Characterization of Background Traffic in Hybrid Network Simulation

Ir. Ben Lauwens, Dr. Ir. Bart Scheers
Royal Military Academy
CISS-TELE
30, Renaissancelaan
B-1000 Brussels
BELGIUM

Dr. Ir. Antoine Van de Capelle
Katholieke Universiteit Leuven
ESAT-TELEMIC
10, Kasteelpark Arenberg
B-3001 Heverlee (Leuven)
BELGIUM

{ben.lauwens, bart.scheers}@rma.ac.be antoine.vandecapelle@esat.kuleuven.ac.be

ABSTRACT

The information transfer across the battlespace is expanding at an ever increasing pace. Simulations have to be carried out to ensure the quality of service of the different applications (sensors, command and control systems, voice communications, ...) all integrated in a dynamic network environment. Two approaches are common: discrete event simulation and fluid approximation. A discrete event simulation generates a huge amount of events for a full-blown battlefield communication network resulting in a very long run-time. The usability of this simulation technique is very limited for a rapid changing combat theater. A faster fluid based approximation lacks the accuracy wanted for a realistic simulation. 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 model. A research area of this hybrid approach is the characterization of the fluid data flow. Classical statistics are inappropriate for real network data due to the heavy-tailed behaviour and the self-similarity of network traffic. This paper discusses the estimates based on the Large Deviations asymptotic, the Central Limit Theorem and a range of scaling laws in between both limits, a so called Moderate Deviations approximation. Based on real network traces the loss probability of a data stream can be estimated for all models. The results of this protocol independent analysis can be incorporated in a discrete-event simulation allowing a considerable reduction of computation time and memory consumption. The desired fine-grained details of the foreground application behaviour can be obtained while taking into account the background traffic. In this article different characterization methods of background traffic in a hybrid network simulation are presented and the results are compared with discrete-event simulations.

1.0 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 links characterized by their transmission capacity. A high accuracy is gained but the simulation becomes very slow due to the computational cost when many events are generated. A large amount of events can be generated for high bandwidth links or for networks with a great number of devices. A parallel discrete-event

Lauwens, B.; Scheers, B.; Van de Capelle, A. (2006) Characterization of Background Traffic in Hybrid Network Simulation. In *Dynamic Communications Management* (pp. 2-1 – 2-22). Meeting Proceedings RTO-MP-IST-062, Paper 2. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.asp.

simulation can increase the scalability of the size of a network but can not accelerate the simulation of a high capacity link due the sequential nature of a traffic flow on a link. The technique of events generated by packets has to be abandoned if a higher simulation speed is required.

Fluid based approaches for estimating the behaviour of network traffic can be used to simplify the simulation. However the level of detail available to the approximation is less and accuracy is sacrificed for a faster run-time. Subtle protocol dynamics can not be studied using this technique. Network emulation systems that interact with real applications running on real networks, are also out of reach.

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

The amount of abstraction introduced by the fluid model determines the precision of the calculation. There are many papers, i.e. [3], considering the transport layer behaviour of traffic streams as a basic approximation of the data flow in a real network. The flow conservation laws and the congestion control techniques used in TCP allow to write down a system of equations describing the traffic behaviour. The traffic source has a limited number of parameters that have to be estimated to get a realistic behaviour of a specific application on a network and this has to be done for each source putting data into the network. In the case the individual traffic streams are not known or too complex to demultiplex, as for a high bandwidth traffic trace, or the traffic flow can not be modeled by the used equations, as for circuit emulated services over Ethernet[1], the transport layer model is not adequate. In this paper the link layer behaviour of traffic is modeled considering the same network elements as the discrete-event network simulation, i.e. buffers and links. The influence of the transport layer is embedded in the analyzed traffic trace. The exact behaviour of the background traffic is secondary to how it influences the foreground packets. A parameter fitting of a traffic source to a traffic trace is not necessary. Only the interaction of the background stream with the queue levels in the event driven simulation has to be modeled. This is done by introducing the buffer flow probability, a one-dimensional stochastic variable that describes the probability that the amount of work in the queue is bigger than a certain value. This technique evades the challenge of modeling the interaction between the packet streams and the fluid flows.

Related work. The hybrid technique in which packet event simulation and fluid flows approximations are combined, is a recent development. Some approaches use different simulators for the foreground traffic and the background stream [22], other separate the network in packet level parts and fluid parts [13]. There exist two simulators [19, 15] that integrate the Monte-Carlo simulation and the background model into the same simulator. Both use the behaviour of the transport layer model to model the fluid flow. Traffic is simulated as an incompressible fluids, 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, not always well documented approach is needed to synchronize the foreground traffic and the background stream and to model the interaction of the the fluid flow approximation with the event-driven packet simulation and vice versa.

Literature. The approach in this paper has inherently a many flows paradigm, which is quite acceptable for the majority of the networks. A solution of the basic queue equation for a general traffic trace is impossible [11]. An asymptotic approximation is the best one can get. In the literature two main directions

[1]The emulation of SDH circuits over Ethernet or IP are hot topics in the carrier world where the Internet Service Providers like to migrate to one link layer.

are developed. The first considers the aggregation of many flows as a Gaussian process and is based on the Central Limit Theorem [2, 4]. The second estimates that the possibility of a buffer overflow is a rare event and uses the Large Deviations theory [9, 10, 18, 11]. Both can have different scaling regimes. The most important are the fast time scaling and the many sources scaling [11]. Fast time scaling looks at a speeded-up version of the arrival process of one basic flow whether the many sources scaling considers the sum of N independent copies of the basic flow. The former gives bad results when the buffer size is small compared to the transmission capacity. This is not acceptable and only the many flows scaling [25, 11] is examined for both the Central Limit theorem and the Large Deviations theory. Both limits have however their strength and their weakness in regard to a traffic trace if no assumptions can be made about the distributions of the individual flows. Several attempts where made to combine both and get "the best of both worlds": analytically tractability and minimal statistical information to be extracted from a traffic trace [26]. The Maximum Variance Asymptotic [16, 6, 5, 7], the Most Probable Path for Gaussian Inputs [27, 1] and the Moderate Deviations scaling [11, 26] answer this problem from a different viewpoint. After translating the different theories in one language the same kind of equation is found but the application domain is very different.

Contribution. In this paper, a unified framework for the queueing behaviour of many flows fluid traffic models is presented. These model are all independent of the higher layer mechanics as TCP or UDP and the results are 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 bibliography. This work gives an uniform approach to the different estimates to obtain a fair analysis of the accuracy and ease of computation. The appropriate domains for each method is examined. The reference for the benchmarks is a Monte-Carlo simulation. Both synthetic trace, periodic on-off sources, and real trace, the famous Bellcore dataset [17] and the Star Wars MPEG-1 trace file [23], are used. The results are presented as optimal domains for each technique based on workload, capacity and buffer size. This is the first systematic evaluation of estimates for background fluid models independent of the higher-layer mechanics and integrating the live behaviour of real traffic traces.

Organization. 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 Theorem (CLT) 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.

2.0 BASIC QUEUEING MODEL

In order to make the discrete event simulation aware of the background traffic, the packet level network traces are translated in a buffer flow occupation probability. This queue indicator gives the probability that the amount of work in a network device due to the fluid data is more than a certain value. It can be considered as a stochastic variable depending on one parameter. For each network element, i.e. each queue, this probability has to be calculated separately because the capacity of the link, the buffer size 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[2] meet, incoming packets are queued up, inspected,

[2] 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.

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.

The basic FIFO[3] queueing system can be quantified using the modified[4] Lindley's recursion [11]

$$Q_n = [Q_{n-1} + A_n - C_n]_0^B \tag{1}$$

where $[x]_0^B = \max(\min(x,B),0)$. Q_n can be interpreted as the amount of work in the queue just after time $n \in \mathbb{Z}$, A_n as the arrival process, i.e. the number of traffic units which arrive in the interval $(n-1,n)$, C_n as the capacity of the link, i.e. the number of traffic units served between $n-1$ and n, and B the buffer size also expressed in traffic units.

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. $\mathbb{E}(A_n) \leq C_n$;
- the queue is empty at time $-\infty$.

The capacity of a network link can be considered to have a constant value[5]. Measurements of traffic traces on real networks have shown that real traffic can be long-range dependent [17] but is mostly stationary. If the mean traffic rate is higher than the capacity, the queue will never empty and the recursion is not stable. The emptiness of the queue at infinity is implicit, i.e. the boot up of the device.

With these constraints,Q_n is also stationary and its distribution is called the steady state distribution of the queue size

$$Q = \sup_{t\in\mathbb{N}} (S_t - Ct) \tag{2}$$

where C is the tranmission capacity, $S_t = \sum_{n=-t+1}^{0} A_n$ and $S_0 = 0$. The queue indicator is the probability that Q is larger than X

$$\Pr(Q \geq X) = \Pr\left(\sup_{t\geq 0} (S_t - Ct) \geq X\right) \tag{3}$$

Only in certain cases[6] 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, try do find an exact asymptotic for the buffer flow probability.

To test the accuracy of the approximations an event-driven simulation of the queue with the same arrival process as the fluid traffic is used. The result of this simulation described in the literature [18] is however not the buffer overflow probability (BOP), i.e. the buffer flow probability with $X = B$, but the loss rate (LR) [18] of packets expressed in traffic units.

$$LR = \frac{\mathbb{E}([Q + A - (C+B)]_0)}{\mathbb{E}(A)} \tag{4}$$

[3]First In First Out

[4]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. $max(x,0)$.

[5]In a real network the capacity can be variable but the speed of change is considered to be much slower than the characteristic time-scales that determine the queueing behaviour.

[6]For basic arrival distributions, i.e Poisson traffic, the calculation can be done but real traffic tends to behave differently [21].

Both the *BOP* and the *LR* calculated for the different approaches and they are compared with the event driven simulation. The *BOP* is obtained directly from an event driven simulation by measuring the total time that the buffer is in overflow compared to the total simulation time.

3.0 LARGE DEVIATIONS

Large Deviations theory has many applications and can successfully be applied to networks. The most interesting[7] limiting regime considers what happens when a queue is shared by a large number of independent traffic flows (also called sources).

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 at time n. Assume that for each i, $\left(A_n^{(i)}, n \in \mathbb{Z}\right)$ is a stationary sequence of random variables, and that these sequences are independent of each other but identically distributed. This can be rephrased in the terminology of previous section

$$Q = \sup_{t \in \mathbb{N}} (S_t - cNt) \tag{5}$$

where $A_n = \sum_{i=1}^{N} A_n^{(i)}$, $S_t = \sum_{n=-t+1}^{0} A_n$ and $S_0 = 0$. So S_t is the total amount of work arriving at the queue in the interval $(-t, 0]$.

The buffer flow probability can be considered as a queue indicator in the event driven simulation

$$\Pr(Q \geq X) = \Pr\left(\sup_{t \in \mathbb{N}} (S_t - cNt) \geq xN\right) \tag{6}$$

here is $X = xN$. The principle of the Largest Term that plays an important role in general Large Deviations theory [11], can be applied to the previous expression

$$Pr(Q \geq X) = \sup_{t \in \mathbb{N}} \Pr(S_t - cNt \geq xN) \tag{7}$$

The Chernoff's bound [25] for the upper bound and the Cramér's theorem [25] for the lower bound can be used

$$-I(x+) \leq \liminf_{N \to \infty} \tfrac{1}{N} \log(\Pr(Q \geq X)) \tag{8}$$

$$\limsup_{N \to \infty} \tfrac{1}{N} \log(\Pr(Q \geq X)) \leq -I(x-) \tag{9}$$

where $S_t^{(1)} = \sum_{n=-t+1}^{0} A_n^{(1)}$, $S_0^{(1)} = 0$, $\Lambda_t^{(1)}(s) = \log \mathbb{E} e^{sS_t^{(1)}}$, the cumulant generating function[8] of $S_t^{(1)}$, $\overline{\Lambda}_t^{(1)}(x+ct) = \sup_{s \in \mathbb{R}^+} \left(s(x+ct) - \Lambda_t^{(1)}(s)\right)$, the convex conjugate[9] of $\Lambda_t^{(1)}(s)$ and $I(x) = \inf_{t \in \mathbb{N}} \overline{\Lambda}_t^{(1)}(x+ct)$, the rate function. The calculation is based on following assumptions:

- for all t, $\Lambda_t^{(1)}(s)$ is finite for s in a neighborhood of the origin;
- $\Lambda^{(1)}(s) = \lim_{t \to \infty} \frac{1}{t} \Lambda_t^{(1)}(s)$ exist and is finite and differentiable, for s in a neighborhood of the origin.

[7] The simpler large buffer asymptotic [11] can not be used due to the possible small values of X.
[8] also known as the logarithmic moment generating function
[9] also known as the Fenchel-Legendre transform

The general ideas behind this formula can be summarized in two principles:

- application of the principle of the Largest Term to move the $\sup_t$ 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 [11]. The principle of the Largest Term 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 approaches nicely the buffer flow probability for low traffic rates but overestimates for traffic rates near the link rate.

As does Cramér's theorem, this result involves both an upper and a lower bound. If $\Lambda_t^{(1)}(s)$ is continuous in s for each t, then the two bounds agree and a straightforward approximations can be obtained

$$Pr(Q \geq X) \approx \mathrm{e}^{-N\left(s_{\hat{t}}(b+c\hat{t})-\Lambda_{\hat{t}}^{(1)}(s_{\hat{t}})\right)} \tag{10}$$

where $s_t = \arg\sup_{s\in\mathbb{R}^+}\left(s(b+ct)-\Lambda_t^{(1)}(s)\right)$ and $\hat{t} = \arg\inf_{t\in\mathbb{N}}\left(s_t(b+ct)-\Lambda_t^{(1)}(s_t)\right)$.

There was assumed that $t^{-1}\Lambda_t^{(1)}(s)$ converges to a limit $\Lambda^{(1)}(s)$, and yet $\Lambda^{(1)}(s)$ does not appear anywhere in the result. This assumption is just a easy way to control the tails so that there are no surprises in the distribution of $S_t^{(1)}$ for large t. This is only needed to prove the upper bound but not the lower bound. All the measured traffic traces allow the convergence of this limit.

The principle of the Largest Term is used for the deduction of both the upper and the lower limit. The motivation behind this principle is the idea that to estimate the probability of a rare event, only the most likely path in which this event can occur, has to be considered. For the lower bound, the fixed time $\hat{t}$ is the most likely duration starting from an empty buffer that the workload in the queue exceeds the value x. The upper bound shows that the probability of a workload exceeding a level x over a time period $t \neq \hat{t}$ is negligible on a logarithmic scale. The principle of the Largest Term can be used to obtain directly a numerical estimate for the buffer flow probability but the calculation for real traffic traces is both memory and processor intensive.

The assumptions that the A_t are independent and identical are overly restrictive. General Large Deviations theory gives a more sophisticated theorem[11], with less restrictions. The queueing behaviour of traffic with a random composition will be dominated by the most bursty source over a certain period t. Experimental setups have shown that the buffer flow probability of real network traces successfully can be approximated by this theory. The knowledge of the number of sources or the different traffic types is not necessary [10]

$$Pr(Q \geq X) \approx \mathrm{e}^{s_{\hat{t}}(B+C\hat{t})-\Lambda_{\hat{t}}(s_{\hat{t}})} \tag{11}$$

where $\Lambda_{\hat{t}}(s)$ is the cumulant generating function of the traffic mix, $s_t = \arg\sup_{s\in\mathbb{R}^+}(s(B+Ct)-\Lambda_t(s))$ and $\hat{t} = \arg\inf_{t\in\mathbb{N}}(s_t(B+Ct)-\Lambda_t(s_t))$. $s_t = \arg\sup_{s\in\mathbb{R}^+}(s(B+Ct)-\Lambda_t(s))$ can easily be computed with the method of Brent exploiting the concavity of Λ_t in s so that $s(B+Ct)-\Lambda_t(s)$ is also concave in s [8]. A linear search over the discrete time intervals can be used to minimize the equation in respect to t.

The cumulant generating function can be transformed in a convenient and intuitive descriptor of the stochastic properties of the traffic stream, called the *Effective Bandwidth* ($\alpha_t^{(1)}(s)$) of the flow [14].

$$\alpha_t^{(1)}(s) = \frac{1}{st}\Lambda_t^{(1)}(s) \tag{12}$$

Suppose the optimum is attained, and the optimizing parameters are $s_{\hat{t}}$ and $\hat{t}$, both strictly positive. A small number of the N flows can be replaced by constant-rate flows of rate $\frac{1}{s_{\hat{t}}\hat{t}}\Lambda^1_{\hat{t}}(s_{\hat{t}})$. Locally, at $(s_{\hat{t}},\hat{t})$, these new flows have the same cumulant generating function as the flows they replace, so the rate functions are the same[10]. In other words, a flow with cumulant generating function $\Lambda^{(1)}_t$ has the same effect at operating point $(s_{\hat{t}},\hat{t})$, as a flow of constant rate $\alpha^{(1)}_{\hat{t}}(s_{\hat{t}})$. The effective bandwidth is additive for independent sources. Alternative, the effective bandwidth can be understood in terms of admission regions. Suppose there are mN flows with effective bandwidth $\alpha_t(s)$ and nN flows with effective bandwidth $\beta_t(s)$. For what values of m and n does the system meet the quality of service constraint γ?

$$\Pr\left(Q^N \geq Nx\right) < \mathrm{e}^{-\gamma N} \tag{13}$$

The admissible region is

$$\bigcap_{t>0}\left\{m,n : \exists s > 0 : m\alpha_t(s) + n\beta_t(s) < c + \frac{b}{t} - \frac{\gamma}{st}\right\} \tag{14}$$

The effective bandwidth gives the trade-off between flows of different types [14].

Let $A^{(1)}$ be a typical input flow in a queue fed by N independent and identically distributed flows and served at rate cN, and let $D^{(1)}$ the corresponding departure flow. $A^{(1)}_t$ is not described via a large deviations principle, neither is $D^{(1)}$. Instead A^N obeys a large deviations principle and so does D^N. The surprising result says that D^N satisfies the same large deviations principle as A^N. In a large deviations sense, the characteristics of a flow of traffic are not changed as it passes through a queue. Consider now a queue fed by many independent flows of different types: N flows like $A^{(1)}$ and N flows like $B^{(1)}$. Let $D^{(1)}$ and $E^{(1)}$ be typical outputs. If the total mean arrival rate is less than the transmission capacity, then D^N satisfies the same large deviations principle as A^N and E^N the same as B^N, which means that in a heuristic sense $D^{(1)}$ $A^{(1)}$ and $E^{(1)}$ $B^{(1)}$. By considering the additive behaviour of the cumulant generating function it is clear that $D^{(1)}$ and $E^{(1)}$ are essentially independent. The contrary might be expected. For example, if $A^{(1)}$ is very bursty and $B^{(1)}$ is rather smooth, on might expect that $D^{(1)}$ be less bursty and $E^{(1)}$ be less smooth. Indeed in a router with a small number of inputs this can happen. But in the many flows scaling regime[11] it is not the case. In other words, $D^{(1)}$ and $E^{(1)}$ 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* [24]. In the large deviations limit, it make 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 and 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. This is nice property allowing straightforward traffic engineering.

The time parameter t corresponds to the buffer busy period before reaching Nx. This is the time scale for buffer overflow probability, $Nx = Nb$. Smoothing of traffic will only be effective at a time scale larger than t because only then it affects the effective bandwidth of the traffic flow [9]. This parameter is also important for the granularity of the samples from a real traffic trace. The sampling time has to be less than t to have an accurate estimate but a value several orders less than t will have a big impact on the performance of the calculation of $\Lambda_t(s)$. A ripple effect can be seen in the effective bandwidth surface for periodic values of t over the full s-range. This indicates a periodic behaviour of the traffic source with a period equal to the time t of the first ripple.

The space scale s is a parameter for the degree of multiplexing [12]. The more s approaches 0 the multiplexing will be more efficient and the effective parameter tends to the mean rate of the traffic stream at the

[10] to the first order and under appropriate smoothness conditions

[11] Experimental setups have shown that already for a rather small number of flows $N = 3$ for two traffic classes, the departure flows are mostly decoupled [24].

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 t, the effective bandwidth is strictly concave going from the mean rate to the peak rate of the traffic flow at the time scale t.

The basic estimate $\Pr\left(Q^N \geq Nx\right) < \mathrm{e}^{-IN}$ can be refined by using the Bahadur-Rao theorem [18]. Conditions can be found which allow a tighter limit on the buffer flow probability. The result is an exact asymptotic solution for the queueing system

$$\Pr\left(Q^N \geq Nx\right) \approx \frac{1}{s_{\hat{t}}\sqrt{2\pi\sigma_{\hat{t}}^2 N}}\mathrm{e}^{-NI(x)} \tag{15}$$

where $I(x) = \inf_{t\in\mathbb{N}} \sup_{s\in\mathbb{R}^+} J_t(s,x)$ with $J_t(s,x) = s(x+ct) - \sum_{i=1}^{n} \rho^i \Lambda_t^{(i)}(s)$, $\rho^i = n^i/N$, n^i the number of flows from class i and $\sigma_{\hat{t}}^2 = \frac{\mathrm{d}^2 J_t(s,x)}{\mathrm{d}s^2}\,|(t=\hat{t}, s=s_{\hat{t}})$. s_t can easily be found as $\arg\left(x+ct = \sum_{i=1}^{n} \rho^i \frac{\mathrm{d}\Lambda_t^{(i)}(s)}{\mathrm{d}s}\right)$ and the expression for $\sigma_{\hat{t}}^2$ simplifies to $-\sum_{i=1}^{n} \rho^i \frac{\mathrm{d}^2\Lambda_t^{(i)}(s)}{\mathrm{d}s^2}\,|(t=\hat{t}, s=s_{\hat{t}})$. A further simplification gives an ad-hoc approximation [20], which is exact for the case of Gaussian arrival processes $\sigma_{\hat{t}}^2 \approx \frac{2J_{I(x)}}{s_{\hat{t}}}$. A final expression for the buffer flow probability can be obtained by entering in equation (15) this result and setting $N = 1$ because the number of flows or the exact composition of the flows are not needed

$$\Pr\left(Q \geq X\right) \approx \frac{1}{\sqrt{4\pi I(X)}}\mathrm{e}^{-I(X)} \tag{16}$$

where $I(X) = \inf_{t\in\mathbb{N}} \sup_{s\in\mathbb{R}^+} \left(s(X+Ct) - \sum_{i=1}^{n} n^i \Lambda_t^{(i)}(s)\right)$

The loss rate can be found for the exact asymptotic by using Petrov's Theorem [18]. Using the same approximation as equation (16), the loss rate becomes

$$LR \approx \frac{1}{s_{\hat{t}}\mu\sqrt{4\pi I(B)}}\mathrm{e}^{-I(B)} \tag{17}$$

where $\mu = \mathbb{E}\left(A_t\right)$ is the mean traffic rate of the aggregated traffic stream. This final equation will be the used to test the accuracy of the large deviations limit.

A trade off is done between exactness and speed of computation:

- the number of flows or the exact traffic mix are not needed;
- the probability is approximated with an exponential;
- a refined estimate based on the Bahadur-Rao theorem and Petrov's theorem is used;
- $\sigma_{\hat{t}}(s_{\hat{t}})$ is approximated by supposing that the aggregated traffic stream is Gaussian;
- decoupling makes it easy to follow a stream passing through a network;
- s gives an indication how well the traffic can be multiplexed on a link;
- the full statistical characteristics of a traffic flow have to be known;
- a sampling procedure is needed to obtain an estimated effective bandwidth or cumulant generating function from a real traffic trace.

4.0 CENTRAL LIMIT

Heavy Traffic[12] 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 [2] and relies on the assumption that the variance of the component traffic streams is uniformly bounded. Denote the component traffic streams in a network by $A_n^{(i)}$, the aggregation by $A_n = \sum_{i=1}^{N} A_n^{(i)}$ and the cumulative arrival process by $S_t = \sum_{n=-t+1}^{0} A_n$ where $S_0 = 0$. S_t can be seen as a vector with index t and every component of this vector has a Gaussian distribution in respect to the Central Limit theorem

$$S_t = \mathcal{N}\left(\mu t, \sigma_t^2\right) \tag{18}$$

where $\mu = \mathbb{E}(A)$ is the mean traffic rate of the aggregated traffic stream, $\mathbb{E}(S_t) = \mu t$ and $\sigma_t^2 = \mathrm{var}(S_t)$.

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. 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 aggregate traffic with a Gaussian distribution. This can be formulated in the basic queueing model

$$Q = \sup_{t \in \mathbb{N}} (S_t - Ct) \tag{19}$$

The buffer flow probability can be chosen as a queue indicator in the event driven simulation

$$\begin{aligned} \Pr(Q \geq X) &= \Pr\left(\sup_{t \in \mathbb{N}} (S_t - Ct) \geq X\right) && (20) \\ &= \Pr\left(\sup_{t \in \mathbb{N}} \left(\frac{S_t - \mu t}{\sqrt{\sigma_t^2}}\right) \geq \frac{X + (C-\mu)t}{\sqrt{\sigma_t^2}}\right) && (21) \end{aligned}$$

Assuming that $\sigma_t^2/t^2 \to 0$ as $t \to 0$, $\sigma_t^2/(X+(C-\mu)t)^2$ attains its maximum value at some finite $t = \hat{t}$ [5] and using the principle of the Largest Term

$$\begin{aligned} \Pr(Q \geq X) &\approx \sup_{t \in \mathbb{N}} \left(\Pr\left(\frac{S_t - \mu t}{\sqrt{\sigma_t^2}} \geq \frac{X + (C-\mu)t}{\sqrt{\sigma_t^2}}\right)\right) && (22) \\ &\approx \Psi\left(\frac{X + (C-\mu)\hat{t}}{\sqrt{\sigma_{\hat{t}}^2}}\right) && (23) \end{aligned}$$

where $\Psi(w) = \frac{1}{2\pi}\int_w^\infty \exp\left(-z^2/2\right) \mathrm{d}z$, the standard normalized Gaussian tail function, and $\hat{t} = \sup \arg_{t \in \mathbb{N}} \sigma_t^2/(X + (C-\mu)t)^2$. The qualitative statement "rare events happen only in the most probably way" suggests that the previous equation is a good lower bound approximation.

The Principle of the Largest Term is used to get this estimate. This seems at first sight a contradiction. How can for a queue filled with Heavy Traffic, i.e. the mean rate close to the transmission capacity, the

[12] Heavy Traffic limit is mostly used for the Fast Time Central Limit Theory or 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.

probability that the work in queue exceeds the buffer size, be a rare event? The answer is related to 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 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 0, a convenient approximation [1] can be used

$$\Pr(Q \geq 0) \approx 2\Pr(A > C) = \sqrt{\frac{2}{\pi\sigma^2}} \int_C^\infty \mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}} \mathrm{d}z$$

where $\mu = \mathbb{E}(A)$ is the mean traffic rate of the aggregated traffic stream and $\sigma^2 = \mathrm{var}(A)$ the variance of the aggregated traffic stream. The heuristic motivation that during busy periods, the buffer goes down roughly as often as it goes up, is straightforward. This can be used to rescale the lower bound.

To ease the computation of the integral, bounds for $\Psi(w)$ can be used to calculate the buffer flow probability.

$$\frac{1-w^{-2}}{\sqrt{2\pi}} w^{-1} \mathrm{e}^{-\frac{w^2}{2}} \leq \Psi(w) \leq \frac{1}{\sqrt{2\pi}} w^{-1} \mathrm{e}^{-\frac{w^2}{2}} \tag{24}$$

The upper bound reappears in the next session where the rare event hypothesis will be quantified.

Most other approaches calculate the limit in the assumption that the base streams are of a specific kind[13] without the principle of the Largest Term. The resulting equation is used to estimate the behaviour of measured traffic. The presented results are mostly very good due to the specific choice of data but in general it is impossible to predict whether the used formula is a good estimate. In this paper the general lower bound is used as queue indicator without any assumptions about the distribution of the basic data streams.

To test the performance of the Heavy Traffic limit, the loss rate has to be evaluated. The approach is quite different compared to the Large Deviations Approximation. The key idea is that the shape of the curves representing the buffer overflow probability and the loss rate are similar [16]. If there is a good estimate of the tail probability an a way to calculate $LR(A)$ then $LR(B)$ can be calculated as

$$LR(Y) = \frac{LR(X)}{\Pr(Q \leq X)} \Pr(Q \leq Y) \tag{25}$$

with the calculation of $\Pr(Q \leq X)$ and $\Pr(Q \leq Y)$ based on previous equations. For $LR(0)$ a convenient formula can be obtained

$$LR(0) = \frac{1}{\mu\sqrt{2\pi\sigma^2}} \int_C^\infty (z-C) \mathrm{e}^{-\frac{(C-\mu)^2}{2\sigma^2}} \mathrm{d}z \tag{26}$$

where $\mu = \mathbb{E}(A)$ the mean traffic rate of the aggregated traffic stream is and $\sigma^2 = \mathrm{var}(A)$ the variance of the aggregated traffic stream. The equation uses the fact that if $B = 0$, $\hat{t} = 1$.

Again a trade-off between exact values an ease of computation is made:

- the number of flows or the exact traffic mix are not needed;
- the aggregated traffic flow is assumed to be Gaussian;

[13] In [27] a Poisson Pareto Burst Process (PPBP) is used for the modeling of Internet traffic.

- a straightforward lower bound is obtained;
- only the mean and the variance of the summed aggregate traffic stream have to be calculated from a traffic trace;
- for small values of X the approach is not accurate;
- a integral that can only be solved analytically in a few specific cases, has to be calculated.

5.0 MODERATE DEVIATIONS

The estimate based on the Central Limit Theorem is appealing because it leads to parsimonious models, i.e. all that is needed to know about an arrival process is its mean an covariance structure. Large Deviations says integrals can be replaced by 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 the two techniques[26]: by the Central Limit Theorem the arrival process can be modeled as a Gaussian process; by Large Deviations theory, it is sufficient to study the most likely paths.

Several theories can be applied to get the same equation:

- Most Probable Path [1]: application of Schindler's theorem, a Gaussian Large Deviations result;
- Moderate Deviations [26]: application of Gärtner-Ellis theorem, also a Large Deviations result, to general traffic streams;
- Maximum Variance Asymptotic based on Extreme Value theory [7]: with key idea that the local behaviour of the normalized variance around the index $\hat{t}$ where the maximum variance is achieved, is found to essentially determine the supremum definition.

The approach in [26] quantifies with a burstiness parameter β the relation with both the Central Limit Theorem ($\beta = 0$) and the Large Deviations theory ($\beta = 1$) . Moderate Deviations concerns a range of scales between these. A typical traffic flows has bursts at all scales $0 \leq \beta \leq 1$, with larger bursts (large β) less frequently than smaller bursts (small β). The parameter β plays the dual role: it measures both burst size and burst frequency.

Let N be the number of stationary, identically distributed arrival processes $A_n^{(i)}$ with mean rate $\tilde{\mu}$, $S_t = \sum_{n=-t+1}^{0} A_n$ where $S_0 = 0$ the cumulative arrival process, C the transmission capacity and X the number of traffic units in the buffer. The following equation can be found

$$\frac{1}{N^{\beta}} \log \Pr\left(N^{\frac{1-\beta}{2}} \left(\frac{S_t}{N} - \tilde{\mu} t \right) \geq x \right) = - \inf_{t \in \mathbb{N}} \frac{(\tilde{x} + \tilde{c} t)^2}{2 \tilde{\sigma}_t^2}$$

where $\tilde{x} = X / N^{\frac{1+\beta}{2}}$, $\tilde{c} = (C - \mu N) / N^{\frac{1+\beta}{2}}$ and $\tilde{\sigma}_t^2 = \mathrm{var}(S_t^{(1)})$. If $\beta = 0$, the probability becomes $\Pr\left(\mathcal{N}\left(\tilde{\mu} N t, \tilde{\sigma}_t^2 N\right)\right)$ with an exponential approximations and if $\beta = 1$, the result resembles a Large Deviations result for a Gaussian distribution[14].

[14]For a Gaussian arrival process, $\Lambda_t(s) = s\mu t + \frac{1}{2} s^2 \tilde{\sigma}_t^2$ and the result follows directly by taking the supremum $s_t = \arg\sup_{s \in \mathbb{R}^+} \left(s(\tilde{x} + \tilde{c} t) - \Lambda_t^{(1)}(s) \right)$.

A straightforward estimate can be found using this equation

$$\Pr(Q \geq X) \approx \mathrm{e}^{-\frac{(X+(C-\mu)\hat{t})^2}{2\sigma_{\hat{t}}^2}} \tag{27}$$

where $\hat{t} = \sup \arg_{t \in \mathbb{N}} \sigma_t^2/(X+(C-\mu)t)^2$ and $\sigma_t^2 = N\tilde{\sigma}_t^2$. This result is independent of N and β. It has to be clear that this approximation is only valid, if the input process, C and X stand in a certain relation but of course the estimate can always be calculated.

As is the case for Large Deviations, some assumptions have to be made to control the tail of the distribution. The details can be found in [11]. An important result is that the full covariance structure γ_t can be recovered from the marginal variances σ_t^2.

$$\begin{aligned}
\gamma_0 &= \sigma_1^2 \\
\gamma_1 &= \frac{1}{2}\left(\sigma_2^2 - \sigma_1^2\right) \\
\gamma_t &= \frac{1}{2}\left(\sigma_{t+1}^2 - 2\sigma_t^2 + \sigma_{t-1}^2\right) \qquad \text{for} \quad t > 1
\end{aligned}$$

In other words, the marginal distributions of the cumulative arrival process fully characterize the process.

The same refinement as for the Large Deviations estimate can be applied [26]

$$\Pr(Q \geq X) \approx \frac{1}{\sqrt{4\pi I(X)}} \mathrm{e}^{-I(X)} \tag{28}$$

where $I(X) = \frac{(X+(C-\mu)\hat{t})^2}{2\sigma_{\hat{t}}^2}$ and $\mu = \tilde{\mu}N$ is the mean traffic rate of the aggregated traffic stream.

The scale procedure based on $\Pr(Q \geq 0) \approx 2\Pr(A > C)$ can be used to correct the buffer flow probability [1].

The loss rate can be found as for the Large Deviations case

$$LR = \frac{1}{s_{\hat{t}}\mu\sqrt{4\pi I(B)}} \mathrm{e}^{-I(B)} \tag{29}$$

or the same procedure as for the Central Limit estimate can be used [16] with or without the refinement.

This is an estimate for an isolated queue. For a network of queues the parameter β can be interpreted as a low-pass filter [26]. The loss probability at a queue of scale β will be of the order of $L^{-\beta}$. Thus the loss probability for a flow through a network will be dominated by the loss probability at the queue along the path with the smallest scale $\beta_{\min}$. 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 $\beta_{\min}$ 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 buffer flow probability or loss rate of a flow through a network can be approximated by estimating the buffer flow probability at the bottleneck link.

The Maximum Variance Asymptotic finds the same equation of the buffer flow probability for a general class of Gaussian processes with stationary increments [4], including a large class of long-range dependent processes with $\alpha = \lim_{t\to\infty} \log(\sigma_t^2)/\log(t)$, assuming that this limit exists. α can not be greater than 2

from the stationary increment property. The interval $\alpha \in [0,2]$ covers the majority of non-trivial Gaussian processes with stationary increments. Following conditions have to be respected:

$$\lim_{t\to\infty} t\frac{d\sigma_t^2}{dt} = \alpha \tag{30}$$

$$\lim_{t^2\to\infty} t\frac{d^2\sigma_t^2}{dt^2} = \alpha \tag{31}$$

$$\sigma_t^2 \sim Kt^\alpha \qquad \text{for some } K > 0 \tag{32}$$

$$\limsup_{t\downarrow 0} \frac{\sigma_t^2}{t^a} < -\infty \qquad \text{for some } a \in (0,\alpha) \tag{33}$$

The first two conditions are a direct result of the definition of α. The third condition is closely related to the self-similarity of S_t[15]. The last condition is about the behaviour of σ_t^2 around $t = 0$, and it is satisfied if σ_t^2 decreases as fast, or faster than t^a for some positive a as $t \downarrow 0$.

For sufficiently large X, $\log\left(\sigma_t^2/(X+(C-\mu)t)^2\right)$ is strictly concave on $[(\alpha-\sqrt{\alpha/2})X/(2-\alpha)(C-\mu), (\alpha+\sqrt{\alpha/2})X/(2-\alpha)(C-\mu)]$, and there is a unique index $\hat{t}$ where it attains its maximum [6]. The buffer flow probability and $\hat{t}$ can be computed by performing a simple local search algorithm starting at $\alpha X/(2-\alpha)(C-\mu)$. Even for fairly small numbers of X this can be used because $\sigma_t^2/(X+(C-\mu)t)^2$ is usually of a distinctly uni-modal shape. For small X the minimizer may not be unique. The formula for the calculation of $\Pr(Q \le X)$ however does not depend on the uniqueness of $\hat{t}$.

The upper bound for the estimate from the Central Limit Theorem $\frac{1}{\sqrt{2\pi}}w^{-1}e^{-\frac{w^2}{2}}$ is exponentially the same as the result from the Moderate Deviations scaling. The prefactor $\frac{1}{\sqrt{2\pi}}w^{-1}$ varies much slower than the exponential part. The fact that for sufficiently large X, $\log\left(\sigma_t^2/(X+(C-\mu)t)^2\right)$ is strictly concave, confirms and strengthens the reasoning to obtain the lower bound in the Central Limit Theorem approach.

It seems that the Moderate Deviations estimate can be the best of both worlds:

- the number of flows or the exact traffic mix are not needed;
- the aggregated traffic flow is assumed to be Gaussian;
- the probability is approximated with an exponential;
- a refined estimate based on the Bahadur-Rao theorem and Petrov's theorem can be used and a smart rescale can be applied;
- the interpretation of the burst parameter β as a low-pass filter makes the calculation of the loss rate in a network straightforward;
- only the mean and the variance of the summed aggregate traffic stream have to be calculated from a traffic trace;
- it is not directly clear if a traffic stream has the right scaling behaviour for a specific queue.

6.0 NUMERICAL RESULTS

To validate the accuracy of the formulas (15, 27, 23), both the *BOP* and the *LR*, both expressed in $\log 10$, are compared with simulations. The reported results have a maximum traffic unit loss probability of the order

[15]The Hurst parameter H is directly related to the parameter α by $2H = \alpha$. The first two condition can be used to obtain an estimate for H.

10^{-6} to obtain a reasonable confidence in simulations, expressed as an 95% confidence interval. Importance sampling needs to be used to get results with a lower traffic unit loss rate. The plots in this paper are only a small sample of the experiments that are conducted during the research for an optimal background fluid model.

The first experiment considers the multiplexing of periodic on-off sources with a period of 30 peak rates denoted by K_i, busy periods by on_i and composition by n_i. The following data are used

Source$_i$	n_i	K_i	on_i
1	$10N$	10	3
2	$4N$	1	10

The influence of N, C and B are respectively shown in the figures 1, 2, 3, 4, 5 and 6. The parameters for N, C and B are visible in the caption of each figure.

The *BOP* is bounded by the LD as upper bound and the CL as lower bound. The Large Deviations Bahadur Rao refinement fits the simulated probability for all values of x when the load is moderated. For a high load, the CL estimate without rescaling approximates very well the simulated buffer occupation probability.

The MD and MD BR estimates for both the *BOP* as the *CLR* are an approximation first for respectively LD and LD BR and if N, c or b increases for respectively CL and CL scaled. The Gaussian approximation becomes less suitable if the inverse of the normalized variance $\frac{\sigma_{\hat{t}}^2}{(B+(C-\mu)\hat{t})^2}$ or the space-scale s are high. This is a logical result. The Central Limit theorem deals with fluctuations of size $O\left(1/\sqrt{N}\right)$, but the fluctuations for a low load can be much higher $O(1)$ and can be calculated by the Large Deviations theory. A high value of the space-scale s indicates less multiplexing and a value of the effective bandwidth closer to the peak rate. Using the values of a Gaussian process in the cumulant generating the normalized variance and the space scale are linked[16].

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 c and x.

In all the results, the characteristic time-scale $\hat{t}$ identical is almost identical. This allows to reduce the expensive linear search for LD estimates by searching around the values of $\hat{t}$ found for the MD or CL approximation.

A second experiment is the multiplexing of 100 MPEG-1 encoded Star wars trace files [23]. The variation of *BOP* and *LR* in function of b are represented in the figures 7 and 8 with a transmission capacity of respectively 30 Mbits and 40 Mbits. The load is for the former 87.5% and for the latter 65.6%.

Both plots are very different and confirm the results from the multiplexing of periodic on-off sources. In regions with a high normalized variance, the MD scaled estimate gives an appropriate upper bound while in regions with a low load the LD BR estimate approaches the values of the simulation.

The plots show clearly that buffering is a limited solution for a heavily loaded link. The real answer to the problem is increasing the transmission capacity, the buffer can only contain a limited burst.

In the last experiment, the Bellcore trace file [17] is used to demonstrated the multiplexing of an unknown number of traffic sources where the individual distributions are not available. The *BOP* is calculated in function of B. $C = 2$ Mbits in the first figure 9 and $C = 4$ Mbits in the second 10.

The results are similar as the previous and indicate the limits of the applicability of a Gaussian approximation for aggregated traffic.

[16]For a Gaussian arrival process, $\Lambda_t(s) = s\mu t + \frac{1}{2}s^2\sigma_t^2$, $s_t = \arg\sup_{s\in\mathbb{R}^+}(s(B+Ct) - \Lambda_t(s)) = \frac{B+(C-\mu)t}{\sigma_t^2}$.

7.0 CONCLUSION

In this paper, the buffer flow probability for a single FIFO queue is estimated by different formulas ranging from a central limit approach to a large deviations limit. The approximations are bounded by LD as an upper bound and CL as an lower bound. For regions with a high normalized variance depending on the characteristic time-scale $\hat{t}$, the difference between the approximations is small. For low load ranges with a smaller degree of multiplexing, $s_{\hat{t}}$ large, only the LD BR gives an accurate estimate. This is shown by event-driven simulations of both synthetic and real traffic sources.

A appropriate designed network has a low *LR* for each queue in regime. If the loss rate is too high, the distribution of the traffic flow is altered and a fluid based approach becomes very difficult. The LD estimates are useful when the loss rate is limited, i.e. a high normalized variance. The computational expensive search for the characteristic time-scale $\hat{t}$ can be reduced by performing a search around the values found for the MD limit. In the case of a high load the CL approximation can be used, knowing that the nice features of LD, decoupling without smoothing and the effective bandwidth interpretation, are no longer supported.

The performance of these formulas can also be measured in a real hybrid fluid-flow event-driven simulation compared to a event-driven simulation of both background an foreground 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 results thus favoring the LD estimates. Nevertheless the simulations can be refined and speeded up to obtain tighter limits for the performance of the approximations. This will be investigated in a future work.

Topics for further research also includes:

- programming a real hybrid fluid-flow event-driven simulation with the LD BR estimate as background descriptor;
- estimating the burst parameter β to find the bottleneck node and adapting, smoothing, the fluid-flows after this node in a hybrid simulation with a high load scenario;
- extending the FIFO queue to priority queueing and generalized processor sharing;
- extending the model to include wireless base-stations and ad-hoc networks.

A fluid flow buffer indicator 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.

8.0 REFERENCES

[1] R. Addie, P. Mannersalo, and I. Norros. Performance formulae for queues with gaussian input. In *ITC 16*, 1999.

[2] R.G. Addie, M. Zukerman, and T.D. Neame. Application of central limit theorem to communications networks. Technical report, USQ, 1998.

[3] S. Bohacek, J.P. Hespanha, J. Lee, and K. Obraczka. A hybrid systems modeling framework for fast and accurate simulation of data communication networks. In *SIGMETRICS '03*, pages 58–69, 2003.

[4] J. Choe and N.B. Shroff. A central limit theorem based approach for analyzing queue behaviour in high speed networks. Technical report, Purdue University, 1998.

[5] J. Choe and N.B. Shroff. Queueing analysis of high-speed multiplexers including long-range dependent arrival processes. In *IEEE INFOCOM'99*, pages 617–624, 1999.

[6] J. Choe and N.B. Shroff. Queueing analysis with gaussian inputs including srd, lrd, and self-similar processes. Technical report, Purdue University, 1999.

[7] J. Choe and N.B. Shroff. Use of the supremum distribution of gaussian processes in queueing analysis with long-range dependence and self-similarity. *Stochastic Models*, 16, 2000.

[8] C. Courcoubetis and V. Siris. Procedure and tools for analysis of network traffic measurements.

[9] C. Courcoubetis, V. Siris, and G. Stamoulis. Application of the many sources asymptotic and effective bandwidths to traffic engineering. *Telecommunication Systems*, 12:167–191, 1999.

[10] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33, 1996.

[11] A. Ganesh, N. O'Conell, and D. Wischik. *Big Queues*. Springer, 2004.

[12] R.J. Gibbens. Traffic characterization and effective bandwidths for broadband network traces. In *Stochastic Networks: Theory and Applications*, volume 4 of *Royal Statistical Society Lecture Notes Series*. Oxford University Press, 1996.

[13] Y. Gu, Y. Liu, and D. Towsley. On integrating fluid models with packet simulation. In *IEEE INFOCOM'04*, 2004.

[14] F. Kelly. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications*, volume 4 of *Royal Statistical Society Lecture Notes Series*, pages 141–168. Oxford University Press, 1996.

[15] C. Kiddle, R. Simmonds, C. Williamson, and B. Unger. Hybrid packet/fluid flow network simulation. In *PADS'03*, 2003.

[16] H. Kim and N. B. Shroff. Loss probability calculations at a finite buffer multiplexer. In *IEEE/ACM Trans. on Networking*, volume 9, pages 765–768, 2001.

[17] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2:1–15, 1994.

[18] N.B. Likhanov and R. Mazumdar. Cell loss asmptotics in buffer fed with large number of independent stationary sources. In *IEEE INFOCOM'98*, 1998.

[19] B. Melamed, S. Pan, and Y.i Ward. Hybrid discrete-continuous fluid-flow simulation. In *SPIE*, volume 4526, pages 263–270, 2001.

[20] M. Montgomery and G. De Veciana. On the relevance of time scales in performance oriented traffic characterizations. In *IEEE INFOCOM'96*, volume 2, pages 513–520, 1996.

[21] V. Paxson and S. Floyd. Wide-area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking*, pages 226–244, 1995.

[22] G. F. Riley, T. M. Jaafar, and R. M. Fujimoto. Integrated fluid and packet network simulations. In *MASCOTS'02*, 2002.

[23] O. Rose. Statistical properties of mpeg videotraffic and their impact on traffic modeling in atm systems. Institute of Computer Science Research Report Series 101, University of Wuerzburg.

[24] D. Wischik. The output of a switch, or, effective bandwidths for networks. *Queueing Syst.*, 32:383–396, 1999.

[25] D. Wischik. Sample path large deviations for queues with many inputs. *Annals of Applied Probability*, 11:379–404, 2001.

[26] D. Wischik. Moderate deviations in queueing theory. 2004.

[27] M. Zukerman, T. Neame, and R. Addie. Internet traffic modeling and future technology implications. In *IEEE Infocom'03*, pages 1–4, 2003.

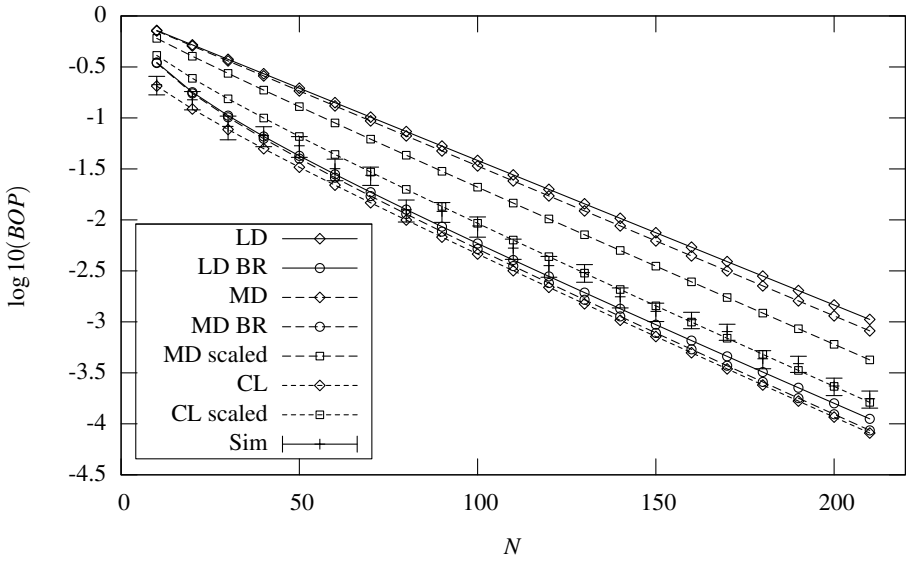

Figure 1: Variation of BOP as a function of N for the multiplexing of two periodic on/off sources where C=12N and B=5N.

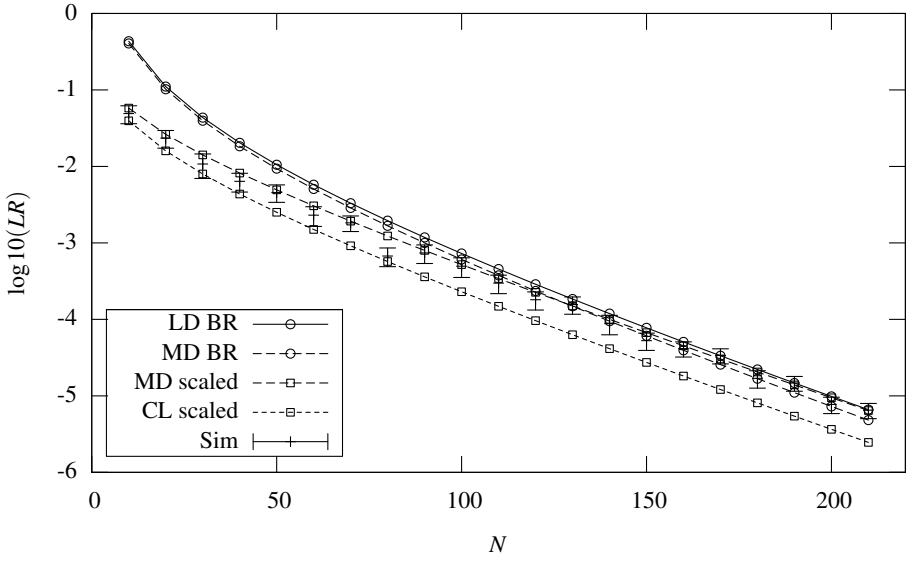

Figure 2: Variation of LR as a function of N for the multiplexing of two periodic on/off sources where C=12N and B=5N.

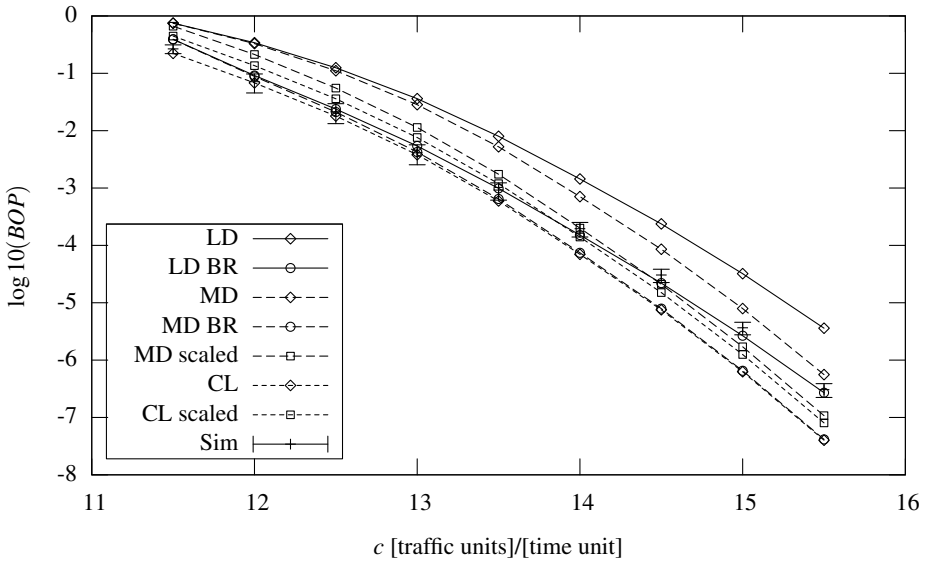

Figure 3: Variation of BOP as a function of c for the multiplexing of two periodic on/off sources where N=80 and B=2N.

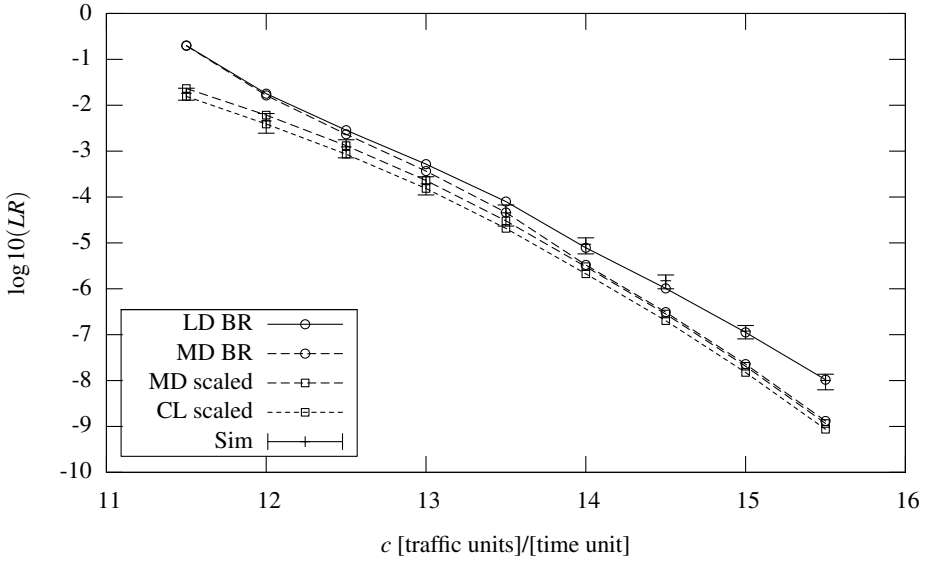

Figure 4: Variation of LR as a function of c for the multiplexing of two periodic on/off sources where N=80 and B=2N.

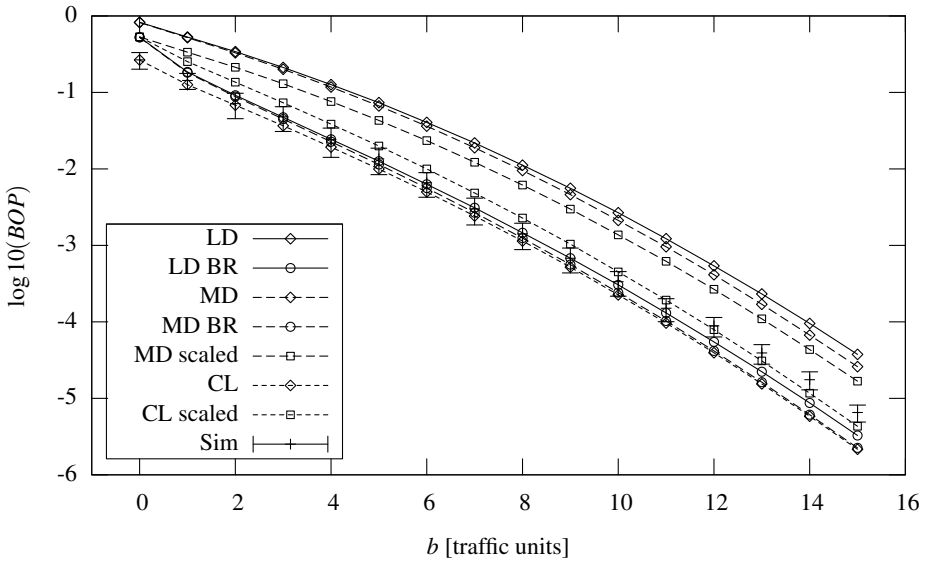

Figure 5: Variation of BOP as a function of b for the multiplexing of two periodic on/off sources where N=80 and C=12N.

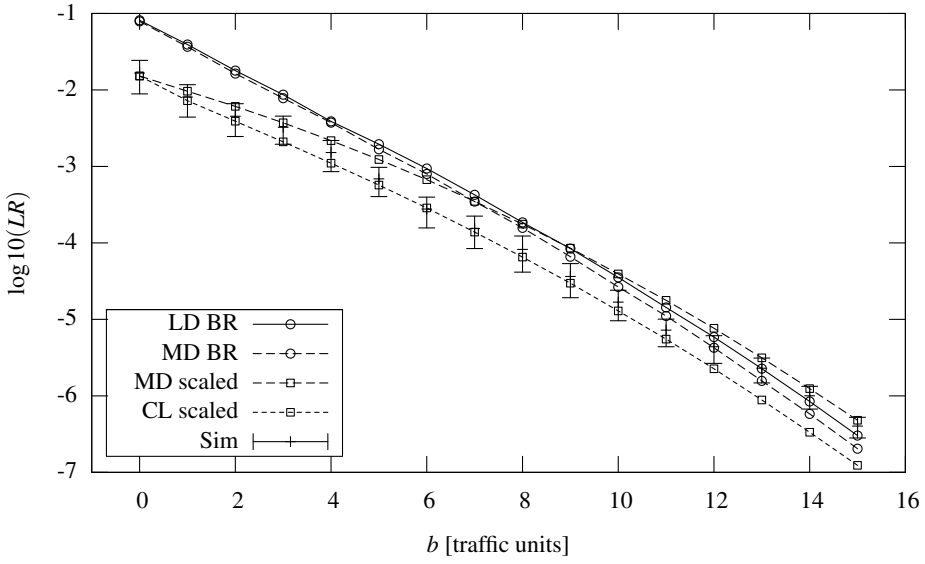

Figure 6: Variation of LR as a function of b for the multiplexing of two periodic on/off sources where N=80 and C=12N.

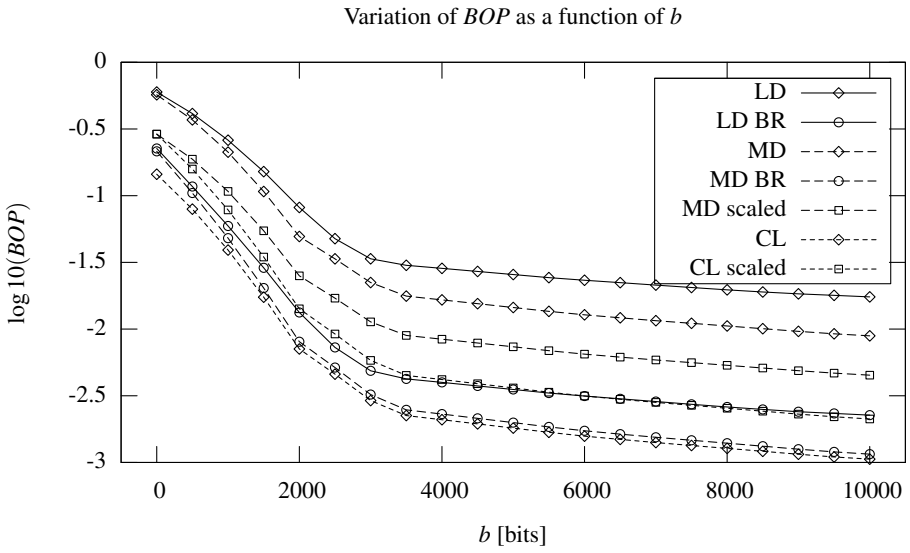

Figure 7: Variation of BOP as a function of b for the multiplexing of MPEG encoded Star wars traces where N=100 and C=30 Mbits.

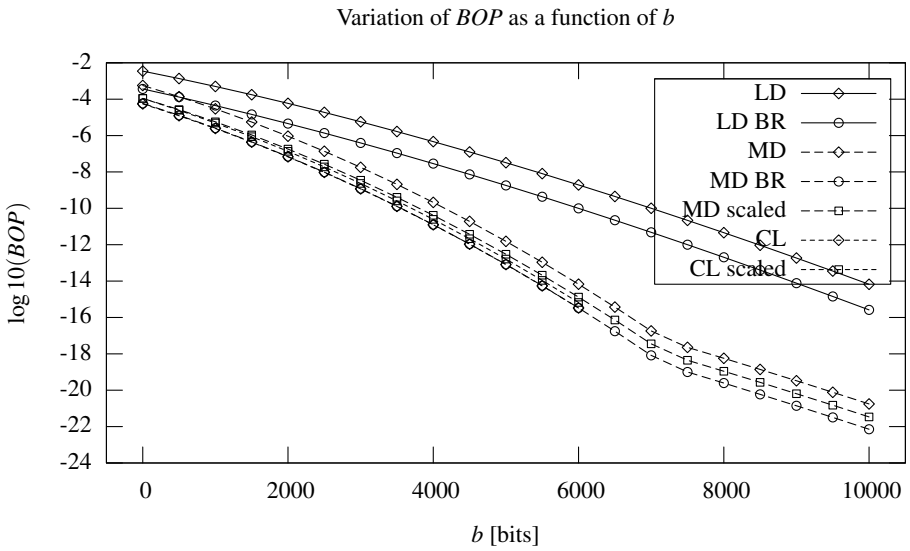

Figure 8: Variation of BOP as a function of b for the multiplexing of MPEG encoded Star wars traces where N=100 and C=40 Mbits.

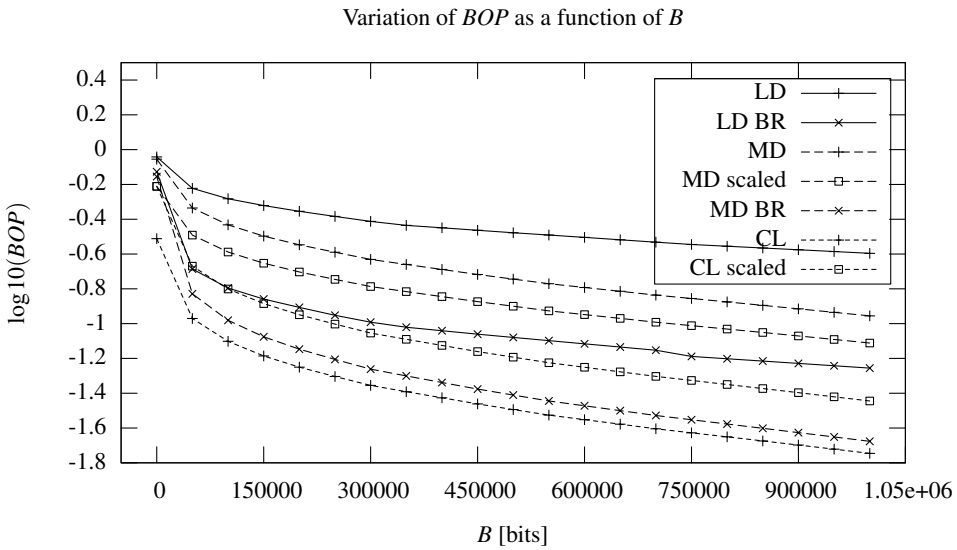

Figure 9: Variation of BOP as a function of B for the Bellcore trace where C=2 Mbits.

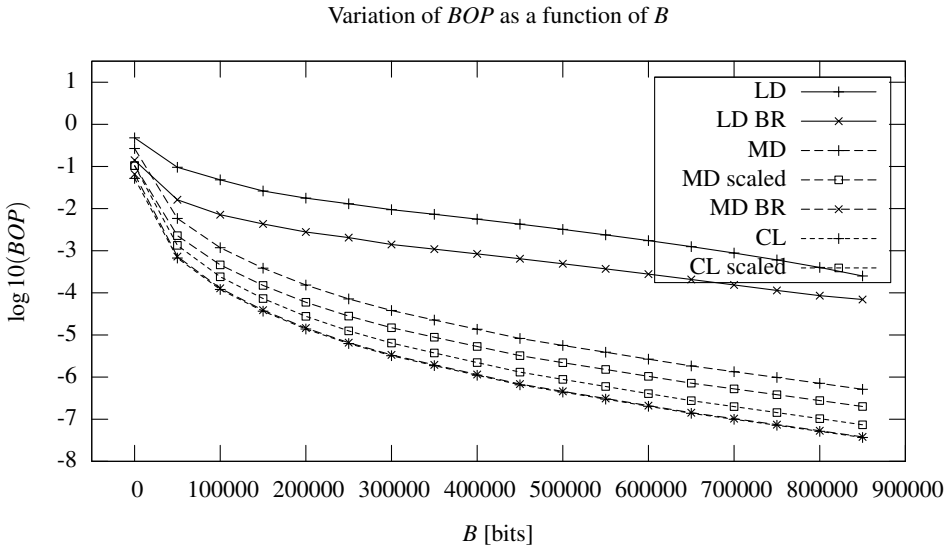

Figure 10: Variation of BOP as a function of B for the Bellcore trace where C=4 Mbits.