Again, results for both homogeneous as well as heterogeneous environments are presented.

For different link-to-peak rate ratios, Fig. 4 plots utilization curves for the various CAC schemes against the source activity factor. The figure also shows the Buffet-Duffield bound on the link utilization.

The results presented in Fig. 4 indicate that the Gibbens-Hunt and generalized Gibbens-Hunt methods run the link at very low utilization values for sources that are highly bursty in nature. Furthermore, for sources with low activity factor, the performance is unacceptable at a high link-to-peak rate ratio. For example, for a source with source activity of 0.1 and a link-to-peak rate ratio of 100, the Gibbens-Hunt method runs the link at 30%, whereas a utilization of at least 67% is possible. However, with more smooth sources (e.g., \( \theta \geq 0.6 \)),
all algorithms run the link at around the highest possible link utilization level.

Fig. 4 also shows that at low values of the source activity factor, the Gaussian approximation kicks in to cover-up the slackness in the effective capacity approximation. However, it still fails to run the link at high utilization level. A maximum deviation of around 13 to 15% from Buffet-Duffield bound is observed in most of the settings presented in this figure.

As can be observed from Fig. 4, the utilization achieved by the proposed CAC never goes below the Buffet-Duffield bound on utilization, which means that, for the example setup considered in this figure, QoS is guaranteed. In other words, at various source activity factors, the proposed CAC scales its performance to the highest possible level without violating the QoS contract.

Fig. 5 complements the above discussion for a heterogeneous environment where VBR.a sources of highly diverse characteristics share the same facility. The figure plots the admissible regions for various CAC algorithms and simulation results, for different settings. Again, it can be deduced from the figure that the Gibbens-Hunt approach and the generalized Gibbens-Hunt method do not scale up the performance based on the mix of the traffic offered to the system. The Gibbens-Hunt with Gaussian approximation also fails to run the link at a reasonably higher utilization level. However, the proposed CAC scheme is superior in terms of its ability to operate the link at a high utilization level, with various traffic mix. Furthermore, the utilization achieved by the proposed CAC never goes below the Buffet-Duffield bound, which means that the proposed CAC guarantees the QoS.

H. Effect of Buffer Size

In Section IV-B, we define buffer size to MBS ratio as one of the metrics to measure CAC scalability. This section presents scalability analysis of various CAC algorithms with respect to this parameter.

Figs. 6 and 7 are intended to evaluate how well various CAC methods accommodate the changing size of the buffer. These figures show the achievable link utilization for the four methods considered in this paper as the buffer size to MBS ratio is varied from 2 to 80. The Buffet-Duffield theoretical bound on the link utilization along with simulation results are also presented to serve as a performance reference.

It can easily be seen from Fig. 6 that Gibbens-Hunt and generalized Gibbens-Hunt perform poorly with low buffers (e.g., for $\frac{B}{M} \leq 40$). This is true of the various link-to-peak rate ratios considered in the figure. For example, at a buffer size to MBS ratio of 10, in Fig. 6(d) a utilization of 37% is observed by these algorithms whereas the link can be operated at 68% without violating the CLR constraint. Gibbens-Hunt with Gaussian approximation improves the situation but such improvements can only be enjoyed either for systems with a very small buffer (e.g., $\frac{B}{M} \leq 10$) or at facilities with large buffers (e.g., $\frac{B}{M} \geq 40$). However, it behaves poorly in between, which is a more likely region of operation. The proposed CAC, on the other hand, consistently yields a close to the highest possible utilization for the entire range of the buffer size to MBS values. Furthermore, the utilization achieved by the proposed CAC never goes below the Buffet-Duffield lower bound on utilization, which means that for the example setup considered in this figure, QoS is guaranteed.

Fig. 7 presents the effect of the buffer size to MBS ratio in a heterogeneous environment where VBR.a and VBR.b type of sources share the same facility. The results for the heterogeneous case are consistent with their homogeneous counterpart. That is, the proposed CAC scheme scales very well with increasing buffer sizes and runs the link at a higher utilization level without violating the QoS constraints.

V. Conclusion

In this paper, we first describe certain scalability requirements of CAC algorithms. We then propose a computationally efficient and scalable CAC algorithm. The algorithm applies to an MPLS ATM switch with a FIFO buffer. The switch carries data from statistically independent VBR sources that asynchronously alternate between ON and OFF states with exponentially distributed periods. The sources may be heterogeneous both in terms of their statistical characteristics as well as their QoS requirements.

For the purpose of performance comparison, we conducted scalability analyses for various CAC schemes. Specifically, we considered the Gibbens-Hunt approach, the generalized Gibbens-Hunt method, and Gibbens-Hunt with Gaussian approximation. These algorithms are based on the effective capacity approximation. Our numerical investigation complements the results obtained in [8] that the ECR computation algorithms that are based on the effective capacity approximation do not scale well with the increasing link-to-peak rate ratios. Furthermore, these algorithms yield poor performance when dealing with sources that are highly bursty in nature, or scenarios with small buffers at the switching facility. The effective capacity approach with Gaussian approximation is found to provide some improvements. However, the results show that it fails to consistently operate the link at a level close to the highest possible link utilization mark. Moreover, we noticed that the improvements are only limited to certain regimes of operation, e.g., in a setup with large link-to-peak rate ratios and/or for the case of more smooth sources and/or in an environment where large buffers are setup at the switching facility.

We discuss the generalized exponential approximations which take the effect of statistical multiplexing into account. However, the main limitation of the generalized exponential approximations is their computational complexity which makes them impractical for real-time applications. We address this limitation by proposing a simple but effective approximation for the asymptotic constant. Consequently, expressions for effective capacity, that takes the effect of statistical multiplexing into account, are provided for both homogeneous as well as heterogeneous environments.

We evaluate the performance of the proposed scheme under numerous resource setups and in homogeneous as well as heterogeneous source environments. We show that the proposed CAC scales well with increasing amounts of resources.
(link capacity as well as switch buffers) and accommodates intelligently the mix of traffic offered by sources of diverse burstiness characteristics. For the purpose of establishing this fact, we compared the performance of the proposed CAC using the Buffet-Duffield bound [24]. Our results indicate that the proposed scheme consistently runs the link at a utilization level very close to the safety region dictated by the theoretical bound. Furthermore, utilization achieved by the proposed CAC never goes below the Buffet-Duffield bound on utilization, which means that the proposed CAC guarantees the QoS. Simulation results also complement these facts for both homogeneous as well as heterogeneous environments.

REFERENCES


Zafar Ali

Zafar Ali is a Technical Leader at Cisco Systems, Inc. where he leads software protocol development for Cisco Systems high speed, carrier-class routing business unit. Zafar Ali has authored a number of IETF drafts. Zafar has a broad range of interests include optical networking, MPLS Traffic Engineering and IP routing. Zafar Ali received his Ph.D. and MS in Electrical and Computer Engineering from Purdue University, West Lafayette, Indiana. He obtained his Bachelor of Engineering in Electrical Engineering from NED University, Karachi, Pakistan where he was awarded the University Gold Medal.

Waseem Sheikh

Waseem Sheikh is a Ph.D. candidate in the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. He received his B.S. in Electronics Engineering from Ghuilam Ishaq Khan Institute of Engineering and Technology, Pakistan, in 2000; and M.S.degree in Electrical and Computer Engineering from Purdue University, in 2002. His research interests include optimization techniques in computer networks.

Edwin K. P. Chong

Edwin Chong received the B.E.(Hons.) degree with First Class Honors from the University of Adelaide, South Australia; and the M.A. and Ph.D. degrees from Princeton University. He joined Purdue University in 1991, where he was named a University Faculty Scholar in 1999, and was promoted to Professor in 2001. Since August 2001, he has been a Professor of Electrical and Computer Engineering and of Mathematics at Colorado State University. His interests are in networks and optimization. He is a Fellow of the IEEE, and served as a Control Systems Society Distinguished Lecturer. He is currently an editor for Computer Networks.

Arif Ghafoor

Arif Ghafoor is currently a Professor in the School of Electrical and Computer Engineering, at Purdue University, West Lafayette, IN, and is the Director of Distributed Multimedia Systems Laboratory. He has been actively engaged in research areas related to database security, parallel and distributed computing, and multimedia information systems. He has served on the editorial boards of various journals. He is a Fellow of the IEEE. He has received the IEEE Computer Society 2000 Technical Achievement Award for his research contributions in the area of multimedia systems.