Refinements to CLT-based MBAC Schemes using Moderate Deviations

Sarang Wagholikar[1], R. Ravikanth[2] and R. Srikant[3]

Abstract

Central limit theorem (CLT) based techniques have been recently proposed to perform measurement-based admission control (MBAC). In this paper, we present refinements to this using moderate deviations expansions of the probability of overflow in a bufferless model.

1 Introduction

Emerging real-time applications have necessitated the support for QoS in packet networks. Even in the Internet, which was originally designed for best-effort communication, QoS support is being planned. QoS is typically defined in terms of parameters such as delay, delay-jitter and loss probability. As already studied extensively in the literature, admission control is a key aspect of QoS provisioning.

In order to provide guaranteed QoS to the users, the service provider needs to perform admission control. This requires knowledge of the bandwidth required by a new user and an estimate of the resources that have to be reserved for the sources already admitted into the network. Thus, admission control is done to ensure that QoS of all admitted sources is maintained at an acceptable level. If the traffic statistics of each source is known *a priori*, then an effective bandwidth can be computed for each source [17] and resources can be reserved for such a flow at all the intermediate network elements. Thus, each flow has a guaranteed QoS associated with it throughout its lifetime and the admission control is based on the *a priori* traffic specification of the flows.

In many situations, a priori traffic characterization may not be available. For example, in the case of online live events, it is difficult to provide the traffic parameters *a priori* as they change continuously with time. In such cases, the admission control mechanism needs to adapt to the overall resource utilization from time to time. Recently, a lot of work has been done in the area of Measurement Based Admission Control (MBAC) methods ([10, 4, 12, 15, 11, 2, 6, 1, 14, 3]) and their performance evaluation [5]. The traffic parameters are estimated online by taking measurements from time to time, and admission control is performed using these estimated parameters. If we consider a large number of users with similar traffic patterns then the aggregate behavior of the traffic tends to behave in a manner predicted by the Central Limit Theorem or large deviations and thus, admission control can be done by monitoring the aggregate traffic profile. This has the advantage that one need not keep track of individual user's traffic. Given that the number of users is very large, the measurement of aggregate traffic is relatively very simple as compared to per-flow measurements.

The admission control scenario considered in this paper adresses the problem of meeting loss probability requirement. We consider the situation where the capacity at a link is fixed, and the admission control problem is to admit flows, while attempting to meet a certain overflow probability requirement. We assume that that the buffer sizes are small. At all times, the network monitors the traffic on each link and based on this measurement, at each link, there is an estimated minimum capacity required to meet the QoS needs of the sources already in service. When a new source ar-

[1]Sarang Wagholikar, is with Cisco Systems, San Jose, CA. Email: sarang@cisco.com. This work was done when he was a graduate student at the University of Illinois.

[2]R. Ravikanth, is with Axiowave Networks, MA. This work was done when he was with the Nokia Research Center, Burlington, MA.

[3]R. Srikant, is with the Coordinated Science Lab. and Dept. of General Engineering, University of Illinois at Urbana-Champaign. Email: rsrikant@uiuc.edu.

rives, its peak rate is declared to the network, and it is admitted if the available capacity is more than the declared peak rate. However, through constant monitoring of the traffic, the total required capacity for all admitted sources is estimated from the measurements. Thus, the capacity allocated to admitted sources would be far less than the sum of their peak rates.

The central limit theorem is a convenient tool to compute the overflow probability since it only requires the estimation of two moments [1, 2, 3]. There are two dimensions to the admission control problem using CLT approximations: a *spatial* dimension and a *temporal* dimension. The spatial dimension deals with errors in the measurements due to the fact that the number of sources may not be large enough to use a limit theorem that is valid for a large number of sources. When using the CLT, there is a spatial dimension problem in getting accurate estimates of the mean and variance and the student-T distribution is used as an approximation to deal with this problem, instead of the Gaussian distribution. However, the spatial dimension problem we consider here is the following: *even if the mean and variance are known exactly, is the capacity requirement suggested by the CLT valid or are corrections required to make it a conservative estimate?*

By the temporal dimension problem, we refer to the problem of dealing with errors in measurements due to time correlations in the traffic and the fact that new flows are entering the system and old flows are leaving the system. This has been studied extensively in [3]. We do not directly address this problem in this paper. However, in our simulations, we do study the effect of the corrections proposed in this paper when the mean and variance are estimated over various time windows and for varying degrees of burstiness in the traffic models.

2 General Moderate Deviations Expressions

We now consider the moderate deviations result for the case of sum of general i.i.d. random variables. It states that if the random variables X_j satisfy Cramer's condition which requires, for some $\alpha > 0$,

$$E\{\exp(\alpha|X_j|)\} < \infty \tag{1}$$

then the following theorem ([8, 7]), holds.

Theorem 2.1 *If $x \geq 0$ and $x = o(n^{\frac{1}{2}})$ as $n \to \infty$, then*

$$\frac{Pr[S_n^* > x]}{1-\Phi(x)} = \exp\left[\frac{x^3}{n^{\frac{1}{2}}}\lambda\left(\frac{x}{n^{\frac{1}{2}}}\right)\right]\left(1+O\left(\frac{x+1}{n^{\frac{1}{2}}}\right)\right) \tag{2}$$

and

$$\frac{Pr[S_n^* < -x]}{\Phi(-x)} = \exp\left[-\frac{x^3}{n^{\frac{1}{2}}}\lambda\left(\frac{-x}{n^{\frac{1}{2}}}\right)\right]\left(1+O\left(\frac{x+1}{n^{\frac{1}{2}}}\right)\right) \tag{3}$$

Here $\Phi(x)$ is the cumulative distribution function for $\mathcal{N}(0,1)$ and $\lambda(z)$ is a power series constructed by means of the cumulants of the random variable X_j. The polynomial $\lambda(z)$ is given as follows,

$$\lambda(z) = \lambda_0 + \lambda_1 z + \lambda_2 z^2 + \ldots \tag{4}$$

$$= \frac{\gamma_3}{6\sigma^3} + \frac{\gamma_4\sigma^2 - 3\gamma_3^2}{24\sigma^6}z + \ldots \tag{5}$$

where

$$\gamma_2 = \sigma^2, \gamma_3 = \mu_3, \gamma_4 = \mu_4 - 3\sigma^4, \gamma_5 = \mu_5 - 10\mu_3\sigma^2, \ldots$$

are the cumulants of the variable X_j and μ_is are the moments of X_j.

If $x_n = o(n^{\frac{1}{6}})$, then we have,

$$\frac{\Pr[S_n^* > x]}{1-\Phi(x)} \sim 1 \tag{6}$$

In particular, if x_n increases faster than $n^{\frac{1}{6}}$ but $x_n = o(n^{\frac{1}{4}})$, then only the first term in the series $\lambda(z)$, i.e. λ_0, matters. In this case, one gets,

$$\frac{\Pr[S_n^* > x]}{1-\Phi(x)} \sim e^{\lambda_0 x^3/\sqrt{n}} \tag{7}$$

In general, if $x_n = o(n^\alpha)$, $0 < \alpha < 0.5$, the truncated power series

$$\lambda^{[s]}(z) = \lambda_0 + \lambda_1 z + \ldots + \lambda_s z^s \tag{8}$$

is sufficient to approximate $\Pr[S_n^* > x]$, where s is chosen to satisfy [7, 9]

$$\frac{1}{2}\frac{s}{s+2} \leq \alpha < \frac{1}{2}\frac{s+1}{s+3}. \tag{9}$$

Thus, if we know higher moments of the random variable X_j under consideration, it would be possible to refine the CLT-based approximation.

3 Bounds on the moderate deviations correction using only the first two moments

As we have seen, corrections to the CLT estimate of the overflow probability requires estimates of higher-order moments. However, higher-order moments are difficult to estimate accurately, and thus, we are motivated to find bounds based only on the first two moments. To allow for traffic heterogeneity, we consider k classes of sources. To relate this to the model in the previous section, we let that

$$X_j = \sum_{i=1}^{k} Y_{ji},$$

where $\{Y_{ji}\}_{i=1,n;j=1,k}$ are independent random variables, and for each j $\{Y_{ji}\}_{i=1,n}$ are identically distributed with mean μ_j and variance σ_j^2. Also, let $\mu = E(X_j) = \sum_{i=1}^{k} \mu_i$, and $\sigma^2 = E(X_j - \mu)^2 = \sum_{i=1}^{k} \sigma_i^2$. Let R_i be the peak rate of source Y_{ji}. Let us assume that $R_{max} := \max_i R_i$ is known *a priori*. When a new source arrives at the network, we need some minimal information to ascertain whether the capacity used by the sources already in the network plus the capacity needed by the new source is smaller than the total capacity at each link. Thus, the information we require for the MBAC scheme is simply the peak rate, and R_{max} is the maximum of the peak rates.

A crude, but simple estimate for λ_0 can be obtained as follows:

$$\begin{aligned}
\lambda_0 &= \frac{\mu_3}{6\sigma^3} \\
&= \frac{\sum_{i=1}^{k} E(Y_{ij} - \mu_j)^3}{6\sum_{i=1}^{k}(E(Y_{ij} - \mu_j)^2)^{3/2}} \\
&\leq max_i(R_i - \mu_i)\frac{E(X_i - \mu)^2}{6(E(X_i - \mu)^2)^{3/2}} \\
&\leq \frac{R_{max}}{6\sigma}. \qquad (10)
\end{aligned}$$

Suppose that the first term of the moderate deviations expansion is significant, then the CLT-based approximation for the overflow probability will not be accurate and we need to multiply it by $\exp(\mu_3 x^3/6\sigma^3\sqrt{n})$. If μ_3 is negative, then the actual overflow probability will be less than that predicted by the CLT-based approximation. Thus, the capacity allocated according to the CLT-based approximation will be conservative in this case. On the other hand, if the third moment is positive then surely CLT-based approximation underestimates the probability of overflow and the use of the upper bound in the (10) would give a more conservative estimate of the probability of overflow. In such a case, one would need to assign more capacity than that assigned according to the CLT-based analysis.

While the above upper bound can be applied to obtain conservative estimates of the required capacity whether the third moment is positive or negative, it is not necessary to correct the CLT estimate when the third moment is negative. Note that μ_3 is positive when all the p_is are less than $1/2$. Since, many traffic studies suggests that traffic sources are very bursty, $p_i < 1/2$ is the case of practical interest. However, it is certainly easier to estimate the sign of the third moment than to estimate it precisely. Thus, by estimating the sign of the third moment, it is possible to deside whether to use the moderate deviations refinement to the CLT or not.

The discussion thus far has focused on identifying the inaccuracies of the CLT-based approximation. Having identified the errors in the CLT based computation, we now provide a way to compute the additional bandwidth required to meet the desired loss rate requirement. The problem can be formally described as follows.

Capacity allocation problem: Let us suppose that there are n k-class ON-OFF sources, each generating traffic that has the mean μ and the variance σ^2. If p_q is the maximum overflow probability to be achieved then, the CLT-based analysis predicts that one should have the capacity C, given by

$$C \approx n\mu + \sqrt{n}\sigma Q^{-1}(p_q) \qquad (11)$$

The problem is to find how much extra bandwidth should be added to the capacity given by (11) to ensure that the probability of overflow, for the new capacity, is less than p_q.

3.1 Proposed Solution

Specifically, for a given p_q and if $\lambda_0 > 0$ and $Q^{-1}(p_q)$ is of the order of $n^{1/6}$, then the capacity assigned according to (11) may not be sufficient and the actual overflow probability for that capacity would be greater than $Q(x)$. In this case, the

following algorithm would give us a better value of the capacity.
Step 1: find the x that satisfies,

$$Q(x)e^{\lambda_0 \bar{x}^3/\sqrt{n}} = p_q \tag{12}$$

Step 2: Assign the capacity given as,

$$C = n\mu + \sqrt{n}\sigma\bar{x} \tag{13}$$

An estimate of $\bar{x}$ can be obtained as follows: Let $x = Q^{-1}(p_q)$, then

$$\bar{x} = x + |\Delta| \tag{14}$$

with

$$\Delta = \frac{\mu_3 x^2}{6\sqrt{n}\sigma^3}. \tag{15}$$

The above estimate, $\bar{x}$, of the required capacity is obtained as follows. We first use the following approximation for the ccdf of a Gaussian random variable:

$$Q(x) \approx \frac{1}{\sqrt{2\pi}x} e^{\frac{-x^2}{2}} \tag{16}$$

Putting $x = \bar{x} - \Delta$, we can express it as,

$$p_q = Q(x) \approx \frac{1}{\sqrt{2\pi}(\bar{x} - \Delta)} e^{\frac{-(\bar{x}-\Delta)^2}{2}} \tag{17}$$

Assuming that Δ is negligible compared to x, we can express $Q(x)$ in terms of $Q(\bar{x})$ as follows,

$$p_q = Q(x) \approx \frac{1}{\sqrt{2\pi}\bar{x}} e^{\frac{-\bar{x}^2 - 2\Delta\bar{x}}{2}} \approx Q(\bar{x})e^{\Delta\bar{x}} \tag{18}$$

This should be equal to the actual overflow probability. The actual overflow probability using the moderate deviaitions expansion is given by

$$Q(\bar{x})e^{\lambda_0 \bar{x}^3/\sqrt{n}}$$

So, we should have,

$$\Delta\bar{x} \approx \frac{\lambda_0 \bar{x}^3}{\sqrt{n}} \tag{19}$$

This gives us

$$\Delta \approx \frac{\lambda_0 \bar{x}^2}{\sqrt{n}} \approx \frac{\lambda_0 x^2}{\sqrt{n}} = |\frac{\mu_3 x^2}{6\sqrt{n}\sigma^3}| \tag{20}$$

We can express (14) in terms of the moments of the aggregate traffic, instead of the moments of the k-class ON-OFF source. Let the first second and third moments of the aggregate traffic profile be denoted by $\bar{\mu}$, $\bar{\sigma}^2$ and $\bar{\mu_3}$, respectively, i.e.,

$$\begin{aligned}
\bar{\mu} &= E(\sum_{i=1}^{i=n} X_i) = n\mu \\
\bar{\sigma}^2 &= E(\sum_{i=1}^{i=n} X_i - n\mu)^2 = n\sigma^2 \\
\bar{\mu}_3 &= E(\sum_{i=1}^{i=n} X_i - n\mu)^3 = n\mu_3
\end{aligned} \tag{21}$$

The required capacity is given by

$$\text{Capacity} = \bar{\mu} + (x + \Delta)\bar{\sigma} \tag{22}$$

From (10), for k-class ON-OFF sources, a conservative estimate of the assigned capacity should be

$$\text{Capacity} = \bar{\mu} + (x+\Delta)\bar{\sigma} \leq \bar{\mu} + x\bar{\sigma} + \frac{R_{max}x^2}{6}. \tag{23}$$

We will refer to this as the moderate-deviations estimate of the required capacity.

This is a simple refinement to the CLT-based estimate of the capacity and it depends only on the maximum peak rate R_{max}. This completely avoids the estimation of $\bar{\mu}_3$ and only the estimates of the mean and the variance of the aggregate traffic profile are used to improve the estimate of the capacity to be assigned in the usual CLT-based MBAC. The above upper bound on $\mu_3/6\sigma^3$ is tight i the following sense. Suppose we consider ON-OFF sources so that $\{Y_{ji}\}$ are independent Bernoulli random variables. Let Y_{ji} be R_i with probability p_i, and is 0 with probability $(1 - p_i)$. Let us suppose that $k = 2$ and $R_1 p_1 = R_2 p_2$. Define $\alpha = R_1/R_2$. Then, we have

$$\begin{aligned}
\lambda_0 &= \frac{\mu_3}{\sigma^3} \\
&= \frac{R_1^3(p_1 - {p_1}^2)(1-2p_1) + R_2^3(p_2 - p_2^2)(1-2p_2)}{[R_1^2 p_1(1-p_1) + R_2^2 p_2(1-p_2)]^{\frac{3}{2}}} \\
&= \frac{\alpha^3(p_1 - p_1^2)(1-2p_1) + (p_2 - p_2^2)(1-2p_2)}{[\alpha^2 p_1(1-p_1) + p_2(1-p_2)]^{\frac{3}{2}}}
\end{aligned}$$

When p_1 is not equal to 1 or 1/2,

$$\lim_{\alpha\to\infty} \lambda_0 = \frac{(1-2p_1)}{\sqrt{p_1(1-p_1)}}$$

The upper bound (21) becomes,

$$\lim_{\alpha\to\infty} \frac{R_1}{\sigma}$$

$$= \lim_{\alpha\to\infty} \frac{R_1}{[{R_1}^2 p_1(1-p_1)+{R_2}^2 p_2(1-p_2)]^{\frac{1}{2}}}$$
$$= \lim_{\alpha\to\infty} \frac{\alpha}{[\alpha^2 p_1(1-p_1)+p_2(1-p_2)]^{\frac{1}{2}}}$$
$$= \frac{1}{\sqrt{p_1(1-p_1)}}$$

Since $R_1 p_1 =$ constant, $p_1 \to 0$ as $\alpha \to \infty$. Thus, we see that,

$$\lim_{\alpha\to\infty} \lambda_0/(\frac{R_1}{\sigma}) = 1. \quad (24)$$

Thus the bound is asymptotically tight in the limit $\alpha \to \infty$.

4 Numerical and Simulation Results

Let us consider 2-class ON-OFF sources in which the probabilities p_is are less than 1/2. We are interested in comparing the CLT approximation and the one-term moderate deviations expansion. In Figure 1, we plot the ratio of the probability of overflow obtained using the one-term moderate deviations expansion and the probability of overflow given by CLT-based approximation varies with different combinations of p_is for a fixed choice of R_i's.

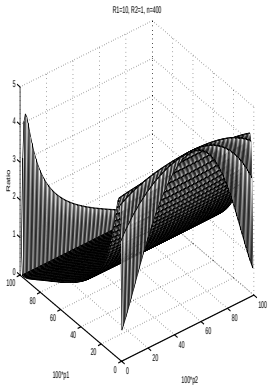

Figure 1: Ratio of the probability of overflow using one-term in the moderate deviations to the CLT estimate

We observe that for p_is less than 1/2, the ratio is greater than one and for some combinations it can go up to 5. Thus for such values of p_is, the CLT-based analysis significantly underestimates the actual probability of overflow. In Figure 2, we see that the variation in p_1 affects the value of the "ratio" more as compared to the variation in p_2. This is due to the fact that R_1 is very large compared to R_2 and hence the first component in the source dominates the "ratio" value.

Table 1: Capacities assigned by different methods for different burstiness factors

$\frac{R_1}{R_2}$	Cpcty by CLT	Cpcty by actual calc.	Cpcty by mod. dev.
3	504	507	513
5	534	542	549
10	590	609	620
15	632	663	677
20	668	710	728
25	700	752	775
30	729	795	819

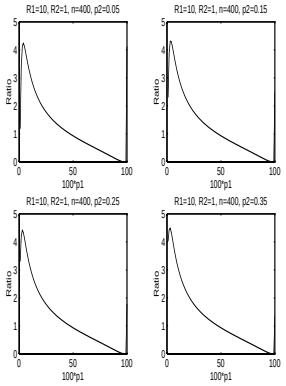

Figure 2: Variation of "ratio" with p_1 value

If R_1/R_2 is increased then the maximum value of the "ratio" in the previous graphs would increase, prompting us to go for asymptotic expansion based approach. Let us see how the ratio R_1/R_2 affects the performance of the capacity estimate proposed in (14). We consider an experiment in which the total number of sources (n) is fixed to 400, and let $R_2 = 1$, $R_1 p_1 = R_2 p_2 = 0.5$ and $p_q = 10^{-5}$. Thus, we are interested in computing the capacity required to keep the probability of overflow less than 10^{-5}. We compute the capacity assigned using CLT-based approach, using exact calculations of overflow probability and the capacity assigned using moderate deviations (23). The results are shown in Table 4.

We observe that as the ratio R_1/R_2 increases, the CLT-based analysis heavily underestimates the actual minimum capacity to be provided. In contrast, the capacity given by the formula (23) always tends to overestimate the required capacity, thus ensuring that QoS requirement is met. The ratio R_1/R_2 gives a measure of the burstiness of the source and the higher moments like μ_3 are large for the bursty sources. Thus in such cases, the terms in the exponent of the moderate deviations expansion become significant.

It is also useful to compare the actual probability of overflow when the capacities are assigned by the three methods as in Table 4. Figure 3 shows the actual probability of overflow andit can be seen that the QoS requirement is always met when the moderate deviations correction is applied to the CLT estimate of the required capacity. The estimated capacity by the CLT alone results in a probability of overflow that is sometimes an order magnitude more the desired value.

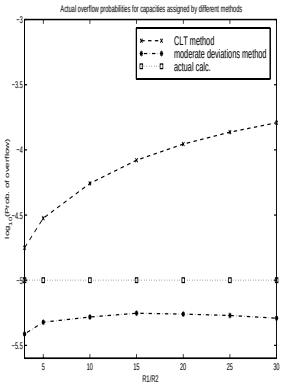

Figure 3: Probabilities of overflow using the different capacity allocation methods

References

[1] D. Tse and M. Grossglauser, Measurement-based call admission control: Analysis and simulation, in *Proceedings IEEE INFOCOM,* 1997.

[2] M. Grossglauser and D. Tse, A framework for robust measurement-based admission control, in *Proceedings ACM SIGCOMM,* 1997.

[3] M. Grossglauser and D. Tse, A time-scale decomposition approach to measurement-based admission control, in *Proceedings IEEE INFOCOM,* 1999.

[4] S. Jamin, P. B. Danzig, S. Shenker and L. Zhang, A measurement-based admission control algorithm for integrated services packet networks, in *Proceedings ACM SIGCOMM,* 1995.

[5] L. Breslau, S. Jamin and S. Shenker, Measurement-based admission control: An empirical performance comparison, preprint

[6] L. Breslau, S. Jamin and S. Shenker, Measurement-based admission control: What is research agenda?, preprint

[7] I. A. Ibragimov and Yu. V. Linnik, *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff Publishing Groningen, 1971.

[8] W. Feller, *Introduction to Probability theory and its Applications*, Wiley Eastern, 3 ed, Vol. 1, 1991.

[9] W. Feller, *Introduction to Probability theory and its Applications*, Wiley Eastern, 3 ed, Vol. 2, 1991.

[10] R. J. Gibbens, F. P. Kelly and P. B. Key, A decision-theoretic approach to call admission control in ATM networks , *IEEE Journal on Selected Areas in Communications ,* pp.1101-1114, 1995.

[11] M. Siler and J. Walrand, Monitoring quality of service: Measurement and estimation, in *Proceedings IEEE Conference on Decision and Control,* 1998.

[12] C. Casetti, J. Kurose and D. Towsley, An adaptive algorithm for measurement-based admission control in integrated services packet networks , in *Int. Workshop on Protocols for High Speed Networks,* 1996.

[13] N. G. Duffield, A large deviation analysis of errors in measurement based admission control to buffered and bufferless resources, *to appear in Queuing Systems.*

[14] N. G. Duffield, Asymptotic sampling properties of effective bandwidth estimation for admission control, in *Proceedings IEEE INFOCOM,* 1999.

[15] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell and F. Toomey, Entropy of ATM traffic streams: a tool for estiamting QoS parameters, *IEEE Journal on Selected Areas in Communications ,* vol.13, pp.981-990, 1995.

[16] Special issue on packet speech and video, *IEEE Journal on Selected Areas in Communications,* vol.7, no.5, June 1989.

[17] G. Veciana and J. Walrand, Effective bandwidths: Call Admission, traffic policing and filtering in ATM networks, *Queueing Systems,* vol.20, pp.37-59, 1995.