ELSEVIER

Computer Communications 25 (2002) 1172–1184

computer communications

www.elsevier.com/locate/comcom

A cell loss upper bound for heterogeneous ON–OFF sources with application to connection admission control

Guoqiang Mao*, Daryoush Habibi

School of Engineering and Mathematics, Edith Cowan University, 100 Joondalup Dr., Joondalup, Western Australia 6027, Australia

Received 10 August 2001; accepted 28 August 2001

Abstract

Bufferless fluid flow model (bffm) is widely used in the literature for cell loss analysis. In this paper, we propose an efficient and effective means of investigating cell loss in the bffm. We define the cell loss rate function (clrf) and use it to characterize the cell loss of traffic sources in the bffm. Properties of clrf are discussed. These properties enable us to decompose the complex analysis of the multiplexing of several traffic sources into the simpler analysis of the individual sources. A cell loss upper bound for heterogeneous ON–OFF sources is proposed using clrf. The proposed cell loss upper bound is tighter than that proposed in previous literature. A connection admission control (CAC) scheme using online measurements is designed based on the cell loss upper bound. Extensive simulations are carried out to study the performance of the CAC scheme. Simulation results indicate that the proposed CAC scheme is time-efficient, can ensure QoS guarantee, and is capable of achieving high network utilization.

Keywords: ATM; QoS; ON–OFF sources; Bufferless fluid flow model

1. Introduction

To date, many connection admission control (CAC) schemes have been proposed [1–8]. These schemes can be classified into two categories: traffic descriptor-based CAC and measurement-based CAC. Traffic descriptor based CAC uses the a priori traffic characterizations provided by sources at connection setup phase to calculate the probabilistic behavior of all existing connections in addition to the incoming one. It achieves high network utilization when traffic descriptions required by the CAC scheme are tight. Measurement-based CAC uses the a priori traffic characterizations only for the incoming connection and uses measurements to characterize existing connections. Under measurement-based CAC scheme, network utilization does not suffer significantly if traffic descriptions are not tight. However, because of the fact that source behavior is not static in general, it is difficult for measurement-based CAC to obtain traffic characteristics accurately from on-line measurements. Measurement-based CAC can only deliver significant gain in utilization when there is a high degree of statistical multiplexing [1].

Considering the difficulty for both traffic descriptor-based and measurement-based CAC to obtain accurate traffic characteristics, the performance of a CAC scheme should not be measured only by the utilization achieved under ideal circumstances where traffic sources are all tightly characterized. Also, one must consider whether enough accurate traffic characteristics can be obtained from sources and/or network practically, and, the robustness of the CAC scheme against the inaccuracies in those traffic characteristics. In addition to high network utilization an ideal CAC scheme should satisfy the following requirements [9,10]:

- Simplicity: the scheme must be both economically implementable and fast. The traffic characteristics required by the CAC scheme should be easily and reliably obtained from the traffic sources and/or network.
- Flexibility: the scheme must not only be able to satisfy the current needs of network services but also be able to adapt to new services which are likely to evolve.
- Robustness: the scheme must be able to handle imperfect assumptions.

Cell loss and cell delay are often adopted as measures of QoS. Cell delay can usually be controlled within a desired bound by engineering the buffer size, hence cell loss is used in most papers as the QoS index.

* Corresponding author.
E-mail address: g.mao@ieee.org (G. Mao).

0140-3664/02/$ - see front matter
PII: S0140-3664(01)00439-X

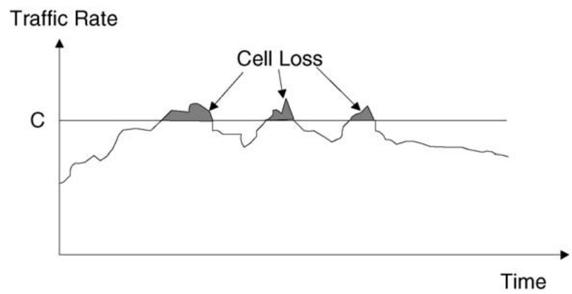

Fig. 1. Cell loss in bufferless fluid flow model.

In this paper, based on the bufferless fluid flow model (bffm) we first define the cell loss rate function (clrf) and use it to study the cell loss in bffm. The properties of clrf are introduced. The conjecture proposed by Rasmussen et al. [11] about the cell loss of ON–OFF sources is proved. Furthermore, we propose a tighter cell loss upper bound than that proposed in previous literature for heterogeneous ON–OFF sources. The upper bound is applied to design a measurement-based CAC scheme. In the design of the CAC scheme, the above guiding principles for CAC schemes are addressed with particular attention given to the impact of inaccuracies in declared traffic parameters.

The rest of the paper is organized as follows: in Section 2 the clrf is defined and its properties are introduced; in Section 3 heterogeneous ON–OFF sources in bffm are studied and the upper bound is developed; the CAC scheme is designed in Section 4; Section 5 presents some simulations using exponential ON–OFF sources; in Section 6 video sources are used to test the performance of our CAC scheme; and finally conclusions are given in Section 7.

2. Evaluating cell loss in the bufferless fluid flow model—cell loss rate function

There are mainly three methods to analyze the cell loss in the network: queueing theory, large deviation approximation and bufferless fluid flow approximation. Queueing theory provides exact analysis of cell loss in the network but its analysis is too complex, hence assumptions have to be made to simplify the analysis. In many studies, the queueing theory depends heavily on the assumption of Poisson processes for cell loss analysis.

Large deviation theory generates approximation for cell loss. In large deviation theory the loss probability P_B in a finite buffer queueing system with buffer size B is approximated by $P(Q > B)$, the tail of the queue length distribution in the corresponding infinite buffer queueing system. For Markovian traffic in the infinite buffer queueing systems it has been shown that $P(Q > B)$ is asymptotically exponential, i.e.

$$P(Q > B) \sim A e^{-\eta B} \quad \text{as} \quad B \to \infty, \tag{1}$$

where η is a positive constant called the *asymptotic decay rate*, A is a positive constant called the *asymptotic constant*, and $f(x) \sim g(x)$ means $\lim_{x\to\infty} f(x)/g(x) = 1$. Eq. (1) applies only to large buffers and therefore limits the usage of large deviation theory method. Effective bandwidth approach assumes that $A = 1$, which is criticized as too conservative.

Bufferless fluid flow model is also used in many literature to analyze cell loss in a network [7,12,13]. Since this approach assumes that there is no buffer in the network, it generates conservative estimate of cell loss. However, the simplicity of bffm enables us to concentrate on the characteristics of traffic sources themselves. In this paper, we use bffm to analyze the cell loss.

Under bffm, cell loss due to overflow occurs if and only if the sum of the traffic rates of all active connections, denoted by X, exceeds the link capacity C. Fig. 1 illustrates the cell loss in bffm.

Let us define a function $F(y)$ as follows:

$$F(y) \triangleq E[(X - y)^+]. \tag{2}$$

We call $F(y)$ the cell loss rate function of X. Our study shows that clrf is an effective tool for studying cell loss in bffm. In this paper, since we are discussing traffic sources, all random variables are non-negative random variables. The clrf has many attractive features, which facilitates the analysis of cell loss in the bffm. For example, $F(C)$ denotes the cell loss rate of a traffic source X on a link with link capacity C. Traffic sources with similar clrf can be regarded as equivalent in the cell loss analysis. From the definition of clrf it can be shown that:

$$F(y) = F(0) - y \qquad \text{for } y < 0 \tag{3}$$

and

$$F(0) = E(X). \tag{4}$$

The cell loss ratio (clr) can be calculated as:

$$\text{clr} = \frac{F(C)}{F(0)}. \tag{5}$$

An important property of the clrf is given as follows.

Property 1. *If $f(x)$ and $g(x)$ denote the traffic density distribution of independent traffic sources X_1 and X_2 respectively then the cell loss rate function of $X_1 + X_2$ equals $F * g(x)$, where $F(x)$ is the clrf of X_1, $*$ denotes convolution.*

Refer to Ref. [14] for a proof.

Definition 1. Given two traffic sources X and Y, we say that X is smaller than Y with respect to the cell loss rate function, written as $X <_{\text{clrf}} Y$, if for the clrf of X and Y, denoted by $F(y)$ and $G(y)$,

$$F(y) \le G(y) \tag{6}$$

holds for any $y \in R$.

$X <_{\text{clrf}} Y$ implies that the cell loss of traffic source X on a link is smaller than or equal to that of Y. If in addition to $X <_{\text{clrf}} Y$, we also have $E(X) = E(Y)$, then not only the cell loss of traffic source X on the link is smaller than or equal to that of Y, but also the clr of traffic source X on the link is smaller than or equal to that of Y. From the definition given above, it can be shown that:

Property 2. *If $X <_{\text{clrf}} Y$ and $Y <_{\text{clrf}} Z$, then $X <_{\text{clrf}} Z$.*

Propterty 3. *If $X <_{\text{clrf}} Y$, then $X + Z <_{\text{clrf}} Y + Z$.*

Proof. Noting that the clrf of a traffic source is always non-negative, the proof is straightforward using Property 1. □

Using the above properties, it can be shown that for traffic sources X_i and Y_i, $i = 1,\dots,n$, $X_i <_{\text{clrf}} Y_i$ implies that $\sum_{i=1}^{n} X_i <_{\text{clrf}} \sum_{i=1}^{n} Y_i$. Another property of the clrf is given in the following:

Property 4. *If $X <_{\text{clrf}} Y$ and $E(X) = E(Y)$, then $\sigma(X) \le \sigma(Y)$, where $\sigma(\cdot)$ denotes the standard deviation of random variable $\cdot$*

Proof. It can be shown that for non-negative random variable X:

$$E(X^2) = \int_0^\infty x^2 \mathrm{d}F_\mathrm{d}(x) = \int_0^\infty 2x\left[\int_x^\infty \mathrm{d}F_\mathrm{d}(y)\right]\mathrm{d}x$$
$$= \int_0^\infty \left(\int_0^x 2\,\mathrm{d}y\right)[1 - F_\mathrm{d}(x)]\mathrm{d}x$$
$$= 2\int_0^\infty \left[\int_y^\infty [1 - F_\mathrm{d}(x)]\mathrm{d}x\right]\mathrm{d}y.$$

In the above equations, $F_\mathrm{d}(x)$ denotes the distribution function of X. Then noting that the clrf of X, denoted by $F(y)$, has:

$$F(y) = E[(X - y)^+] = \int_y^\infty (1 - F_\mathrm{d}(x))\mathrm{d}x.$$

Therefore from $X <_{\text{clrf}} Y$, it can be shown that

$$\int_0^\infty \left[\int_y^\infty [1 - F_\mathrm{d}(x)]\mathrm{d}x\right]\mathrm{d}y \le \int_0^\infty \left[\int_y^\infty [1 - G_\mathrm{d}(x)]\mathrm{d}x\right]\mathrm{d}y,$$

where $G_\mathrm{d}(x)$ denotes the distribution function of Y. So it can be concluded that $E(X^2) \le E(Y^2)$. Then the proof is completed by using the relationship:

$$\sigma^2(X) = E(X^2) - [E(X)]^2. \square$$

This property of the clrf will be used in the traffic measurement analysis shown later.

3. Heterogeneous ON–OFF sources in the bufferless fluid flow model

For simplicity, let us consider ON–OFF traffic sources. According to the probability density distribution of ON and OFF periods. ON–OFF source models can be further classified into exponential ON–OFF source model, periodic ON–OFF source model, Pareto ON–OFF source model, etc. They are widely used for cell loss analysis in CAC schemes [7,11,12,15–18]. ON–OFF source model, have been successfully used to characterize the ON/OFF nature of an individual source or source element, like packetized voice and video [15,19,20]. They provide the worst case analysis of traffic sources in terms of cell loss. Recent studies indicate that ON–OFF source models are also suitable for modeling self-similar traffic [21,22].

Many current modeling techniques model each traffic source or source element by an ON–OFF source. These techniques fail to apply in high speed networks, simply because of the exploded input state space when a large number of diverse sources are multiplexed on each link. Efforts have been made to reduce the input state space. Rasmussen et al. [11] proposed an upper bound for heterogeneous ON–OFF sources. Based on a conjecture they propose that, for n heterogeneous ON–OFF sources which have peak cell rates $\text{pcr}_1,\dots,\text{pcr}_n$, and, the sum of their mean cell rates is S, then the case of n homogeneous ON–OFF sources, each source with peak cell rate $\text{pcr} = \max\{\text{pcr}_1,\dots,\text{pcr}_n\}$ and mean cell rate $\text{mcr} = S/n$, will constitute the cell loss upper bound of the n heterogeneous sources. His approach results in great reduction in input state space, hence great computational savings. However, if sources with large bandwidth demands and sources with small bandwidth demands are multiplexed, the upper bound will be too conservative. Hwang et al. [23] proposed a method of input state space reduction. In their approach, 2-state Markov Chains are built to statistically match with the power spectrum function and probability density function (pdf) of the aggregate traffic. The state space of each traffic type on the link can be reduced to 30 using their method. This is still too large to implement practically.

Lee et al. [12] proposed an algorithm which is suitable for real-time estimation of cell loss of the multiplexing of heterogeneous ON–OFF sources. However, their approach can only be applied to traffic descriptor-based CAC, hence tight characterization of mean cell rate (*mcr*) in traffic descriptors is necessary in this approach. In real networks, it is very difficult for all traffic sources to tightly characterize their mcr, so their approach only has limited use. Moreover, choosing the quantization unit remains a problem in their approach. Small quantization unit will result in huge computation efforts and large quantization unit will result in too conservative cell loss estimate.

In this section, we try to solve these problems in a different approach. Using clrf defined in Section 2, we develop a cell loss upper bound suitable for real-time calculation for heterogeneous ON–OFF sources. The upper bound is then applied to design a measurement-based CAC.

An ON–OFF source generates cells at a peak cell rate (*pcr*) denoted by pcr in active periods. In idle periods no cells are generated. Let mcr denote the mean cell rate of the ON–OFF source. The activity parameter of an ON–OFF source is defined as the ratio of mcr to pcr:

$$p \triangleq \frac{\mathrm{mcr}}{\mathrm{pcr}} \tag{7}$$

Then the probability that an ON–OFF source is active or idle is given by p or $1-p$, respectively.

Assume there are n independent ON–OFF sources $X_1, \ldots, X_n$ on the link, where $X_{i(i=1,\ldots,n)}$ has peak cell rate pcr_i, mean cell rate mcr_i and activity parameter $p_i = \mathrm{mcr}_i/\mathrm{pcr}_i$. The probability mass function (pmf) of $X_{i(i=1,\ldots,n)}$ can be expressed as:

$$f_{1,(\mathrm{pcr}_i)}^{(p_i)}(x) = \begin{cases} p_i & x = \mathrm{pcr}_i \\ 1-p_i & x = 0 \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

and the pmf of $\sum_{i=1}^{n} X_i$ can be expressed as:

$$f_{n,(\mathrm{pcr}_1,\ldots,\mathrm{pcr}_n)}^{(p_1,\ldots,p_n)}(x) = f_{1,(\mathrm{pcr}_1)}^{(p_1)} * \cdots * f_{1,(\mathrm{pcr}_n)}^{(p_n)}(x). \tag{9}$$

In this paper we use subscript n, subscript (pcr) and superscript (p) to denote the number of ON–OFF sources, pcr of sources and their activity parameters, respectively, when we need to emphasize the dependence of one function on these parameters.

Since an ON–OFF source or the multiplexing of several ON–OFF sources, denoted by X, is a discrete random variable, definition in Eq. (2) becomes

$$F(y) \triangleq E[(X-y)^+] \triangleq \sum_x (x-y)^+ f(x), \tag{10}$$

where $f(x)$ is the pmf of discrete random variable X.

Eq. (5) shows the method of calculating the clr of the aggregate traffic $\sum_{i=1}^{n} X_i$ using the clrf. For ON–OFF sources, the clrf can also be used to calculate the clr of individual connections, i.e. the clr of $X_{k(k=1,\ldots,n)}$ on a link with link capacity C is upper bounded by [13]:

$$\mathrm{clr}_k = \frac{E\left[\left(\sum_{i=1, i\neq k}^{n} X_i + \mathrm{pcr}_k - C\right)^+\right]}{C} = \frac{F(C-\mathrm{pcr}_k)}{C}, \tag{11}$$

where the function $F(y)$ is the clrf of $\sum_{i=1,i\neq k}^{n} X_i$. It can be shown that the clr of the aggregate traffic $\sum_{i=1}^{n} X_i$ is related to clr_k by:

$$\mathrm{clr} = \frac{p_k \times F(C-\mathrm{pcr}_k) + (1-p_k)F(C)}{E\left(\sum_{i=1}^{n} X_i\right)} = \frac{p_k \times C \times \mathrm{clr}_k + (1-p_k)F(C)}{E\left(\sum_{i=1}^{n} X_i\right)}.$$

Denoting the clr objective of X_k by $\mathrm{clr}_{k,\mathrm{obj}}$, if

$$\mathrm{clr} \le \frac{p_k \times C \times \mathrm{clr}_{k,\mathrm{obj}} + (1-p_k)F(C)}{E\left(\sum_{i=1}^{n} X_i\right)} \tag{12}$$

is satisfied, the clr QoS of connection X_k can be guaranteed. We can further remove the term $(1-p_k)F(C)$ from Eq. (12), i.e. if

$$\mathrm{clr} \le \frac{p_k C}{E\left(\sum_{i=1}^{n} X_i\right)} \times \mathrm{clr}_{k,\mathrm{obj}} \tag{13}$$

is satisfied, the QoS requirement of connection X_k can be guaranteed. Eq. (13) gives more conservative clr objective for the aggregate traffic than that given by Eq. (12), however Eq. (13) is much easier to implement practically. Since $F(C) < C\mathrm{clr}_{k,\mathrm{obj}}$, when p_k is not a very small value, the clr objective for the aggregate traffic given by Eq. (13) is close to that given by Eq. (12). In the rest of this paper, we limit our discussion to the cell loss performance of the aggregate traffic.

Proposition 1. *Assume X is an arbitrary traffic source such that $E(X) = \mathrm{mcr}$ and $\|X\|_\infty = \mathrm{pcr}$, where $\|X\|_\infty = \inf\{x : \Pr\{X > x\} = 0\}$. Let Y represent an ON–OFF source with mean cell rate mcr and peak cell rate pcr. Then $X <_{\mathrm{clrf}} Y$ and $E(X) = E(Y)$.*

A result to the same effect of Proposition 1 was proved in Appendix B of Ref. [24].

This proposition implies that among traffic sources with the same mcr and pcr, ON–OFF source constitutes the worst-case in cell loss analysis.

We shall now introduce a lemma on the multiplexing of two independent ON–OFF sources. This lemma is used in many parts of this paper. The comparison of the multiplexing of two independent ON–OFF sources constitutes one of the basic cases for loss performance analysis of ON–OFF sources, from which we can extend to the comparison of the multiplexing of many ON–OFF sources.

Lemma 1. *Let X_1, X_2 be two independent ON–OFF sources with peak cell rates $\alpha_1 \times \mathrm{pcr}$, $(1-\alpha_1) \times \mathrm{pcr}$, and activity parameters p_1, p_2 respectively, where $1 > \alpha_1 \ge 0.5$. Let Y_1 and Y_2 be two independent ON–OFF sources with* *peak cell rates* $\alpha_2 \times \mathrm{pcr}$, $(1-\alpha_2)\times \mathrm{pcr}$, *and activity parameters* q_1, q_2 *respectively, where* $1 > \alpha_2 \geq 0.5$. *Moreover,* $E(X_1+X_2)=E(Y_1+Y_2)$ *and* $\alpha_1 \leq \alpha_2$. *Then* $X_1+X_2 <_{\mathrm{clrf}} Y_1+Y_2$ *if and only if* $p_1p_2 \leq q_1q_2$ *and* $p_1+p_2-p_1p_2 \geq q_1+q_2-q_1q_2$.

See Ref. [25] for a proof. We shall now introduce an important theorem about heterogeneous Bernoulli sources. Bernoulli sources are ON–OFF sources with the same pcr.

Theorem 1. *Let* $X_1,\ldots,X_n$ *be n independent heterogeneous Bernoulli sources with activity parameters* $p_1,\ldots,p_n$ *respectively. Their activity parameters are subject to* $p_1+\cdots+p_n=P$. *Let* $Y_1,\ldots,Y_n$ *represent n independent homogeneous Bernoulli sources, where* Y_i *has the same peak cell rate as* $X_{i(i=1,\ldots,n)}$ *and an activity parameter* $p=P/n$. *Then* $E(\sum_{i=1}^n X_i)=E(\sum_{i=1}^n Y_i)$ *and* $\sum_{i=1}^n X_i <_{\mathrm{clrf}} \sum_{i=1}^n Y_i$.

See Ref. [14] for a proof. Theorem 1 states that homogeneous Bernoulli sources generate more cell loss than that of heterogeneous Bernoulli sources. Theorem 1 was first proposed as a conjecture by Rasmussen et al. [11]. Based on Theorem 1, they proposed that the cell loss of *n* heterogeneous ON–OFF sources, whose maximum peak cell rate is pcr, is upper bounded by that of *n* homogeneous ON–OFF sources with peak cell rate pcr, where the sum of mean cell rates remains the same. In addition to [11], Theorem 1 is also used for loss performance analysis in many other literature [7,26,27].

In real networks, many traffic sources of the same type have the same pcr. However, because of their specific application circumstances, they have different mean cell rates. Theorem 1 is very useful for analyzing this kind of traffic sources.

We shall now introduce another theorem about ON–OFF sources.

Theorem 2. *Let X and Y be two independent ON–OFF sources with the same mean cell rate denoted by mcr. X and Y have peak cell rates* pcr_X *and* pcr_Y *respectively. If* $\mathrm{pcr}_X \leq \mathrm{pcr}_Y$, *then* $X <_{\mathrm{clrf}} Y$.

See Ref. [14] for a proof. In Ref. [12], a result to the same effect of Theorem 2 was proved using another approach. Here it is proved using the clrf. Comparing the proof here and that in Ref. [12], the advantage of using clrf for cell loss analysis in the bffm becomes clear.

Theorem 2 is used by Lee et al. [12] in their CAC scheme. Based on it, they design a CAC scheme capable of real-time estimation of cell loss of the multiplexing of heterogeneous ON–OFF sources. Theorem 2 will be used for loss performance analysis and CAC scheme design in this paper.

On the basis of the theorems, lemmas and properties shown previously, we shall now introduce an important theorem about independent heterogeneous ON–OFF sources.

Theorem 3. *Let* $X_1,\ldots,X_n$ *be n independent heterogeneous ON–OFF sources with peak cell rates* $\mathrm{pcr}_1,\ldots,\mathrm{pcr}_n$ *and mean cell rates* $\mathrm{mcr}_1,\ldots,\mathrm{mcr}_n$ *respectively. Let* $Y_1,\ldots,Y_m$ *represent m independent homogeneous ON–OFF sources with peak cell rate pcr and mean cell rate mcr, where* $\mathrm{pcr} \geq \max\{\mathrm{pcr}_1,\ldots,\mathrm{pcr}_n\}$, $m=\lceil \sum_{i=1}^n \mathrm{pcr}_i/\mathrm{pcr}\rceil$ *and* $\mathrm{mcr}=\sum_{i=1}^n \mathrm{mcr}_i/m$. *Then*

$$\sum_{i=1}^n X_i <_{\mathrm{clrf}} \sum_{i=1}^m Y_i \quad \text{and} \quad E\left(\sum_{i=1}^n X_i\right)=E\left(\sum_{i=1}^m Y_i\right). \tag{14}$$

See Appendix A for a proof. This theorem provides a cell loss upper bound for heterogeneous ON–OFF sources. It states that the cell loss of heterogeneous ON–OFF sources is less than or equal to that of corresponding homogeneous ON–OFF sources given in the theorem. The sum of mean cell rates of the homogeneous sources remains the same as that of heterogeneous sources, and the sum of peak cell rates of the homogeneous sources is also substantially the same as that of the heterogeneous sources. It is not difficult, using the clrf, to prove that our upper bound is tighter than the upper bound proposed by Rassussen et al. [11].

The proposed upper bound can also be explained intuitively as follows: substituting *n* independent heterogeneous ON–OFF sources with $\lceil \sum_{i=1}^n \mathrm{pcr}_i/\mathrm{pcr}\rceil$ independent homogeneous ON–OFF sources, the sum of mean cell rates does not change, i.e. the traffic load is unchanged. However, with the decrease in the number of multiplexed ON–OFF sources the aggregate traffic becomes more bursty. Therefore for the same utilization cell loss will increase.

4. CAC scheme

The proposed upper bound can be applied to either traffic descriptor-based CAC or measurement-based CAC. In real networks, owing to policing functions and other regulatory mechanisms, it is reasonable to assume that the declared peak cell rates are tight. However, it is very difficult for all traffic sources to tightly characterize their mean cell rates. Since the upper bound involves both pcr and mcr, we adopt CAC scheme using on-line measurements.

Measurement-based CAC has been a hot research topic in recent years [1,4,7,8,28–32]. Shiomoto et al. [29] used a low pass filter to obtain the instantaneous VP utilization from crude measurements. A residual bandwidth is derived from the maximum of the observed instantaneous VP utilization. If the bandwidth requirement of the new connection is smaller than the residual bandwidth, the new connection is admitted, otherwise it is rejected. Gibbens et al. [7] used a decision-theoretic approach for call admission control to explicitly incorporate call-level dynamics into the model. In their work, call acceptance decisions are based on whether the current measured load is less than a precomputed threshold. In a study by Dziong et al. [30], Kalman filter is used to obtain an optimal estimate of mcr and variance. An aggregate equivalent bandwidth is then derived from the mean and variance estimate. A spare bandwidth is set in the aggregate equivalent bandwidth to account for the estimation errors. Grossglauser et al. [28] studied a robust measurement-based admission control with emphasis on the impact of estimation errors, measurement memory, call-level dynamics and separation of time scales. Their work [28,31] also identifies a *critical time-scale* $\tilde{T}_h$ such that aggregate traffic fluctuation slower than $\tilde{T}_h$ can be tracked by the admission controller and compensated for by connection admissions and departures, fluctuations faster than $\tilde{T}_h$ have to be absorbed by reserving spare bandwidth on the link. Using Gaussian and heavy traffic approximations, the critical time scale is shown to scale as $T_h/\sqrt{n}$, where T_h is the average flow duration and n is the size of the link in terms of number of flows it can carry.

In this section, based on the theoretical analysis in the previous sections, we develop a CAC scheme using parameters from both traffic descriptors and measurements. The principles introduced in Section 1 is used to govern the CAC scheme design. Robustness, flexibility and simplicity become major concerns in our CAC design and later simulation validation.

Let us select a traffic rate unit u such that u is greater than or equal to the maximum pcr of all traffic sources on the link. From now on in this paper all traffic rates are normalized with respect to u unless otherwise specified. Without loss of generality we assume that link capacity C is an integer multiple of u.

Suppose there are n independent heterogeneous ON–OFF sources, denoted by $X_1, \ldots, X_n$, on the link. The declared pcr of ON–OFF source $X_{i(i=1,\ldots,n)}$ is pcr$_i$. We keep a list of the declared pcr of all connections on the link and denote the sum of the pcr by PCR, i.e. $\mathrm{PCR} = \sum_{i=1}^{n} \mathrm{pcr}_i$. Realizing that mcr can not be tightly characterized by traffic sources, we obtain the sum of mcr from on-line measurements. This is of course the mcr of the link. We denote the measured mcr of the link by r.

From Theorem 1,

$$\sum_{i=1}^{n} X_i <_{\mathrm{clrf}} \sum_{i=1}^{m} Y_i,$$

where

$$m = \lceil \mathrm{PCR} \rceil, \tag{15}$$

and $Y_{i(i=1,\ldots,m)}$ is an independent ON–OFF sources with pcr 1 and activity parameter p. The choice of p should make the mcr of $\sum_{i=1}^{m} Y_i$ and the mcr of $\sum_{i=1}^{n} X_i$ equal. We will describe later the method of estimating p from measurements. The pmf of $\sum_{i=1}^{m} Y_i$ is given by the following binomial distribution:

$$f(x) = \begin{cases} \begin{pmatrix} k \\ m \end{pmatrix} p^k (1-p)^{m-k} & x = k \\ 0 & else \end{cases} \tag{16}$$

the clrf of $\sum_{i=1}^{m} Y_i$ is calculated as follows:

$$F(k) = \begin{cases} m \times p & k = 0 \\ F(k-1) - 1 + \sum_{i=0}^{k-1} f(i) & k \geq 1 \end{cases} \tag{17}$$

Eq. (17) comes directly from the definition of clrf for discrete random variable given in Eq. (10). The clr of $\sum_{i=1}^{m} Y_i$ on a link is estimated using Eq. (5). As introduced before, clr of $\sum_{i=1}^{m} Y_i$ is greater than or equal to that of $\sum_{i=1}^{m} X_i$.

4.1. Estimation of the activity parameter *p*

We shall now describe the estimation of p. Since we obtain the mcr of $\sum_{i=1}^{n} X_i$ from on-line measurements, p can be directly estimated as follows:

$$\hat{p} = \frac{r}{m} \tag{18}$$

Increasing the measurement window size, T_m, will increase the accuracy of the measured mcr r and the accuracy of estimation of p. Yamada et al. [33] introduced a method for finding the measurement window size. Estimation of p using Eq. (18) is simple however for accurate estimation it requires a large measurement window size. Here we use another approach.

In an ATM network, traffic can only arrive in integer multiples of an ATM cell. Therefore we first choose a sampling period T_s such that the impact of such granularity on traffic rate measurements taken over T_s can be ignored. In our analysis, T_s is chosen to be 100 cell unit time. One cell unit time is the time required to transmit an ATM cell on the link. Denote the mean and the variance of the traffic rate sample r_1 measured over T_s by S_T and σ_T^2, and the mean and the variance of the traffic rate sample r_K measured over a sampling period of $K \times T_s$ by S_K and σ_K^2, respectively.

Assuming that the aggregate traffic is stationary, then it can be shown that S_K, σ_K, S_T and σ_T are related by $S_K = S_T$ and

$$\sigma_K^2 = \sigma_T^2 \left[\frac{1}{K} + \frac{1}{K^2} \sum_{i=1}^{K} (K-i)\rho_i \right],$$

where ρ_i is the autocorrelation between traffic rate samples taken over $[0,T_s]$ and over $[(i-1)T_s, i \times T_s]$. σ_K decreases with the increase of K. In the above analysis we ignored the impact of call level dynamics, i.e. we assume no call is admitted into the network or depart from the network during the measurement window. The impact of call level dynamics is discussed later at the end of this subsection.

Assume that the aggregate traffic is Gaussian. If the following equation is chosen as an estimate of the mcr:

$$\hat{r} = r_K + \epsilon \times \sigma_K,$$

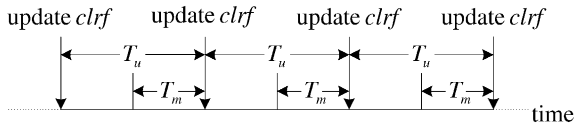

Fig. 2. Relationship between update interval and measurement interval.

where ϵ is a constant, in order to satisfy the estimation objective

$$P(\hat{r} \geq S_K) \geq 0.95, \tag{19}$$

we have to choose $\varepsilon = 1.65$. If the aggregate traffic is not Gaussian, a larger ε can be obtained using the Chebyshev's inequality. In the rest of the paper, we only consider the case when the aggregate traffic is Gaussian traffic.

An estimate of σ_K can either be obtained directly from on-line traffic measurements, or can be obtained from on-line estimation of σ_T and autocorrelation of function ρ. However, on-line estimation of these parameters is not an easy task, so we take another approach. Using Property 4, Theorem 3 and Proposition 1, it can be shown that $mS_T/m(1 - S_T/m)$ is the maximum variance of the aggregate traffic. Thus instead of measuring the variance directly, we estimate σ_T as follows:

$$\hat{\sigma}_T = \sqrt{m\frac{r_K}{m}\left(1 - \frac{r_K}{m}\right)}. \tag{20}$$

When a large enough K is chosen such that traffic fluctuation with time scale larger than $K \times T_s$ are small, the estimated σ_T is larger than its true value despite possible fluctuations in r_K. $\hat{\sigma}_K^2$ is then obtained as follows:

$$\hat{\sigma}_K^2 = \frac{\hat{\sigma}_T^2}{K^\delta}, \tag{21}$$

where δ is a constant between 0 and 1. δ can be obtained by inspecting the variance–time plot obtained from traffic measurements [34,35]. In contrast to [34] and [35] which study the self-similarity in network traffic, our interest is mainly in the variance–time curve in a relatively small time region from $10T_s$ to $1000T_s$. In this region even for short-range dependent traffic, a δ much smaller than 1 may be observed. In our simulations shown later, δ is a very stable value. This is because the traffic mix in the simulation is almost time-invariant. As a result, the autocorrelation in network traffic does not change significantly. However, our analysis on traffic measurement data from real ATM networks[1] shows that the value of δ will change slowly with time within a day. In that case, the smallest value of observed δ should be used in Eq. (21). For simulations using exponential ON–OFF sources shown later, δ is chosen to be 0.35; for simulations using Motion-JPEG encoded video sources, δ is chosen to be 0.4. The advantage of this method is that the value of δ can be obtained from off-line traffic analysis, therefore on-line estimation of the second order statistics is avoided. Here we would like to comment that this method will not give an accurate estimate of σ_K. However, generally the estimated value of σ_K is larger than its true value, which will satisfy the estimation objective in Eq. (19).

Then an estimate of p can be obtained: $\hat{p} = \hat{r}/m$.

To summarize the above analysis, if the measurement window size is chosen to be $T_m = K \times T_s$, an estimate of p can be obtained as follows:

$$\hat{p} = \frac{r}{m} + \alpha\sqrt{\frac{\frac{r}{m}\left(1 - \frac{r}{m}\right)}{m}}, \tag{22}$$

where

$$\alpha = 1.65 \times K^{-\delta/2}, \tag{23}$$

and r is the mean traffic rate measured over T_m.

This approach was first mentioned in Ref. [7]. The introduction of the safety margin α enables us to greatly reduce the required measurement window size while maintaining the robustness of the estimation. Moreover, the safety margin introduces additional benefits. Cell loss analysis presented in this paper gave a cell loss upper bound of the aggregate traffic based on bffm. The measurement scheme shown above also gives a robust estimate of p. Therefore, the above measurement scheme will give a tight QoS guarantee. However, for a network with large buffers, clr will decrease due to large buffer size. The proposed CAC scheme is conservative for network with large buffers. In this case, safety margin α can be chosen to be smaller than that given in Eq. (23), or even zero, to achieve higher network utilization. Our simulations show that for a fixed buffer size, controlling the safety margin α can control the clr, as well as utilization. Choosing a safety margin α to adapt to networks with large buffers is subject to further study.

Therefore the introduction of α brings some flexibility into the CAC scheme which enables us to efficiently utilize network resources. For the above reasons we use Eq. (22) to estimate p in our CAC scheme instead of [18]. The clrf of $\sum_{i=1}^{m} Y_i$ is updated using Eqs. (16) and (17) every T_u second. Fig. 2 shows the relationship between the update period T_u and measurement period T_m. We suggest choosing a T_u in the range $2T_m \sim 10T_m$, depending on the network state. When call level dynamics is high, i.e. connections enter and leave the network very often, T_u should be chosen close to $2T_m$. On the other hand, for a network where call level dynamics is low, T_u should be chosen close to $10T_m$.

Call level dynamics affects the measurements. If a new connection request is admitted in the measurement window, the new connection possibly generates traffic only during

[1] These traffic traces are collected by Waikatp Applied Network Dynamics group at the University of Auckland since November 1999. The time-stamp of measurement data in the trace has an accuracy of 1 µs. More details about the traces can be found at their webpage http://moat.nlanr.net/traces/Kiwitraces.

part of the measurement window, or does not generate traffic at all. Therefore, the new connection will make the measured mcr of the link smaller than its actual value, which will affect the robustness of the CAC scheme with regards to QoS guarantees. To solve this problem, when the measurement window size T_m is much smaller than the connection setup time, we block the admission of new connection during the measurement window. The admission will be delayed till the end of the measurement window. In the worst case, this will introduce a delay of T_m to the connection setup time. We use this method in our simulations. Alternatively, if the delay caused by T_m to the connection setup time becomes a concern, one can add the sum of the declared mean cell rates, if available in the traffic descriptor, or the sum of declared peak cell rates of the connections admitted in the measurement window, divided by m, to the estimated p in Eq. (22). In this case, we do not need to block the admission of new connections in the measurement window.

It is possible that during the measurement window some existing connections are released, thus affecting the measured mcr. Departing connections contribute to the measured mcr of the link. However, PCR is the sum of peak cell rates of connections on the link at the instant of updating clrf, not including peak cell rates of the departing connections. So if there are some connections that are released during the measurement interval they will make the estimated p larger, which in turn makes the CAC scheme more conservative, assuring that QoS guarantees are not affected. Another alternative is to update the clrf only when no existing connections depart during the measurement interval. We do not adopt this approach in our CAC scheme because it may result in the clrf not being updated for a long time, thus affecting the performance of the CAC scheme.

In our CAC scheme, we do not update the clrf for departing connections. The departing connections will be caught up by periodic updates of clrf.

4.2. Cell loss ratio estimation

We shall now show the method of estimating the clr when a new connection request arrives. Suppose there are M connections, denoted by $Z_1,\ldots,Z_M$, admitted into the network since the last measurement interval. Let us use $X_1,\ldots,X_n$ to denote the connections existing in the network at the instant when the last measurement interval finished, and $F(y)$ to be clrf of the corresponding upper bound of $X_1,\ldots,X_n$, $\sum_{i=1}^{m} Y_i$, which is calculated using Eqs. (15)–(17) and (22). When a new connection request arrives, denoted by Z_{M+1}, if $\mathrm{PCR} + \sum_{i=1}^{M+1} \mathrm{pcr}_{Z_i} \le C$, where $\mathrm{pcr}_{Z_i(i=1,\ldots,M+1)}$ is the declared pcr of Z_i, the new connection can be admitted directly and no cell loss will occur. Otherwise, the following method is used to estimate the clr if the new connection is accepted and to determine whether the new connection should be accepted:

Denote the updated clrf after the admission of Z_M by $F_M(y)$. The procedure of updating clrf is described later. Denote the pcr and mcr of the new connection request Z_{M+1} by $\mathrm{pcr}_{Z_{M+1}}$ and $\mathrm{mcr}_{Z_{M+1}}$. Z_{M+1} is smaller than, with respect to the clrf, an ON–OFF source λ with a pcr 1 and mcr $\mathrm{mcr}_{Z_{M+1}}$. The clrf of $\sum_{i=1}^{m} Y_i + \sum_{i=1}^{m} Z_i + \lambda$, denoted by $F_M G(y)$ can be calculated as:

$$F_M G(y) = F_M \times g(y)$$

where $g(y)$ is the pmf of λ. Therefore, the clr, if the new connection is admitted, is estimated as follows:

$$\mathrm{clr} = \frac{(1 - \mathrm{mcr}_{Z_{M+1}}) F_M(C) + \mathrm{mcr}_{Z_{M+1}} F_M(C-1)}{F_M(0) + \mathrm{mcr}_{Z_{M+1}}}. \tag{24}$$

If the estimated clr is less than the cell loss ratio objective then the connection is admitted; otherwise the connection is rejected. If the connection is admitted, clrf will be updated:

$$F_{M+1}(k) = \begin{cases} F_M(0) + \mathrm{mcr}_{Z_{M+1}} & k = 0 \\ (1 - \mathrm{mcr}_{Z_{M+1}}) F_M(k) + \mathrm{mcr}_{Z_{M+1}} F_M(k-1) & k \ge 1 \end{cases}. \tag{25}$$

For the special case of $M = 0$, the clrf $F_0(y)$ is actually $F(y)$, the clrf of $\sum_{i=1}^{m} Y_i$.

Alternatively, one can also take $\hat{r} + \sum_{i=1}^{M+1} \mathrm{mcr}_{Z_i}$ as an estimate of the sum of mean cell rates of all connections in the network if the new connection request Z_{M+1} is admitted; and calculate the clrf $F_{M+1}(y)$ using Eqs. (15)–(17) and (22). Accordingly, in Eq. (15), PCR now means the sum of peak cell rates of all connection in the network if connection request Z_{M+1} is admitted.

Updating of the clrf using Eq. (25) is computationally much more efficient. However, since Eq. (25) actually takes the pcr of $Z_1,\ldots,Z_M,Z_{M+1}$ as 1, it will generate more conservative results.

Noting that in the estimation of cell loss ratio only the computation of $F(y)$ and $f(x)$ from 0 to C is needed, we do not need to calculate all values of $F(y)$ and $f(x)$ from 0 to $\lceil \mathrm{PCR} \rceil$.

5. Simulation study

In this section, we study the performance of our CAC scheme using simulations. Eq. (25) is used to update the value of clrf. The aim of our simulation study is to evaluate the performance of our CAC scheme with respect to network utilization and its effectiveness in terms of its ability to guarantee the QoS constrains required by the connections.

The simulations are carried out using OPNET. The following parameters are used for our simulations: cell loss ratio objective is set to be 10^{-4}; switching speed of the ATM switch is set to be infinity, hence every incoming cell is placed immediately in the output buffer; the output buffer size is set to be 20 cells to absorb cell level congestion [36,37]. The link utilization and clr are observed in our

Table 1
Parameters of the three traffic types in saturation scenario

	λ (s^{-1})	pcr (Kb/s)	*Burstiness*	L (cells)
Type 1	10	100	10	100
Type 2	50	50	5	50
Type 3	100	10	2	20

simulations. Link utilization is calculated as the ratio of instantaneous link traffic rate to link capacity; clr is calculated as the ratio of the total observed cell loss to the total cells offered to the link in a moving window with size T_c. More specifically, clr at time t is the ratio of the number of cell loss occurred in the interval $(t - T_c, t)$ to the total number of cells offered to the link for transmission in the same interval, where T_c equals 500 s. The utilization data is collected every 0.1 s, so it is actually the average utilization over a 0.1 s interval. Clr data is collected every 1 s. In each scenario, there are several types of traffic sources multiplexed onto the link. Each type of traffic has an exponentially distributed arrival rate with the mean of λ calls per second. The connection holding time for all traffic types is exponentially distributed with a mean of 100 s.

5.1. Simulation model

In this section exponential ON–OFF source model is used in the simulations. The duration of the ON and OFF periods are independent and exponentially distributed with means β and γ, respectively. During each ON period an exponentially distributed random number of cells, with mean L, are generated at pcr. During off periods no cells are generated. We define the burstiness of a traffic source as:

$$\text{Burstiness} = \frac{\text{pcr}}{\text{mcr}} = \frac{\beta + \gamma}{\beta} \tag{26}$$

Furthermore, the following parameters are used for our simulations: the link capacity is set to be 10 Mb/s, and the measurement window size is chosen to be 0.08 s. Clrf update period is chosen to be 0.2 s. Safety margin α is chosen to be 1.0. In the simulations, three types of traffic sources are multiplexed onto the link. Traffic rate unit u is set to be 100 Kb/s, which is the maximum pcr of the three traffic types.

In this simulation scenario, referred to as the saturation scenario, the call arrival rate is chosen to be very high. The high call arrival rate and long call holding time mean that the system is continually receiving new connection requests. Thus, the CAC scheme is expected to achieve the maximum utilization in the saturation scenario. This scenario is used to establish the performance of our CAC scheme with regards to QoS guarantees, because if calls are offered at a very high rate, the rate at which calls are admitted in error becomes very large too [7].

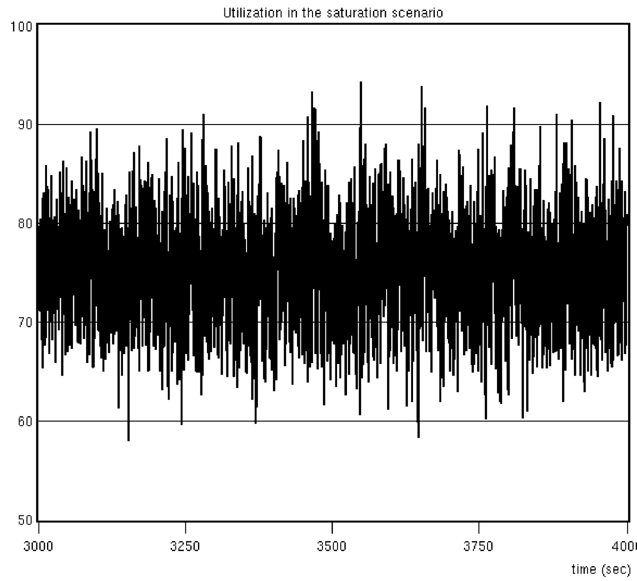

Fig. 3. Utilization achieved in the saturation scenario.

5.2. Saturation scenario

The parameters of the three traffic types of the saturation scenario are listed in Table 1.

The mean burst length is chosen to be several times larger than the buffer size. This is used as a trial to establish the performance of the CAC scheme using on-line measurement.

The simulation was run for 10,000 s. An average utilization of 0.76 is achieved by our CAC scheme. Fig. 3 shows the observed utilization during the period 3000–4000 s. Fig. 4 shows the observed clr. Fig. 5 shows the number of each connection type on the link. Fig. 6 shows the admissible region for the three types of calls as well as the number of calls actually admitted by the CAC scheme. The numbers of the admitted calls are close to the boundary of the admissible region whilst within the admissible region. Therefore, the CAC scheme is robust with regards to QoS guarantee, and is capable of achieving a high network utilization.

This is also verified by the observed clr shown in Fig. 4. The observed clr is within the same order of the clr objective.

6. Application of the CAC scheme to real traffic sources

In Section 5, we study the performance of our CAC scheme using the exponential ON–OFF source model. In this section, we will further study the performance of our CAC scheme using variable bit rate video sources. Eight Motion-JPEG (M-JPEG) encoded movies are used in the simulation. Accordingly, there are eight traffic types. Connections of each type has an exponentially distributed duration with a mean of 100 s, and connection requests arrive exponentially with a mean rate of 1 call/s. When a

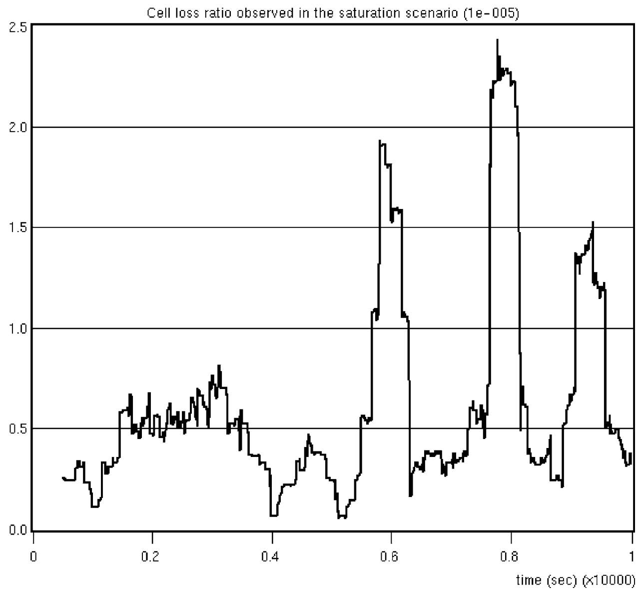

Fig. 4. Cell Loss Ratio observed in the saturation scenario.

connection of a type is admitted, it starts reading the corresponding M-JPEG encoded movie file from the beginning and generates traffic according to the movie file. The statistics of the M-JPEG encoded movie are shown in Table 2. The frame rate of the M-JPEG encoded movies is 30 frames/s. Details about the M-JPEG encoded movies can be found in Ref. [38]. Traffic rate unit u is chosen to be 7.288 Mbps, which is the maximum pcr of the movie sources. $OC3$ link is used in the simulation. The measurement window size is chosen to be 0.10 s. Clrf update period is chosen to be 0.2 s. Safety margin α is chosen to be 0.5. The simulation

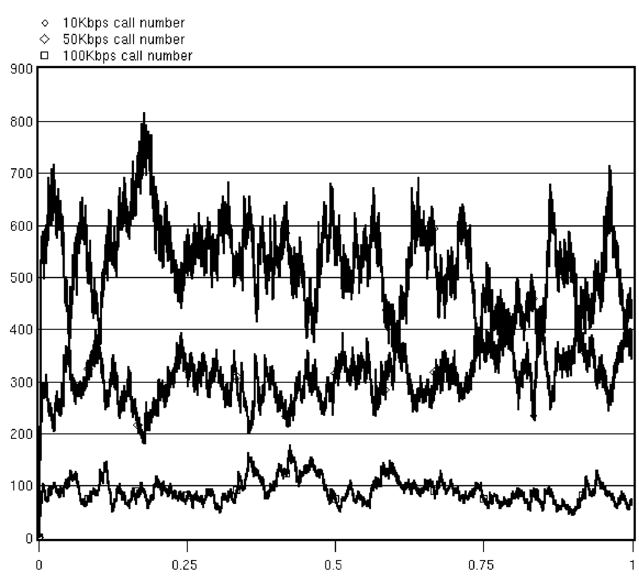

Fig. 5. The number of the three traffic types on the link observed in the saturation scenario.

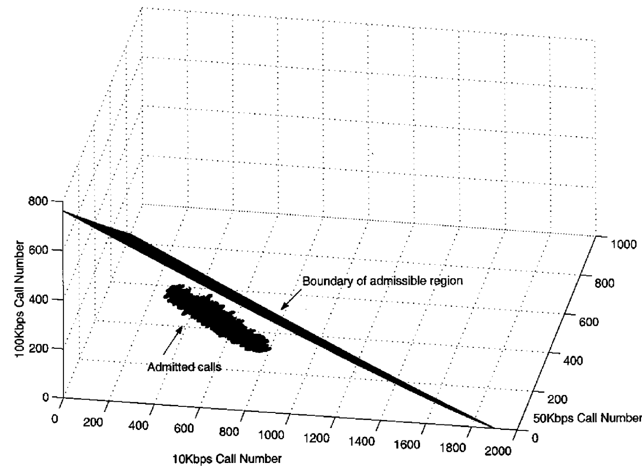

Fig. 6. Comparison between admissible region and actually admitted calls.

was run for 6000 s. Fig. 7 shows the clr observed in this scenario.

An average network utilization of 0.65 is achieved in this scenario while QoS is guaranteed. The utilization achieved is smaller than that achieved using the exponential ON–OFF sources. There are mainly two factors affecting the utilization. First, the pcr of the traffic sources are close to 10% of the link rate. As a rule of thumb, when the pcr of the traffic sources are close to 10% of the link rate, the statistical multiplexing gain which can be achieved is very small. Second, ON–OFF source considered in the CAC scheme is the worst case among all traffic sources with the same peak and mean. In this scenario, each video source is modeled by an ON–OFF source, which is actually the worst case of the real source. This will also result in lower utilization. It is worth noting that the aggregate traffic in this simulation scenario presents significant self-similar behavior with a *Hurst* parameter of 0.7 [35]. Fig. 8 shows the variance–time plot of the aggregate traffic rate. Therefore, it is possible that the proposed measurement-based CAC scheme can be applied to self-similar traffic. Further study is required to clarify this problem.

Table 2
Traffic rate of the M-JPEG encoded movies (bytes/frame)

Type	Name	Peak rate	Mean rate
1	Sleepless in Seattle	16617	9477.6
2	Crocodile Dundee	19439	10772.9
3	Home Alone, II	22009	11382.8
4	Jurassic Park	23883	11363.0
5	Rookie of the Year	27877	12434.9
6	Speed	29385	12374.4
7	Hot Shots, Part Deux	29933	12766.1
8	Beauty and the Beast	30367	12661.5

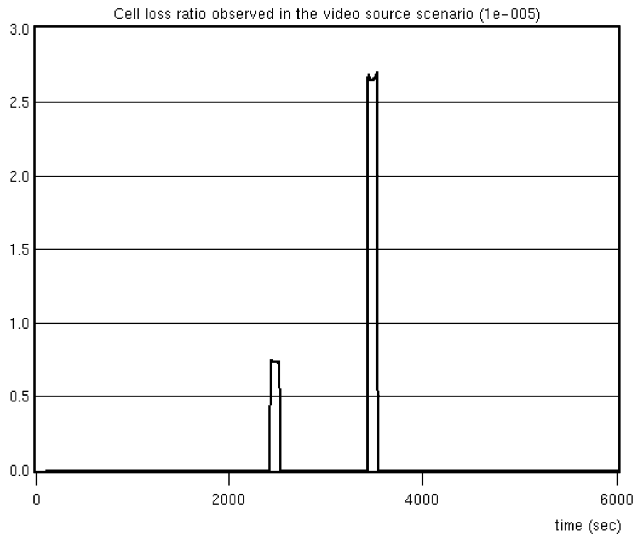

Fig. 7. Cell loss ratio observed in the video source scenario.

7. Conclusions

In this paper we proposed the clr function for studying cell loss in the bffm. The clrf enables us to decompose the complex analysis of the multiplexing of traffic sources into simpler analysis of individual sources. Using the clrf, cell loss of heterogeneous ON–OFF sources was studied. A cell loss upper bound for heterogeneous ON–OFF sources was proposed which may greatly simplify both theoretical analysis and computation. We believe this theoretical analysis forms a good basis for studying ON–OFF sources in the bffm. The methodology presented in this paper also constitute a good starting point for studying other traffic source models in the bffm.

The proposed upper bound can be applied to either parameter-based CAC or measurement-based CAC. In this paper, we used the upper bound to design a measurement-based CAC. Many practical factors were considered in the design of the CAC scheme. Extensive simulation studies were carried out which were indicative of good performance of the CAC scheme. Our CAC scheme proved to be time efficient, and capable of achieving high network utilization. The proposed CAC scheme has many attractive features which make it suitable for implementation in real ATM networks.

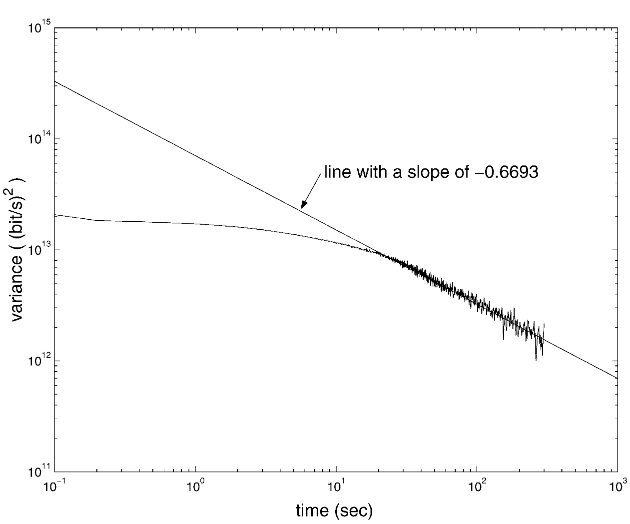

Fig. 8. Variance time plot of the aggregate traffic rate.

The loss performance analysis presented in this paper is a steady-state analysis, thus the proposed upper bound constitutes the worst case in terms of average loss performance. Theoretically speaking, the average loss performance constraints may not be meaningful, if the aggregate traffic exhibits long range dependence. Specifically, if the aggregate traffic exhibits long range dependence although average performance may be deemed to be fine, there may be rare periods of time in which performance is consistently poor.

However, there is considerable debate about the impact of long range dependent traffic on bandwidth allocation and network performance [39–42], especially the impact of long range dependent traffic on the performance of a measurement-based CAC scheme [28,31], because the time scale of interest in real applications is limited. For measurement-based CAC, Grossglauser et al. [28,31] identified a critical time-scale $\tilde{T}_h$ such that aggregate traffic fluctuations slower than $\tilde{T}_h$ can be tracked by the admission controller and compensated for by connection admissions and departures. Following their analysis, only those traffic fluctuations with time scale greater than measurement window size T_{m} but less than $\tilde{T}_h$ will threaten the performance of our CAC scheme. Our CAC scheme is based on an upper bound, i.e. it will over-allocate bandwidth. Thus we consider that if traffic fluctuations with time scale greater than measurement window size T_{m} but less than $\tilde{T}_h$ are negligible, or their bandwidth requirement can be satisfied by over-allocated bandwidth, then long range dependent traffic will not affect the performance of the proposed CAC scheme. Otherwise, the proposed CAC scheme will fail to provide QoS guarantee in the presence of long-range dependent traffic. The impact of the long range dependent traffic on the performance of our measurement-based CAC is in fact also problem for almost all measurement-based CAC. Further studies are required to clarify this problem.

Acknowledgements

The authors wish to thank Wu-chi Feng of the Ohio State University for making his video traces available for simulation purpose. They are also grateful to Joerg B. Micheel of Waikato Applied Network Dynamics group at Waikato University for making his traffic measurement data available for traffic analysis.

Appendix A

In this appendix, we shall prove Theorem 3. First we introduce an important lemma which will be used in the proof of Theorem 3.

Lemma A1. *Let* X_1,X_2 *be two independent ON–OFF sources with peak cell rates* $\alpha_1 \times \mathrm{pcr}$, $(1-\alpha_1)\times \mathrm{pcr}$, *and activity parameters* p_1, p_2 *respectively, where* $0.5 \le \alpha_1 < 1$. *Let* α_2 *be any value satisfying* $\alpha_1 \le \alpha_2 < 1$. *Then there exists two independent ON–OFF sources* Y_1,Y_2 *with peak cell rates* $\alpha_2 \times \mathrm{pcr}$, $(1-\alpha_2)\times \mathrm{pcr}$, *and* $E(Y_1+Y_2) = E(X_1+X_2)$, *such that* $X_1 + X_2 <_{\mathrm{clrf}} Y_1 + Y_2$.

See Chapter 3 of Ref. [25] for the proof. This lemma is used in proving Theorem 3, which is about heterogeneous ON–OFF sources. We shall now begin the proof of Theorem 3.

Proof. It is easy to show that $E(\sum_{i=1}^{n} X_i) = E(\sum_{i=1}^{m} Y_i)$. We prove $\sum_{i=1}^{n} X_i <_{\mathrm{clrf}} \sum_{i=1}^{m} Y_i$ using the induction method.

(1) Let us consider when $n=1$, then $m=1$. We must show that $X_1 <_{\mathrm{clrf}} Y_1$, where X_1 is an ON–OFF source with peak cell rate pcr_1 and mean cell rate mcr_1, and Y_1 is an ON–OFF source with peak cell rate $\mathrm{pcr}_1 \ge \mathrm{pcr}_1$, and mean cell rate mcr_1. This in fact is a direct application of Theorem 2.

(2) Suppose inequality (14) holds for the case when $n=k$, we must show that inequality (14) also holds for the case when $n=k+1$.

First we point out that since $\max\{\mathrm{pcr}_1,\dots,\mathrm{pcr}_k\} \le \max\{\mathrm{pcr}_1,\dots\mathrm{pcr}_{k+1}\}$, from our supposition that inequality (14) holds for $n=k$ and $\mathrm{pcr} \ge \max\{\mathrm{pcr}_1,\dots\mathrm{pcr}_k\}$, it naturally follows that it also holds for $n=k$ and $\mathrm{pcr} \ge \max\{\mathrm{pcr}_1,\dots,\mathrm{pcr}_{k+1}\}$. Now let us consider the two ON–OFF source X_k and X_{k+1}, we shall consider the following two cases:

(a) When $\mathrm{pcr}_k + \mathrm{pcr}_{k+1} \le \mathrm{pcr}$, using Proposition 1, it can be shown that

$$X_k + X_{k+1} <_{\mathrm{clrf}} Z$$

where Z is an ON–OFF source with peak cell rate $\mathrm{pcr}_z = \mathrm{pcr}_k + \mathrm{pcr}_{k+1}$ and mean cell rate $\mathrm{mcr}_z = \mathrm{mcr}_k + \mathrm{mcr}_{k+1}$. Then applying our supposition for k ON–OFF sources $X_1,\dots,X_{k-1},Z$, we can show that inequality (14) holds for the $k+1$ ON–OFF sources $X_1,\dots,X_{k+1}$, i.e.

$$X_1 + \cdots + X_{k-1} + X_k + X_{k+1} <_{clrf} X_1 + \cdots + X_{k-1} + Z$$

$$<_{clrf} Y_1 + \cdots + Y_{m_{k+1}}$$

where

$$m_{k+1} = \left\lceil \frac{\sum_{i=1}^{k-1} \mathrm{pcr}_i + \mathrm{pcr}_z}{\mathrm{pcr}} \right\rceil = \left\lceil \frac{\sum_{i=1}^{k+1} \mathrm{pcr}_i}{\mathrm{pcr}} \right\rceil$$

and $Y_{i(i=1,\dots,m_{k+1})}$ is an independent ON–OFF source with peak cell rate pcr and mean cell rate $\mathrm{mcr} = \sum_{i=1}^{k+1} \mathrm{mcr}_i / m_{k+1}$.

(b) When $\mathrm{pcr}_k + \mathrm{pcr}_{k+1} > \mathrm{pcr}$, using Lemma A1, we are able to find two independent ON–OFF source Z_1 and Z_2 and Z_1 has peak cell rate $\mathrm{pcr}_{z_1} = \mathrm{pcr}_k + \mathrm{pcr}_{k+1} - \mathrm{pcr}$ and mean cell rate mcr_{z_1}. Z_2 has peak cell rate pcr and mean cell rate mcr_{z_2}, $\mathrm{mcr}_{z_1} + \mathrm{mcr}_{z_2} = \mathrm{mcr}_k + \mathrm{mcr}_{k-1}$, such that

$$X_k + X_{k+1} <_{\mathrm{clrf}} Z_1 + Z_2.$$

Here $\mathrm{pcr} \ge \max\{\mathrm{pcr}_k, \mathrm{pcr}_{k+1}\}$ is equivalent to the condition in Lemma A1 that $\alpha_1 \le \alpha_2$.

From $\mathrm{pcr} \ge \max\{\mathrm{pcr}_1,\dots,\mathrm{pcr}_{k+1}\}$, it is guaranteed that $\mathrm{pcr} \ge \mathrm{pcr}_{z_1}$.

Then apply our supposition for the k independent ON–OFF sources $X_1,\dots,X_{k-1},Z_1$, we get:

$$X_1 + \cdots + X_{k-1} + Z_1 <_{\mathrm{clrf}} \lambda_1 + \cdots + \lambda_{m_k'}$$

where

$$m_k' = \left\lceil \frac{\sum_{i=1}^{k-1} \mathrm{pcr}_i + \mathrm{pcr}_{z_1}}{\mathrm{pcr}} \right\rceil = \left\lceil \frac{\sum_{i=1}^{k+1} \mathrm{pcr}_i}{\mathrm{pcr}} \right\rceil - 1$$

$\lambda_i (i=1,\dots,m_k')$ is an independent ON–OFF source with peak cell rate pcr and mean cell rate

$$\mathrm{mcr}_\lambda = \frac{\sum_{i=1}^{k-1} \mathrm{pcr}_i + \mathrm{mcr}_{z_1}}{m_k'}.$$

Here we note that $\lambda_i (i=1,\dots,m_k')$ and Z_2 have the same peak cell rate pcr, so using Theorem 1, it can be shown that

$$\lambda_1 + \cdots + \lambda_{m_k'} + Z_2 <_{\mathrm{clrf}} Y_1 + \cdots + Y_{m_{k+1}}$$

where

$$m_{k+1} = m_k' + 1 = \left\lceil \frac{\sum_{i=1}^{k+1} \mathrm{pcr}_i}{\mathrm{pcr}} \right\rceil$$

and $Y_{i(i=1,\dots,m_{k+1})}$ is an independent ON–OFF source with peak cell rate pcr and mean cell rate

$$\mathrm{mcr} = \frac{\sum_{i=1}^{k-1} \mathrm{pcr}_i + \mathrm{mcr}_{z_1} + \mathrm{mcr}_{z_2}}{m_{k+1}} = \frac{\sum_{i=1}^{k+1} \mathrm{pcr}_i}{m_{k+1}}$$

So, from the supposition that inequality (14) holds for$n=k$, we arrive at the conclusion that it should also holds for $n=k+1$.

Combining 1 and 2, we conclude that Eq. (14) holds for all n. □

References

[1] J. Sugih, B. Danzig Peter, J. Shenker Scott, Z. Lixia, A measurement-based admission control algorithm for integrated service packet networks, IEEE/ACM Transactions on Networking 5 (1) (1997) 56–69.

[2] D. Mitra, M.I. Reiman, J. Wang, Robust dynamic admission control for unified cell and call qos in statistical multiplexers, IEEE Journal on Selected Areas in Communications 16 (5) (1998) 692–707.

[3] C. Hao, L. Sanqi, Fast algorithms for measurement-based traffic modeling, IEEE Journal on Selected Areas in Communications 16 (5) (1998) 612–626.

[4] J. Lewis, R. Russell, F. Toomey, B. McGurk, S. Crosby, I. Leslie, Practical connection admission control for atm networks based on on-line measurements, Computer Communications 21 (17) (1998) 1585–1596.

[5] K. Shiomoto, S. Chaki, Adaptive connection admission control using real-time traffic measurements in atm networks, IEICE Transactions on Communications E78-B (4) (1995) 458–464.

[6] R. Bolla, F. Davoli, M. Marchese, Bandwidth allocation and admission control in atm networks with service separation, IEEE Communications Magazine 35 (5) (1997) 130–137.

[7] R.J. Gibbens, F.P. Kelly, P.B. Key, A decision-theoretic approach to call admission control in atm networks, IEEE Journal on Selected Areas in Communications 13 (6) (1995) 1101–1114.

[8] S. Lee, J. Song, A measurement-based admission control algorithm using variable-sized window in atm networks, Computer Communications 21 (2) (1998) 171–178.

[9] M. Fontaine, D.G. Smith, Bandwidth allocation and connection admission control in atm networks, Electronics and Communication Engineering Journal 8 (4) (1996) 156–164.

[10] A.E. Eckberg, B-ISDN/ATM Traffic and Congestion Control, IEEE Network Magazine 6 (1992) 28–37.

[11] C. Rasmussen, J.H. Sorensen, K.S. Kvols, S.B. Jacobsen, Source-independent call acceptance procedures in atm networks, IEEE Journal on Selected Areas in Communications 9 (3) (1991) 351–358.

[12] T.-H. Lee, K.-C. Lai, S.-T. Duann, Design of a real-time call admission controller for atm networks, IEEE/ACM Transaction on Networking 4 (5) (1996) 758–765.

[13] M. Reisslein, K.W. Ross, S. Rajagopal, Guaranteeing statistical qos to regulated traffic: the single node case, IEEE Infocom' 1999, 1999.

[14] G. Mao, D. Habibi, A tight upper bound for heterogeneous on–off source, IEEE Globecom (2000) 636–640.

[15] B. Maglaris, D. Anastassiou, P. Sen, G. Kaelsson, J.D. Robbins, Performance models of statistical multiplexing in packet video communication, IEEE Transactions on Communications 36 (7) (1998) 834–844.

[16] K. Sohraby, On the theory of general ON–OFF sources with applications in high-speed networks, INFOCOM, 1993, pp. 401–410.

[17] A. Baiocchi, N.B. Melazzi, M. Listanti, A. Roveri, R. Winkler, Loss performance analysis of an atm multiplexer loaded with high-speed on–off sources, IEEE Journal on Selected Areas in Communications 9 (3) (1991) 388–393.

[18] T. Murase, H. Suzuki, S. Sato, T. Takeuchi, A call admission control scheme for atm networks using a simple quality estimate, IEEE Journal on Selected Areas in Communications 9 (9) (1991) 1461–1470.

[19] P.T. Brady, A statistical analysis of on–off patterns in 16 conversations, Bell System Technical Journal (1968) 73–91.

[20] R.L. Easton, P.T. Hutchinson, R.W. Kolor, R.C. Moncello, R.W. Muise, TASI-E communication system, IEEE Transactions in Communications 30 (1982) 803–807.

[21] W. Willinger, M.S. Taqqu, R. Sherman, D.V. Wilson, Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level, IEEE/ACM Transactions on Networking 5 (1) (1997) 71–86.

[22] M.S. Taqqu, W. Willinger, R. Sherman, Proof of a fundamental result in self-similar traffic modeling, Computer Communication Review 27 (1997) 5–23.

[23] H. Chia-lin, L. San-qi, On input state space reduction and buffer noneffective region, Proceedings of the IEEE INFOCOM'94 8b.1.1 (1994) 1018–1028.

[24] Z.-L. Zhang, J. Jurose, J. Salehi, D. Towsley, Smoothing, statistical multiplexing and call admission control for stored video, Umass cmpsci technical report um-cs-96-29, Department of Computer Science, University of Massachusetts, Amherst, available via FTP from gaia.cs.umass.edu in pub/Zhan96: Smoothing.ps.gz, 1996.

[25] G. Mao, Statistical multiplexing and connection admission control in ATM Networks, PhD thesis, School of Engineering and Mathematics, Edith Cowan University, Australia, June 2001.

[26] N.G. Bean, Statistical multiplexing in broadband communication, PhD thesis, Univ. Cambridge, Cambridge 1993.

[27] R. Griffiths, P. Key, Adaptive call admission control in ATM networks, Proceedings of the 14th International Teletraffic Cong.-ITC 14, 1994.

[28] M. Grossglauser, D.N.C. Tse, A framework for robust measurement-based admission control, IEEE/ACM Transactions on Networking 7 (3) (1999) 293–309.

[29] K. Shiomoto, S. Chaki, N. Yamanaka, A simple bandwidth management strategy based on measurements of instantaneous virtual patch utilization in atm networks, IEEE/ACM Transactions on Networking 6 (5) (1998) 625–634.

[30] Z. Dziong, M. Juda, L.G. Mason, A framework for bandwidth management in ATM networks—aggregate equivalent bandwidth estimation approach, IEEE/ACM Transactions on Networking 5 (1) (1997) 134–147.

[31] M. Grossglauser, D.N.C. Tse, A time-scale decomposition approach to measurement-based admission control, Proceedings of the IEEE INFOCOM, 1999.

[32] S.M. Srinidhi, W.H. Thesling, V.K. Konangi, An adaptive scheme for admission control in atm networks, Computer Networks and ISDN System 29 (1997) 569–582.

[33] H. Yamada, S. Sumita, A traffic measurement method and its application for cell loss probability estimation in atm networks, IEEE Journal on Selected Areas in Communications 9 (3) (1991) 315–324.

[34] M.E. Crovella, A. Bestavros, Self-similarity in world wide web traffic: evidence and possible causes, IEEE/ACM Transactions on Networking 5 (6) (1997) 835–846.

[35] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, On the self-similar nature of ethernet traffic (extended version), IEEE/ACM Transaction on Networking 2 (1) (1994) 1–15.

[36] J.W. Roberts, Variable-bit-rate traffic control in B-ISDN, IEEE Communications Magazine 29 (9) (1991) 50–56.

[37] J.Y. Hui, Resource allocation for broadband networks, IEEE Journal on Selected Areas in Communications 6 (9) (1998) 1598–1608.

[38] W.-c. Feng, Video-on-demand services: efficient transportation on decompression of variable bit rate video, PhD thesis, Univ. of Michigan, available at http://www.cis.ohio-state.edu/wuchi/ April 1996.

[39] M. Grossglauser, J.-C. Bolot, On the relevance of long-range dependence in network traffic, IEEE/ACM Transaction on Networking 7 (5) (1999) 629–640.

[40] B.K. Ryu, A. Elwalid, The importance of long-range dependence of vbr video traffic in atm traffic engineering: myths and realities, Computer Communication Review, 24 (4) 1996.

[41] Y. Kim, S.-q. Li, Timescale of interest in traffic measurement for link bandwidth allocation design, Proceedings of the IEEE INFOCOM'96, 1996.

[42] D.P. Heyman, T.V. Lakshman, What are the implications of long-range dependence for VBR-video traffic engineering?, IEEE/ACM Transactions on Networking 4 (3) (1996) 301–317.