

Monitoring Quality of Service: Measurement & Estimation

Matthew Siler and Jean Walrand
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
siler@eecs.berkeley.edu, wlr@eecs.berkeley.edu

Abstract

To provide statistical guarantees of QoS, the Internet requires a measurement infrastructure for estimating available resources from actual traffic. In this paper, we outline an algorithm that collects a histogram of the occupancy of a single-server FCFS queue at packet arrival times, and infers the loss rate and delay distribution from such measurements. Direct estimation of such QoS parameters typically leads to estimators with a large variance. To reduce this variance, we fit a buffer occupancy model, a sum of exponentials, to the histogram using a weighted least-squares algorithm. Furthermore, we compute batch means to minimize the bias due to the positive correlation between measurements. In this manner, we provide an efficient and robust approach to QoS estimation.

1 Introduction

To deliver realtime multimedia, Internet applications rely on compression and backoff algorithms to try to provide smooth images and clear sound. However, the “best-effort” nature of the Internet limits the quality of these applications. If the Internet continues to increase at its current rate, the network will have to provide differentiated services, either through Class of Service (CoS) or Quality of Service (QoS) performance guarantees. While CoS-based contracts are typically arranged and paid for in advance, QoS-based contracts can be set up at the time of connection, so that customers are billed for what they use. Although a traffic descriptor may be provided, it may be too conservative for the actual traffic, or too complex to estimate usage accurately. In such cases, the network should measure network traffic and estimate current usage in

terms of the loss rate, L , and the distribution of the packet delay, D . However, an important question is how to estimate these parameters from measurements of network traffic.

In this paper, we outline a measurement-based approach for estimating these QoS parameters indirectly by fitting a function to the buffer occupancy distribution. These estimators are designed for Markov-modulated traffic sources in a first-come-first-serve (FCFS) queue. In our approach, we fit a distribution function to measurements of buffer occupancy, and then use this fitted function to infer the QoS parameters. By fitting the buffer occupancy function at multiple points, we can trade off complexity for speed of convergence. Although this approach uses more sophisticated measurement and estimation algorithms than more direct methods, we believe that the increased value and efficiency of the network will be worth the additional complexity.

In Section 2, we investigate an appropriate queuing model based on Markov-modulated traffic in a FCFS single-server queue. Section 3 details our approach, including the appropriate measurements and fitting algorithm that we use to estimate this distribution. Lastly, we support our approach with a simulated example of Poisson traffic in Section 4.

2 Modeling the Buffer Occupancy Distribution

We consider the problem of estimating L and $P(D \leq d)$ for stationary and ergodic Markov-modulated traffic in a single-server FCFS queue. Below, we describe our model in more detail, and summarize important results.

2.1 Queue and Traffic Model

Consider a G/G/1 FCFS queue, with rate R bytes per second and a maximum buffer size of B bytes, as shown in Figure 1.

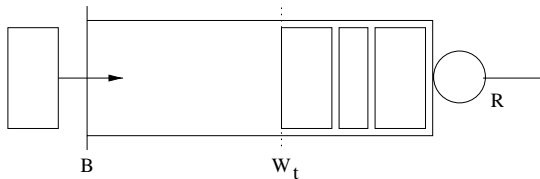


Figure 1: Single-server FCFS queue with rate R and buffer size B .

Packets with a random length enter the queue and are serviced at a constant rate. Denote the arrival process by the random variables $((S_n, T_n), n \geq 1)$, where $S_n > 0$ is the length, in bytes, of the n -th packet, and $T_n > 0$ is the interarrival time between the n -th and the $(n + 1)$ -st packets. We assume that (S_n, T_n) are modulated by a Markov process. More specifically, let Z_n be an irreducible and aperiodic Markov chain on a finite state space \mathcal{S} . Then, $(S_n, T_n) = h(Z_n)$, for some function $h(\cdot)$ and for every packet n , and therefore the subsequent packet lengths and interarrival times of the traffic process are functions of Z_n . Since Z_n is irreducible and positive recurrent, and assuming that Z_1 has a stationary distribution, the sequence $(Z_n, n \geq 1)$ is stationary and ergodic. Hence, the sequence of packet lengths and interarrival times are also stationary and ergodic, and we may define S and T to be a typical packet length and interarrival time, respectively. The buffer occupancy process is defined as the following. Define τ_n to be the n -th packet arrival time, so that

$$\tau_{n+1} = \tau_n + T_n.$$

Let W_t^B be the buffer occupancy, in bytes, at time $t \in [0, \infty)$ for a queue with a buffer size of B bytes. Furthermore, let $W_{\tau_n}^B$ be the number of bytes in the queue immediately after the n -th packet arrives. Then, we may define the queue occupancy process by the set of equations

$$W_{\tau_n}^B = W_{\tau_{n-1}}^B + S_n 1\{W_{\tau_{n-1}}^B + S_n \leq B\}$$

and

$$W_t^B = (W_{\tau_n}^B - (t - \tau_n)R)^+, \quad t \in [\tau_n, \tau_{n+1})$$

where $(\cdot)^+ \equiv \max(\cdot, 0)$. As defined, the occupancy of the actual queue at time t is W_t^B since packets are dropped if the queue occupancy exceeds B

bytes. Furthermore, we make the assumption that the queue is always stable, so that

$$E[S] < E[T]R.$$

2.2 Stationary Distribution

The QoS that packet n receives depends on the random variable $W_{\tau_n}^B + S_n$ since this decides whether the packet is lost, and if not, what its delay in the queue will be. Therefore, the QoS of the collection of packets can be determined from the stationary distribution of $W_{\tau_n}^B + S_n$. More precisely, we define F^B to be

$$\begin{aligned} F^B(x) &\equiv \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n 1\{W_{\tau_i}^B + S_i \leq x\} \quad (1) \\ &= P(W_{\tau_n}^B + S \leq x). \end{aligned}$$

The distribution F^B is the workload seen by an arriving packet (including its own work to be done) if the packet were added to the queue. An important related distribution is

$$\begin{aligned} \bar{F}^B(x) &\equiv P(W_{\tau_n}^B + S \leq x \mid W_{\tau_n}^B + S \leq B) \\ &= F^B(x)/F^B(B). \end{aligned}$$

\bar{F}^B is the workload as seen by packets that have not been dropped, and therefore it is used to determine the delay distribution.

Because of our modeling assumptions, F^B (and, by definition, \bar{F}^B) exists and is unique.

Theorem 1 *If $E[S] < E[T]R$, then F^B exists and is unique for all $0 \leq B \leq \infty$.*

This follows from the existence and uniqueness of F^∞ , which is the classic result due to Loynes [1], and the fact that W_t^B is stochastically smaller than W_t^∞ for finite B . We will refer to F^B as F and W_τ^B as W_τ , and only use the superscript when necessary for clarification. Under this model, we may compute all of our QoS performance measures directly from F . The loss rate, i.e., the fraction of packets that will be dropped by the queue because of overflow, is defined as

$$L \equiv \Pr(W_{\tau_n}^B + S > B) = 1 - F(B).$$

Furthermore, the distribution of D is defined as

$$P(D \leq d) \equiv \bar{F}(Rd).$$

This is the delay distribution for packets that are not dropped from the queue. In particular, the moments of D may be computed from this equation, including the average packet delay.

2.3 Regeneration Times

Under these modeling assumptions, it is possible to find regeneration times for the process. In this context, the process regenerates whenever the traffic process renews and the queue returns to zero. For any state $z \in \mathcal{S}$, we may define

$$\gamma_m^z = \inf\{n > \gamma_{m-1}^z : W_{\tau_n} = 0, Z_n = z\} \quad (2)$$

for $m \geq 1$, where γ_0^z is the first such time. Because of our modeling assumptions, the moments of the regeneration times exist and are finite for at least one state $\bar{z} \in \mathcal{S}$.

Lemma 1 *Let $(\gamma_m^z, m \geq 0)$ be defined as in (2). Then, there exists a state $\bar{z} \in \mathcal{S}$ such that $E[\gamma_m^{\bar{z}}] < \infty$ and $\text{Var}[\gamma_m^{\bar{z}}] < \infty$.*

This follows from the fact that $(\gamma_{m+1}^{\bar{z}} - \gamma_m^{\bar{z}}, m \geq 1)$ are i.i.d. and are upper bounded by an appropriate geometric random variable. The importance of Lemma 1 is that it is possible to find regeneration times with intervals of bounded mean and variance, and therefore we can apply a central limit theorem to the samples we collect. If we were to assume that the traffic process exhibits long-range dependence, then this would not be possible, and the estimator would exhibit such a large variance as to be unreliable.

2.4 Approximating Family of Distributions

To approximate the distribution F , we consider a collection of distribution functions on $[0, \infty)$ of the form

$$G(\theta; x) = 1 - \sum_{j=1}^J \alpha_j e^{-\beta_j x}, \quad x \geq 0 \quad (3)$$

where $\theta \in \Theta$ is a vector of real, non-negative parameters given by

$$\theta = (\alpha_1, \dots, \alpha_{J-1}, \beta_1, \dots, \beta_J)^T$$

and $\alpha_J \equiv 1 - \sum_{j=1}^{J-1} \alpha_j$. We restrict $\alpha_j \in [\alpha_{\min}, \alpha_{\max}]$ and $\beta_j \in [0, \beta_{\max}]$, for some $\beta_{\max} > 0$, so that Θ is a compact subset of \mathfrak{R}^{2J-1} . Moreover, $G(\theta; x)$ is continuous and infinitely differentiable in θ . Let $\mathcal{G}_J = \{G(\theta; x) : \theta \in \Theta\}$ be the set of all approximating distribution functions of F , for a given J . Although various models are possible, the sum of exponentials model corresponds well to certain classes of Markov-modulated traffic

in a FCFS queue since the waiting time distribution has a matrix-geometric form. In particular, if the packet lengths have an exponential distribution, then F^∞ is precisely a finite sum of exponentials [5].

3 Measurement and Estimation

In the previous section, we saw that F could be approximated with the function $G(\theta)$ for some $\theta \in \Theta$. In this section, we detail our approach for estimating θ given a histogram of F . We begin with an overview of model selection techniques, and then discuss a maximum likelihood estimator of θ as well as an implementable algorithm based on batch means. For details on convergence results, please see [7].

3.1 Model Selection

Model selection techniques focus not only on the estimation of model parameters, but also on how well particular models describe the behavior of a stochastic process. In our case, the model selection problem is, given the set \mathcal{G}_J , to select the distribution function from this set that best approximates F . Furthermore, since \mathcal{G}_J represents a distinct model for each J , we also need to choose J based on a suitable criterion, discussed below.

How we select $G(\theta)$ depends on how we measure the disparity between two distribution functions, defined by a discrepancy measure $\Delta(G, F)$. In particular, we are interested in minimizing $\Delta(G(\theta), F)$ with respect to $\theta \in \Theta$ for some appropriate discrepancy measure. Many types of discrepancy measures are available in the literature depending on what properties of distributions from \mathcal{G}_J are most important [4, 6]. Using this discrepancy measure, we may compute

$$\theta^* = \arg \min_{\theta \in \Theta} \Delta(G(\theta), F). \quad (4)$$

Then, $G(\theta^*)$ would be the best estimate of F , with respect to the discrepancy measure $\Delta(G(\theta), F)$. Furthermore, we may construct an estimate of θ^* based on \hat{F} , an empirical estimate of F . In this case, the empirical discrepancy measure is defined as $\hat{\Delta}(G(\theta), \hat{F})$, and the corresponding estimate of θ^* is $\hat{\theta} \in \Theta$, where

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \hat{\Delta}(G(\theta), \hat{F}). \quad (5)$$

To choose the appropriate J , we compare a suitable criterion, or estimate of $E_F[\Delta(G(\hat{\theta}), F)]$, for each model. This way, multiple models may be compared to choose the one that best approximates F .

Furthermore, once a model has been chosen, the quality of the fit must be evaluated. Since $\hat{\Delta}(\hat{\theta})$ is a random variable, we may test the likelihood of it occurring under its limit distribution.

3.2 Histogram

A key difficulty in model selection problems is finding an appropriate discrepancy measure. Although several are available, some are more appropriate than others. For example, computing \hat{F} based on samples from n packets requires that n data pairs be stored. This is not possible due to the lack of resources within the network. Furthermore, with such a large amount of data being generated, transferring this data to a remote processor would use a non-negligible amount of the network bandwidth. Lastly, a fitting algorithm is needed to minimize $\hat{\Delta}(G(\theta), \hat{F})$, and discrepancy measures which fit at each point in the distribution function lead to complex algorithms. To mitigate these problems, we propose to use a histogram when storing, transferring, and fitting the data. That is, we approximate \hat{F} with a histogram \hat{H} , with K bins, of the observations.

Given the number of thresholds K , we construct the histogram in the following manner. The bin thresholds are fixed, and are denoted by the random vector

$$b = (b_1, b_2, \dots, b_K)^T \quad (6)$$

where the k -th bin threshold is denoted by b_k , and where $0 < b_1 < b_2 < \dots < b_K < \infty$. Furthermore, let $(X_n, n \geq 1)$ be a sequence of $\{0, 1\}^K$ -valued random vectors, such that

$$X_n = (X_{n,1}, \dots, X_{n,K})^T$$

and

$$X_{n,k} = \begin{cases} 1 & W_{\tau_n-} + S_n \leq b_k \\ 0 & \text{otherwise} \end{cases}.$$

Then, $X_{n,k}$ is the indicator of the event that the number of bytes in the queue, after the n -th packet arrival, exceeds the threshold b_k . Let \hat{H}_n be the histogram at time n , where

$$\hat{H}_n = (\hat{H}_{n,1}, \dots, \hat{H}_{n,K})^T$$

and

$$\hat{H}_{n,k} = \frac{1}{n} \sum_{i=1}^n X_{i,k}. \quad (7)$$

The histogram \hat{H}_n approximates \hat{F}_n at precisely K points on the distribution function. Furthermore, because the traffic process is stationary and ergodic, then $\hat{H}_n \xrightarrow{a.s.} H$. Also, because it is possible to construct regeneration points of the process, where the lengths of the regenerative cycles are i.i.d. with finite variance, then

$$\sqrt{n} (\hat{H}_n - H) \xrightarrow{d} N(0, C) \quad (8)$$

for some covariance matrix $C \geq 0$. The importance of this result is the fact that the error $(\hat{H}_n - H)$ is asymptotically normal with zero mean and covariance $\frac{1}{n}C$, and therefore we can construct maximum likelihood estimators that minimize this asymptotic error.

3.3 Maximum Likelihood Estimation

Since the error is asymptotically normal, the maximum likelihood estimator uses the discrepancy measure $\Delta(\theta) \equiv \Delta(G(\theta), H)$, where

$$\Delta(\theta) = V(\theta)^T C^{-1} V(\theta)$$

and

$$V(\theta) \equiv (G(\theta; b) - H).$$

Furthermore, we assume that C is positive definite (if not, an appropriate pseudo-inverse may be used instead). As before, an appropriate empirical discrepancy measure is $\hat{\Delta}_n(\theta) \equiv \hat{\Delta}_n(G(\theta), \hat{H}_n)$, where

$$\hat{\Delta}_n(\theta) = V_n(\theta)^T C_n^{-1} V_n(\theta)$$

and

$$V_n(\theta) \equiv (G(\theta; b) - \hat{H}_n).$$

We assume that C_n^{-1} exists and is an estimate of C^{-1} . Under our assumptions, we can define a sequence of estimators $(\hat{\theta}_n, n \geq 1)$, defined by (5), such that

$$\begin{aligned} \hat{\theta}_n &\xrightarrow{a.s.} \theta^* \\ \hat{\Delta}_n(\hat{\theta}_n) &\xrightarrow{a.s.} \Delta(\theta^*). \end{aligned}$$

Furthermore, since $(\hat{H}_n - H)$ is asymptotically normal by (8), then $(\hat{\theta}_n - \theta^*)$ and therefore $(G(\hat{\theta}_n; b) - G(\theta^*; b))$ will also be asymptotically normal since $G(\theta)$ and $\Delta(\theta)$ are continuous and differentiable functions with respect to θ . We can also view the

modeling error $V_n(\hat{\theta}_n)$ as asymptotically normal, and therefore minimizing $\hat{\Delta}_n(\theta)$ with respect to θ maximizes the likelihood of the asymptotic distribution.

To estimate C , we need to know the correlation structure of the traffic process. If full knowledge of the traffic process is available, then we can determine the traffic process regeneration times given by $(\gamma_m, m \geq 1)$. Since

$$C = \lim_{n \rightarrow \infty} \text{cov} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i, \frac{1}{\sqrt{n}} \sum_{j=1}^n X_j \right),$$

we can divide the sum into i.i.d. busy cycles and compute an unbiased estimator C_n of C . Of course, the difficulty with this approach is that the regeneration times are typically not known and cannot be determined from the process. (The exception to this is when Z_n is a renewal process, so that the regeneration times are simply the times when the queue is empty before the next packet arrival.)

As mentioned in [7], if we assume that $G(\theta^*) \approx F$ in the sense that modeling error is insignificant with respect to Δ , then a suitable criterion for evaluating $G(\hat{\theta}_n) \in \mathcal{G}_J$ is

$$E_F[\Delta(\hat{\theta}_n)] \approx E_F[\hat{\Delta}_n(\hat{\theta}_n)] - \frac{K}{n} + \frac{2(2J-1)}{n}$$

for large n . In other words, the decision to increase J and use a more sophisticated model depends on whether doing so will reduce the discrepancy measure by more than $\frac{4}{n}$. Furthermore, $\hat{\Delta}_n(\hat{\theta}_n)$ is approximately $\frac{1}{n}\chi_p^2$ distributed for large n , with $p = K - (2J - 1)$ degrees of freedom. Therefore, the likelihood of the empirical discrepancy measure can be estimated, so that the overall quality of the fit can be evaluated.

3.4 M-Block IID Estimation

As mentioned before, the key difficulty with the maximum likelihood approach is computing C_n because, in general, the regeneration times of the process are not available. Instead, we adopt an approach based on batch means [2]. This approach allows us to estimate the correlation structure and achieve similar performance to the maximum likelihood estimator.

In this case, we define the discrepancy measure to be

$$\bar{\Delta}_n(\theta) = V_n(\theta)^T W_n V_n(\theta)$$

where W_n is a weight matrix defined below. Using $\bar{\Delta}_n(\theta)$, we can construct a sequence of estimators $(\bar{\theta}_n, n \geq 1)$, defined by (5), such that $\bar{\theta}_n \xrightarrow{a.s.} \theta^*$ as before. However, in this case, the bias of the fit will depend on the bias of W_n as an estimator of C^{-1} .

Suppose we divide the process $(X_n, n \geq 1)$ into blocks of $M \geq 1$ packets, and average over each block. That is, we compute

$$\bar{X}_i = \frac{1}{M} \sum_{n=(i-1)M+1}^{iM} X_n.$$

Then, we can rewrite C as

$$C = \lim_{m \rightarrow \infty} A_m(0) + 2 \sum_{l=1}^{m-1} \binom{m-l}{m} A_m(l),$$

where

$$A_m(l) = \frac{M}{m-l} \sum_{i=1}^{m-l} \text{cov}(\bar{X}_i, \bar{X}_{i+l}).$$

Therefore, $A_m(l)$ is an estimate, based on the first m blocks of length M , of the autocovariance of \bar{X}_i separated by l blocks. The covariance matrix that is used in the fit is $\bar{C}_m = A_m(0)$, and therefore $W_n \equiv \bar{C}_m^{-1}$ for $n = mM$, m an integer. One can show that, by increasing M , \bar{C}_m converges to C_n , and so we can get arbitrarily close to the maximum likelihood estimator. However, it is desirable to keep M small so that we generate a significant number of samples.

The Sherman-Morrison formula [3] can be used to compute updates to the weight matrix W_n . Since

$$\bar{C}_m^{-1} = \left[\left(\frac{m-1}{m-2} \right) \bar{C}_{m-1} + \frac{1}{m} U_m U_m^T \right]^{-1}$$

for

$$U_m = (\bar{X}_m - \bar{H}_{(m-1)M}),$$

the weight matrix may be updated very simply after each block.

As before, we use the same criterion as in the maximum likelihood estimator. To evaluate the quality of the fit, we adopt the following approach. We can write, for $n = mM$ where m is an integer,

$$C_n \geq \bar{C}_m + \frac{2(m-1)}{m} A_m(1).$$

Then,

$$\begin{aligned} C_n^{-1} &\leq \left[\bar{C}_m + \frac{2(m-1)}{m} A_m(1) \right]^{-1} \\ &= W_n - E_n. \end{aligned}$$

where E_n is a matrix that can be computed from the Sherman-Morrison formula, and depends only on the matrices W_n and $A_m(1)$. The matrix E_n provides an estimate of the biasing of estimates due to the positive correlation in the measurements. Therefore, when we compute $\bar{\Delta}_n(\bar{\theta}_n)$, we can use the matrix E_n to test the validity of the fit since

$$V_n(\bar{\theta}_n)^T C_n^{-1} V_n(\bar{\theta}_n) \leq \bar{\Delta}_n(\bar{\theta}_n) - V_n(\bar{\theta}_n)^T E_n V_n(\bar{\theta}_n).$$

In this manner, we can estimate the quality of the fit by comparing the upper bound above to the $\frac{1}{n} \chi_p^2$ distribution, as before. By computing E_n in tandem with W_n , we provide an implementable and robust algorithm where the parameter M may be adjusted, based on E_n , to provide the desired fit.

4 Simulation

In this section, we motivate our approach with a simple example. Suppose the traffic process is Poisson(λ), with packet lengths that are exponentially distributed with rate μ . Then, $F \in \mathcal{G}_1$, and so

$$F(x) = 1 - e^{-\mu(1-\rho)x}, \quad 0 \leq x \leq B$$

where $\rho = \frac{\lambda}{\mu R}$. If we assume that $\lambda = 840$ packets/sec, $1/\mu = 100$ bytes, $B = 10^4$ bytes, and $R = 10^6$ bytes/sec, then the loss rate is approximately $L \approx 10^{-6.95}$. For $K = 8$ bins, we tested the maximum likelihood estimator (MLE) and the M-Block IID estimator for block sizes $M = 1, 8, 64$. The results are presented in Figure 2.

The graph shows the averaged estimates of the loss rate for the first 600 seconds of 100 simulation runs for each estimator. Although the estimators have not yet converged, the effect of using batch means is clear from the figure. As M get larger, the fit using batch means gets closer to the MLE. Furthermore, even without using batch means, as is the case with $M = 1$, the estimates are still very close to the MLE. In fact, the variability of the estimates over time increases with M , so that choosing M small has a smoothing effect on the estimates.

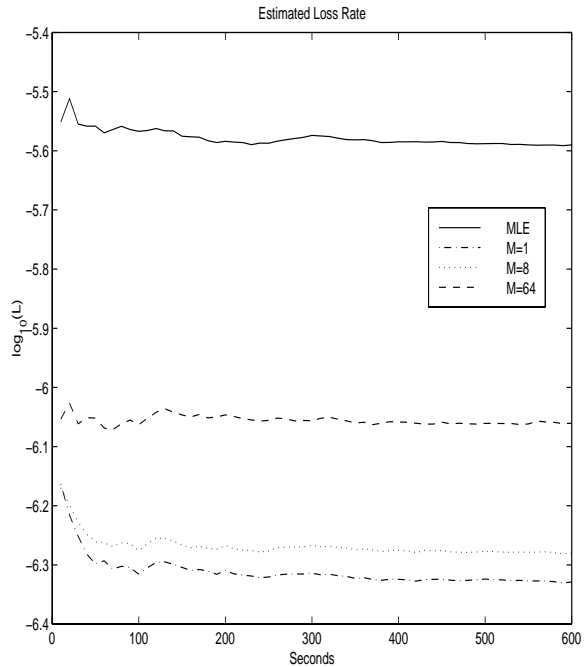


Figure 2: Estimated loss rate for MLE and M-Block IID estimators.

References

- [1] F. Baccelli and P. Bremaud. *Elements of Queueing Theory*. Applications of Mathematics. Springer-Verlag, 1994.
- [2] P. Bratley, B. Fox, and L. Schrage. *A Guide to Simulation*. Springer-Verlag, 1983.
- [3] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins Series in the Mathematical Sciences. The Johns Hopkins University Press, 1989.
- [4] H. Linhart and W. Zucchini. *Model Selection*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1986.
- [5] M. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins Series in the Mathematical Sciences. The Johns Hopkins University Press, 1981.
- [6] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1980.
- [7] M. Siler. Performance monitoring and call admission using measured buffer occupancy. Master's thesis, University of California at Berkeley, May 1998.