Buffer Occupation Probability of Trace-Driven Background Streams in Hybrid Simulation

Ben Lauwens and Bart Scheers
Royal Military Academy
CISS/TELE
30, Renaissancelaan
B-1000 Brussels
BELGIUM
E-mail: {ben.lauwens,bart.scheers}@rma.ac.be

Antoine Van de Capelle
Katholieke Universiteit Leuven
ESAT/TELEMIC
10, Kasteelpark Arenberg
B-3001 Heverlee (Leuven)
BELGIUM
E-mail: antoine.vandecapelle@esat.kuleuven.ac.be

Abstract—In this paper, numerical calculation schemes of the buffer occupation probability for a buffer which is fed by a large number of independent and stationary sources are evaluated. The resulting buffer occupation probability density function can be sampled in order to model the impact of a background traffic stream on the foreground traffic in a hybrid fluid-flow packet-event simulation. The calculation methods are based on the Large Deviations asymptotic or on the Central Limit theorem. An estimate combining both approaches is also studied. The so called Moderate Deviations scaling gives a Large Deviations result for a Gaussian source. The paper concludes with a quantitative comparison of the calculation schemes with simulations. Periodic ON/OFF-sources with a heavy tailed distribution are used as arrival processes for performance evaluations. The methods are also applied to the well known Bellcore LAN traffic traces.

I. INTRODUCTION

The basic events for most packet-event network simulations are the arrivals of packets in a network node and the departure of packets on a carrier. The model of the network consists of queues having a limited buffer size and of links characterized by their transmission capacity. High accuracy is gained but the simulation can become very slow when many events are generated. A large amount of events has to be generated for high bandwidth links or for networks with a large number of devices. Parallel packet-event simulations can increase the scalability of the latter but this approach can not be used for the former due to the sequential nature of a traffic flow on a link. The technique which uses events originating from the arrival or departure of packets, has to be abandoned if a higher simulation speed is required.

Fluid-flow based approaches for estimating the queueing behaviour of network traffic can be used to accelerate the simulation. The level of detail is however less and precision is sacrificed for a faster run-time. Subtle protocol dynamics can not be studied using this technique. Network emulation systems which interact with applications running on real networks, are also not feasible.

The combination of packet-event simulations and fluid-flow approximations can answer the challenge of getting packet level details in a reasonable amount of time. In the majority of scenarios, a simulation is carried out to get the performance parameters of a specific application when the data from that application is multiplexed with the background traffic on a network. A hybrid simulator separates the traffic in two classes. The packets of the foreground traffic, for which fine-grained performance details are needed, are simulated by an event-driven approach, while the background traffic, for which less detailed information is required, is approximated by a fluid-flow model.

These hybrid techniques in which packet-event simulation and fluid-flow approximations are combined, are a recent development. Some approaches use different simulators for the foreground traffic and the backgroun stream [1], others separate the network in packet parts and fluid parts [2]. Two simulators [3], [4] integrate the Monte-Carlo simulation and the fluid model into the same simulator. Both use the behaviour of the transport layer to model the fluid-flow. Traffic is simulated as an incompressible fluid, flowing among storage tanks (the buffers). Open-loop (UDP) traffic is generated by an Exponential ON/OFF traffic generator and the closed-loop (TCP) source model is a simplified version of TCP Reno. A complex approach is needed to synchronize the foreground traffic and the background stream and to model the interaction of the fluid-flow approximation with the packet-event simulation.

In this paper, a unified framework for the queueing behaviour of many fluid-flows is presented. The numerical calculation methods are independent of the higher layer mechanisms as TCP or UDP and the results can be computed using real traffic traces. The estimates based on the Large Deviations asymptotic, the Central Limit theorem and the Moderate Deviations limit, are described with the emphasis rather on the key ideas behind the approximations than on the rigorous mathematical details. The latter can be read in the papers found in the references. A detailed overview is given in a previous paper [5]. This work gives an uniform approach to the different estimates to obtain a fair analysis of the accuracy and the ease of computation. The reference for the benchmarks is a Monte-Carlo simulation with periodic ON/OFF-sources as arrival process. To test the feasibility of the calculations for a real traffic trace, the results of the methods are also shown for
the well known Bellcore LAN trace [6].

This paper is organized as follows. In section 2 the basic queueing model is introduced. An estimate based on the Large Deviations (LD) theory is developed in section 3 whereas section 4 details the Central Limit (CL) approximation. Section 5 gives an overview of the different approaches to Moderate Deviations (MD). All estimates are compared with a discrete-event simulation and the numerical results are presented in section 6.

II. QUEUEING MODEL

In order to make the discrete event simulation aware of the background traffic, the packet level network traces are transformed in a buffer occupation probability density function (PDF). This distribution is calculated by estimating the probability that the amount of work in a network device due to the fluid-flow data is higher than a certain value. It can be considered as a stochastic variable depending on one parameter, the number of traffic units in the buffer. For each network element, i.e. each queue, these probabilities have to be calculated separately because the capacity of the link and the composition of the traffic stream can be different.

Packet-based data networks are easily modeled by queueing systems. Data is parceled up into packets and these are sent over wires. At points where several wires meet, incoming packets are queued up, inspected, and sent out over the appropriate wire. When the total number of traffic units, i.e. bytes, packets or cells, reaches the buffer size, incoming packets are discarded. A basic FIFO queueing system can be quantified in discrete time using the modified Lindley’s recursion [7]

\[
Q_n = [Q_{n-1} + A_n - C_n]^B_0
\]

where \( [x]^B_0 = \max(\min(x, B), 0) \). \( Q_n \) can be interpreted as the amount of work in the queue in bin \( n \in \mathbb{Z}, A_n \) as the arrival process, i.e. the number of traffic units from packets arriving in timestep \( n \), \( C_n \) as the number of traffic units served during the interval \( n \), and \( B \) the buffer size expressed in traffic units. The bin duration is considered to be fixed and is called the fluid step \( \Delta t \).

To solve this recursion equation the following assumptions are made:

- \( C_n \) is independent of \( n \);
- the arrival process is stationary, i.e. \((A_{-n}, \ldots, A_0)\) has the same distribution as \((A_{-n-m}, \ldots, A_{-m})\) for every \( n \) and \( m \);
- the mean traffic rate is lower than the capacity, i.e. \( \mathbb{E} (A_n) \leq C \);
- the queue is empty at time \(-\infty\).

If these constraints are met, \( Q_n \) is also stationary and its distribution is called the steady state distribution of the queue size

\[
Q = \sup_{t \geq 0} (S_t - Ct)
\]

where \( Q \) is the stationary amount of work in the queue, \( C \) is the transmission capacity of the link, \( S_t \) is the stationary cumulative arrival function with \( S_{t,n} = \sum_{j=0}^{m-1} A_{n+j}, S_{0,n} = 0 \) and \( m = \frac{4}{\Delta t} \). The buffer occupation probability is the probability that \( Q \) is larger than \( B \).

\[
\Pr(Q \geq B) = \Pr \left( \sup_{t \geq 0} (S_t - Ct) \geq B \right)
\]

Only in certain cases this probability can be calculated analytically. For real traffic traces numerical approaches have to be used to find an estimation. All models that are detailed in this paper give other approximations for the buffer occupation probability.

The buffer occupation PDF \( f(Q = B) \) can directly be derived from the buffer occupation probability.

\[
f(B) = \Pr(Q \geq B) - \Pr(Q \geq B + 1)
\]

III. LARGE DEVIATIONS ASYMPTOTIC

Large Deviations theory has many applications and can be applied successfully to networks. The most interesting limiting regime considers what happens when a queue is shared by a large number of independent traffic flows.

A. Description

Consider the single server queue as before, with \( N \) sources and constant service rate \( C = cN \). Let \( A_n^{(i)} \) be the amount of work arriving from source \( i \) in fluidstep \( n \). Assume that for each \( i \), \( (A_n^{(i)}, n \in \mathbb{Z}) \) is a stationary sequence of random variables, and that these sequences are independent of each other. The stationary amount of work in a queue can be expressed in the terminology of previous section

\[
Q = \sup_{t \geq 0} (S_t - cNt)
\]

where \( A_n = \sum_{i=1}^{N} A_n^{(i)} \), \( S_{t,n} = \sum_{j=0}^{m-1} A_{n+j}, S_{0,n} = 0 \) and \( m = \frac{4}{\Delta t} \). So \( S_t \) is the stationary amount of work arriving at the queue in the interval \((-t, 0]\).

The buffer occupation probability can be estimated [9] by

\[
\Pr(Q \geq B) \approx \frac{1}{s_t \sqrt{2\pi \sigma_t^2}} e^{-Nt(b)}
\]

where \( B = Nb, S_t^{(i)} \) the stationary amount of work of one source with \( S_{t,n}^{(i)} = \sum_{j=0}^{m-1} A_{n+j}, S_{0,n}^{(i)} = 0 \) and \( m = \frac{t}{\Delta t} \), \( \Lambda_t^{(i)}(s) \) the cumulant generating function [6] of \( S_t^{(i)} \)

\[
\Lambda_t^{(i)}(s) = \log \mathbb{E} e^{sS_t^{(i)}}
\]

[1] A wireless network can be regarded as one generalized processor sharing queueing system for all the devices in transmission range. A paper is being prepared to be submitted in 2007 about an extension of the theory to wireless base stations and ad-hoc networks.

[2] First In First Out

[3] The original Lindley’s recursion (\( Q_t = (Q_{t-1} + A_t - C_t)^+ \)) is extended to take into account the limited buffer size. \( x^+ \) denotes the positive part of \( x \), i.e. \( \max(x, 0) \).

[4] For basic arrival distributions, i.e. Poisson traffic, the calculation can be done but real traffic tends to behave differently [8].

[5] The large buffer asymptotic [7] can not be used due to the possible small values of the buffer occupation level B.

[6] also known as the logarithmic moment generating function.
\( I(b) \) the rate function

\[
I(b) = \inf_{t \geq 0} \sup_{s \in \mathbb{R}^+} J_t(s, b)
\]

with \( J_t(s, b) = s(b + ct) - \sum_{i=1}^{n} \rho_i^s \Lambda_i^{(t)}(s), \rho_i^s = N_i^s/N \) and \( N^s \) the number of flows from class \( i \), and \( \sigma_i^2 \) the variance

\[
\sigma_i^2 = \frac{d^2 J_t(s, b)}{ds^2} \bigg|_{(t = \hat{t}, s = s_t)}
\]

\( s_t \) can easily be found as

\[
s_t = \arg_{s} \left( b + ct = \sum_{i=1}^{n} n_i^s \frac{d\Lambda_i^{(t)}}{ds} \right)
\]

and the expression for \( \sigma_i^2 \) simplifies to

\[
\sigma_i^2 = -\sum_{i=1}^{n} \rho_i^s \frac{d^2 \Lambda_i^{(t)}}{ds^2} \bigg|_{(t = \hat{t}, s = s_t)}
\]

A further simplification gives an ad-hoc approximation [10], which is exact for the case of Gaussian arrival processes \( \sigma_i^2 \approx \frac{2J_i(b)}{s_i^t} \). A final expression for the buffer flow probability can be obtained without the knowledge of the number of flows or the exact composition of the flows [11], [12]

\[
\Pr(Q \geq X) \approx \frac{1}{\sqrt{4\pi I(B)}} e^{-I(B)}
\]

Equation 6 will be called the Large Deviations Bahadur-Rao (LD BR) result.

### B. Remarks

The general idea behind formula (6) can be summarized in two principles:

- application of the principle of the Largest Term to move the \( \sup_{s} \) out of the probability;
- approximation of the tail of the distribution with an exponential, based on a large deviations result.

This shows that inaccuracy can occur in two ways [7]. The principle of the Largest Term\(^7\) is very accurate for a moderate traffic rate compared to the transmission capacity and provides a lower bound for higher traffic rates. The exponential curve underestimates the buffer flow probability for low traffic rates but overestimates for traffic rates near the link rate.

It might be expected that when bursty and smooth traffic are multiplexed on a link that the bursty traffic will be smoothed and the smooth traffic will be more bursty. In a router with a small number of inputs this can indeed happen. But in the many flows scaling regime\(^8\) this is not the case. In other words, the individual traffic flows do not depend on the traffic mix at the router as long as the queue empties regularly with high probability. This is known as decoupling [14]. An intuitive descriptor of the stochastic properties of an individual traffic stream, called the Effective Bandwidth \( \alpha_i(s) \) of the fluid-flow [15], can be introduced.

\[
\alpha_i(s) = \frac{1}{st} \Lambda_i(s)
\]

In the large deviations limit, it makes sense to talk about the effective bandwidth of a single flow through a network as long as in each network device the service rate is higher than the mean arrival rate. In that case the effective bandwidth of the departure flow at the last device will exactly be the same as the effective bandwidth of the arrival flow at the first device. The time parameter \( t \) corresponds to the buffer busy period before reaching a buffer occupation level \( B \). Smoothing of traffic will only occur in a buffer with size \( B \) at a timescale larger than the corresponding \( \hat{t} \). Only then it affects the effective bandwidth of the traffic flow [11]. This parameter is also important for the granularity of the samples from a real traffic trace. The sampling time has to be less than \( \hat{t} \) to give an accurate estimate. A value several orders less than \( \hat{t} \) will have a big impact on the performance of the calculation of \( \Lambda_i(s) \).

The space-scale \( s \) is a parameter for the degree of multiplexing [16]. More \( s \) approaches zero, more the multiplexing will be efficient and more the effective parameter will tend to the mean rate of the traffic stream at the appropriate time-scale. If \( s \) becomes large, the flows will not multiplex very well and the effective bandwidth will be close to the peak rate of the traffic stream at the corresponding time-scale. For a fix \( \hat{t} \), the effective bandwidth is strictly concave going from the mean rate to the peak rate of the traffic flow at this time-scale.

### IV. Central Limit Theorem

Heavy Traffic\(^9\) theory relies on the fact that the amount of traffic in a certain sampling interval due to the aggregation of many flows onto a link tends to approach a Gaussian distribution. This is an application of the Central Limit theorem and relies on the assumption that the variance of the component traffic streams is uniformly bounded.

#### A. Description

Denote the component traffic streams in a network by \( A^{(i)} \), the aggregation by \( A = \sum_{i=1}^{N} A^{(i)} \) and the stationary cumulative arrival process \( S_t \) with \( S_{t,n} = \sum_{j=1}^{m} A_{n+j} \), \( S_{0,n} = 0 \) and \( m = \frac{t}{\Delta t} \). \( S_t \) can be seen as a stationary vector

\( ^7\)The principle of the Largest Term states qualitatively that rare events happen only in the most probable way.

\( ^8\)Experimental setups have shown that already for a rather small number of flows \( N = \beta \) for two traffic classes, the departure flows are mostly decoupled [14].

\( ^9\)Heavy Traffic limit is mostly used for the Diffusion Approximation which has a rather bad performance for a strong correlated source. In this paper the term Heavy Traffic is also used for the Many Flows Central Limit Theory.
with index $t$ and where every component of this vector has a Gaussian distribution

$$S_t = N(\mu_t, \sigma^2_t)$$

where $\mu = \mathbb{E}(A)$ is the mean traffic rate of the aggregated traffic stream, $\mathbb{E}(S_t) = \mu_t$ and $\text{var}(S_t) = \sigma^2_t$.

It can be shown that as the aggregation takes place, that virtually all performance measures of multiplexers and switches in a network individually or in combination, including loss, delay and jitter, must approach the performance measures of a communication system carrying Gaussian traffic with the same second order statistics [17]. If sufficient independent sources of traffic contribute to a network, it is reasonable to analyze the network using a Gaussian model, matching the second order statistics of the model to the real traffic data.

Consider the single server queue as before, with constant service rate $C$ and the aggregated traffic with a Gaussian distribution. This can be reformulated in the basic queueing model

$$Q = \sup_{t \geq 0} (S_t - C t)$$

The buffer occupation probability can then be estimated by [18], [19]

$$\Pr(Q \geq B) \approx \Psi\left(\frac{B + (C - \mu)t}{\sqrt{\sigma^2_t}}\right)$$

where $\Psi(w) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-z^2/2) \, dz$, the standard normalized Gaussian tail function, and

$$\hat{t} = \sup_{s \geq 0} \frac{\sigma^2_t}{(B + (C - \mu)t)^2}$$

Equation 10 is noted as CL.

B. Remarks

The principle of the Largest Term suggests that the previous equation is a good lower bound approximation. This seems at first sight a contradiction. How can the probability that the work in queue exceeds the buffer size be small for a queue filled with traffic which has a mean rate close to the transmission capacity? The answer is similar to the interpretation of the space-scale $s$ of the Large Deviations Asymptotic. A low $s$ means a high multiplexing gain and the effective bandwidth is close to the mean rate. The same phenomenon happens in the Heavy Traffic case. When many independent flows are aggregated and their mean is less than the transmission capacity, the mixture has a better performance than the individual traffic streams even for a limited buffer size. The supremum time-scale $\hat{t}$ is however in the case of a small buffer not necessary unique and multiple parallel paths to overflow can exist.

If the buffer size becomes zero, a convenient approximation [19] can be used

$$\Pr(Q \geq 0) \approx 2 \Pr(A > C) = \sqrt{\frac{2}{\pi \sigma^2}} \int_{C}^{\infty} e^{-\frac{(x-C)^2}{2\sigma^2}} \, dx$$

where $\mu = \mathbb{E}(A)$ is the mean traffic rate of the aggregated traffic stream and $\sigma^2 = \text{var}(A)$ the variance of the aggregated traffic stream. This heuristic is motivated by the fact that during a busy period the buffer goes down roughly as often as it goes up. This can be used to scale equation 10 and the result is noted as (CL sc).

V. Moderate Deviations Estimate

The estimate based on the Central Limit Theorem is appealing because it leads to parsimonious models, i.e. all you need to know about an arrival process is its mean and covariance structure. Large Deviations theory says that the calculation of an integral can be replaced by finding a supremum. Both approaches can be analytically intractable: the first because integrals are needed to calculate probabilities and the second because the full statistical characteristics of a traffic flow are unwieldy. It is tempting to combine both techniques [20]: by the Central Limit theorem, the arrival process can be modeled as a Gaussian process; by Large Deviations theory, it is sufficient to look at the exponential behaviour of the tail of the distribution.

A. Description

The approach in [20] quantifies with a burstiness parameter $\beta$ the relation between the Central Limit theorem ($\beta = 0$) and the Large Deviations theory ($\beta = 1$). Moderate Deviations scaling concerns a range of scales between these extremes. A typical traffic flows has bursts at all scales $0 \leq \beta \leq 1$, with larger bursts (large $\beta$) less frequently than smaller bursts (small $\beta$). The parameter $\beta$ plays the dual role: it measures both burst size and burst frequency.

Let $N$ be the number of stationary, identically distributed arrival processes $A^{(i)}$ with mean rate $\bar{\mu}$, $S_t$ the stationary cumulative arrival process with $S_{t,n} = \sum_{j=0}^{n-1} A_{t+j}$, $S_{0,n} = 0$ and $m = \frac{1}{N}C$, $B$ the transmission capacity and $B$ the number of traffic units in the buffer. The following equation can be found [20]

$$\frac{1}{N^\beta} \log \Pr\left(\frac{N^{1+\beta}}{N} \left(\frac{S_t}{N} - \bar{\mu} \frac{t}{N}\right) \geq \bar{b}\right) = -\inf_{\hat{t} \geq 0} \frac{(\hat{b} + \hat{c}t)^2}{2\hat{\sigma}^2_t}$$

where \( \hat{b} = B/N^{1+\beta}, \quad \hat{c} = (C - \mu N)/N^{1+\beta} \) and $\hat{\sigma}^2_t = \text{var}(S_t)/t$ are the moments of a large deviation. If $\beta = 0$, the probability becomes $\Pr(N \bar{\mu} N t, \hat{\sigma}^2_t N)$ with an exponential approximations and if $\beta = 1$, the result resembles a Large Deviations result for a Gaussian distribution [10].

A straightforward estimate can be derived

$$\Pr(Q \geq B) \approx \frac{1}{4\pi \sqrt{B}} e^{-1/(B)}$$

where $\sigma^2_t = \text{var}(S_t)$.

$$\hat{t} = \sup_{s \geq 0} \frac{\sigma^2_t}{(B + (C - \mu)t)^2}$$

For a Gaussian arrival process, $\Lambda_t(s) = s \bar{\mu} t + \frac{1}{2} s \sigma^2_t$ and the result follows directly by taking the supremum $s_t = \arg \sup_{s \geq 0} (s (\hat{b} + \hat{c}t) - \Lambda_t(s))$.
and $I(B) = (B + (C - \mu) \hat{t})^2 / 2\sigma_t^2$. This result is independent of $N$ and $\beta$. It is clear that this approximation is only valid, if the input process, $C$ and $X$ stand in a certain relation but this estimate can always be calculated. Equation 11 without prefactor is noted as (MD) and with prefactor (MD BR).

B. Remarks

Equation 11 is an estimate for an isolated queue. For a network of queues the parameter $\beta$ can be interpreted as a low-pass filter [20]. The loss rate at a queue of scale $\beta$ will be of the order of $L^{-\beta}$. Thus the loss rate for a fluid flow through a network will be dominated by the buffer occupation probability at the queue along the path with the smallest scale $\beta_{\text{min}}$. This can be called the bottleneck link of the flow. By the low-pass filter result, traffic is essentially unchanged at scales less than or equal to $\beta_{\text{min}}$ until it reaches the bottleneck link. The approximation for loss rate can be used at the bottleneck link without taking into account any smoothing. The loss rate of a flow through a network can be approximated by estimating the loss rate at the bottleneck link.

The Maximum Variance Asymptotic finds the same equation of the buffer occupation probability for a general class of Gaussian processes with stationary increments [22], including a large class of long-range dependent processes with $\kappa = \lim_{t \to \infty} \log(\sigma_t^2) / \log(t)$, assuming that this limit exists. $\kappa$ can not be greater than 2 from the stationary increment property. The interval $\kappa \in [0, 2]$ covers the majority of non-trivial Gaussian processes with stationary increments. For a sufficiently large buffer occupation level $B$, $\log (\sigma_t^2/(B + (C - \mu) \hat{t}))$ is strictly concave on the interval $[(\kappa - \sqrt{\kappa^2}/2) B/(2 - \kappa)(C - \mu), (\kappa + \sqrt{\kappa^2}/2) B/(2 - \kappa)(C - \mu)]$, and there is a unique index $\hat{t}$ where it attains its maximum. The buffer flow probability and the characteristic time-scale $\hat{t}$ can be computed by performing a simple local search algorithm starting at $\kappa B/(2 - \kappa)(C - \mu)$. Even for fairly small numbers of $X$ this can be used because $\sigma_t^2/(B + (C - \mu) \hat{t})^2$ is usually of a distinctly uni-modal shape. For small $B$ the minimizer may not be unique. The formula for the calculation of $\Pr(Q \leq B)$ however does not depend on the uniqueness of $\hat{t}$.

The same scaling as for the Heavy Traffic case can be applied to (MD) [19] resulting in an estimate noted by (MD sc).

VI. NUMERICAL RESULTS

To validate the accuracy of the different methods, the results of a Monte-Carlo simulation are used as a reference. The numerical results described here are an example illustrating the findings of several experiments with different traffic processes.

A FIFO queue is fed by 2 different periodic ON/OFF-sources. The period is $p_1 = 29$ traffic units for the first and $p_2 = 31$ for the second. The peak rates of these sources in traffic units per time slot are $a_1 = 10$ and $a_2 = 1$. The durations of the ON states are $a_{11} = 3$ time units and $a_{21} = 1$ time units. The basic stream consists of $n_{11} = 10$ flows of the first source and $n_{21} = 4$ flows of the second. For the experiment $N = 80$ basic streams are multiplexed. The simulation results are given with a 95% confidence interval.

The calculation for both $s_i^2 = \text{var}(S_i)$ and $s_i^2 = \text{var}(S_i)$ is straightforward for a periodic ON/OFF-source. Let $a$ be the peak number of traffic units per time unit, on the number of time units that the source is generating traffic units and $p$ the period of the source.

\[
\sigma^2_t = \text{var}(S_i) = e^{a \cdot \text{Ons. m. a} \cdot s_i^2(S)}
\]

where $r = (t - 1) \mod{p} + 1$ and $n = \lfloor (t - 1)/d \rfloor$. Once the values for one period are calculated the others can be obtained directly.

Figures 1 and 2 show the buffer occupation probability as a function of $B$ for different values of $C$. The load of the link $\mu / C$ where $\mu$ is the mean rate of the aggregated data stream varies between 98% and 84%.

For a high load, the Heavy Traffic estimate without rescaling approximates very well the simulated buffer occupation probability. The scaled Moderate Deviations approximation over-estimates the probability for all values of $B$. The other methods are acceptable for larger values of $B$. The Moderate Deviations and Large Deviations Bahadur-Rao estimates have an unphysical value for $B = 0$. The exponential approximation for the buffer occupation probability becomes higher than 100%.

The Large Deviations Bahadur-Rao refinement fits the simulated probability for all values of $B$ when the load is moderated. For high values of $B$ it is actually a lower bound for the buffer occupation probability.

The scaled versions of the Central Limit approximation and
the Moderate Deviations estimate don’t give consistent results when the load changes. These two methods are only acceptable for the calculation of the buffer occupation distribution in a limited range of values for $B$ and $C$.

The Moderate Deviations Bahadur-Rao refinement approximates for low values of $B$ the Large Deviations result and for high values the Central Limit estimate. But for low $B$ the Central Limit approximation is better and for high $B$ the Large Deviations asymptotic is closer to the simulated buffer occupation probability. This behaviour is directly related to the degree of multiplexing and the value of $s_j$.

Without knowledge of the distribution of the source the following empirical rule gives a tight fit to the simulation:

- for the MD BR estimate close to the LD BR asymptotic, the CL approximation is used and the loss rate will have an impact on the departure process;
- for the MD BR estimate close to the CL approximation, the LD BR asymptotic is the best estimate and the effective bandwidth of the traffic stream is unaltered after the queue.

Figure 3 shows the buffer occupation probability of the simulation and the estimate based on the empirical rule for the different values of $B$ and $C$. The same multiplexing scheme is used for figure 1 and 2. The estimates are calculated in a fraction of the time of the simulated results and there are no significant differences between both.

The Large Deviations calculation is the most computational intensive but the value of $\hat{t}$ is always less or equal than the value of $t$ for the Moderate Deviations or the Central Limit estimate which is easier to calculate. This limits the search procedure and speeds up the calculation of the estimates.

Figure 4 shows the buffer occupation probability of a discrete-time simulation and the Large Deviations estimate for the well known Bellcore LAN trace [6]. The value of the transmission capacity varies between 8.0 Mbps and 12.5 Mbps. The estimates are very similar to the simulations. The step behaviour of the simulated results is due to the distribution of the packet sizes in the Bellcore LAN trace. The complementary cumulative packet size distribution is shown in figure 5. A large step means a high probability that the corresponding number of traffic units are in the buffer. If a packet size is very probable and the number of packets in the buffer is small, the buffer occupation probability for the number of traffic units, that equals a highly probable packet
size, will be very high. In a hybrid simulation, this behaviour can be included by sampling the distribution of the packet sizes to fill up the sampled buffer occupation level. When multiple traffic traces are multiplexed, this step behaviour will fade out and the Large Deviations asymptotic will closely resemble the simulated results.

VII. CONCLUSION

In this paper, the buffer occupation probability for a single FIFO queue is estimated by different formulas ranging from a Central Limit approach to a Large Deviations limit. For a high load the Central Limit theorem gives the best approximation. For a moderate load where the degree of multiplexing is less, only the Large Deviations Bahadur-Rao refinement gives an accurate estimate for a large range of traffic units in the buffer. This is shown by Monte-Carlo simulations of periodic ON/OFF-sources. The Large Deviations asymptotic for the Bellcore LAN trace is also calculated. The estimated buffer occupation probability follows closely the behaviour of a discrete-time simulation of the same traffic trace.

In a future work the performance of these formulas will directly be evaluated in a hybrid fluid-flow packet-event simulation. The results can be compared to an event-driven simulation of both foreground and background traffic. The loss rate for the foreground traffic stream is however so small that importance sampling is needed. The most common importance sampling methods for networks are based on Large Deviations theory, so that favoring the LD estimates. This paper gives a validation domain for the applicability of the fast simulation by importance sampling.

Topics for further research also include:

- extending the FIFO queue to priority queueing and generalized processor sharing;
- adapting the model to include wireless base-stations and ad-hoc networks.

A fluid-flow buffer indicator, the buffer occupation probability density function, can be found when many traffic sources are aggregated but the performance measures of the resulted traffic stream have not always a Gaussian behaviour. The degree of multiplexing plays an important role and determines whether a Central Limit or a Large Deviations estimate is applicable.

REFERENCES