On the Necessity of Transient Performance Analysis in Telecommunication Networks

H.-P. Schwefel\textsuperscript{a} \& L. Lipsky\textsuperscript{b} and M. Jobmann\textsuperscript{c}

\textsuperscript{a}Siemens AG
81359 München, Germany

\textsuperscript{b}Department of Computer Science and Engineering,
University of Connecticut, Storrs, CT 06269-3155, USA

\textsuperscript{c}Institut für Informatik, Technische Universität München,
80290 München, Germany

Performance Analysis of analytic models of telecommunication networks is mostly based on steady-state methods. This paper discusses potential drawbacks of steady-state parameters, in particular of the steady-state buffer-overflow or cell-loss probability. The importance of transient performance analysis is demonstrated for long-range dependent (multiplexed) ON/OFF traffic. A transient parameter pair is proposed as replacement for the steady-state overflow or loss probabilities. The in-depth discussion of the behavior of those transient parameters reveals surprising results that allow for characterization and understanding of the fluctuations that are being observed in actual network behavior under traffic loads with long-range dependent properties.

1. Introduction

The use of analytic performance models has become increasingly popular for the analysis of telecommunication networks. However, frequently the analysis of such analytic models is limited to steady-state behavior, e.g. steady-state buffer-overflow probabilities. Although such steady-state analysis provides fundamental insights into the system's behavior, its applicability can be questionable in the scenario of telecommunication networks for several reasons: Measurements of network traffic have revealed properties such as \textit{self-similarity} and \textit{long-range dependence} (LRD), see [12]. For such traffic in systems with large buffers, convergence to the steady-state can be very slow, far slower, in fact, than say, for Poisson traffic. Thus, what may be an adequate time-span for \textit{normal} traffic to stabilize to its steady-state might be nowhere near convergence for LRD traffic, see [8], [5].

This paper demonstrates the need for transient analysis for network traffic with LRD properties, using an analytic queueing model with ON/OFF traffic. Here, we focus on buffer-overflow and cell-loss events, but the discussion can be extended to other performance parameters. It is shown that the correlation between the individual overflow/loss events can be dramatic, but this correlation is not captured in steady-state overflow/loss probabilities. A replacement of the individual steady-state probability by a transient parameter pair is proposed, and the behavior of those transient parameters is analyzed in our analytic queueing model. Finally, those results are validated by simulations.

\textsuperscript{a}A major part of this research was performed while the author was at Technische Universität München. The research was supported by T-Nova, Deutsche Telekom AG.
In addition to the vast literature on steady-state analysis of analytic telecommunication models (e.g. [?] ), [Manfred, please insert the reference to that book, requested by review 3] the need for transient analysis was already discussed earlier: [15] looks at the transient queue-length probabilities in a fluid flow model with $N$ exponential ON/OFF sources. [21] and [20] also use fluid queues, but with more general, Markov modulated arrival rates. Finally, [10] investigate the transient behavior of queues with a particular MMPP arrival process, but still restricted to exponential state-times. So far, little work was done with respect to the transient behavior for arrival processes with LRD properties.

2. Performance Model

2.1. Traffic Model: 1-Burst Model

The traffic model in this paper is a single ON/OFF source: cells arrive with Poisson rate $\lambda_p$ during ON periods (bursts) only. Each ON period is followed by a subsequent OFF period with no traffic. Later on in Section 5 the results are generalized for multiplexed traffic from $N$ sources ($N$-Burst). Reference [7] is among the earliest works in the literature that use such ON/OFF models, but only with exponentially distributed ON- and OFF-periods.

Let $\kappa$ be the mean rate of the ON/OFF source (the average for the ON- and OFF-times together). The ‘degree of burstiness’ of the source can be expressed by the parameter

$$b := 1 - \frac{\kappa}{\lambda_p}.$$  

Consequently, $0 < b < 1$, and the higher $b$, the more bursty is the traffic of the source.

The 1-Burst model is an extension of most ON/OFF models in the literature (e.g. [7]), since it allows more general, so-called Matrix-Exponential distributions (see [14]), for the duration of the ON periods. However, analytic queueing models with this more general input are still tractable for numerical evaluation.

(Truncated) Power-Tail Distributions

The distribution of the ON periods turns out to be critical for (steady-state) performance of queueing models with 1-Burst input, see e.g. [16]. Various measurements of actual network traffic ([3,22]) suggest that so-called Power-Tail (PT) distributions are appropriate for the ON periods. Such Power-Tail distributions show the following asymptotic behavior of their complementary distribution function, $R(x)$:

$$R(x) := \text{IP} \{ X > x \} \sim \frac{\text{CPT}}{x^\alpha} \quad \text{for large} \ x, \quad \text{where} \ 0 < \text{CPT} < \infty,$$

and $\alpha$ is called the Power-Tail exponent. $f(x) \sim g(x)$ is a short-hand notation for $\lim_{x \to \infty} f(x)/g(x) = 1$. The log-log plot of such a function $R(x)$ looks linear with slope $-\alpha$, since $\log(x^\alpha) = \alpha \cdot \log(x)$.

If Power-Tails with infinite variance ($0 < \alpha < 2$) are used for the ON-time distribution in the 1-Burst model, Long-Range Dependence (also called asymptotic second order self-similarity) in the inter-arrival process and in the counting process is the result [6].

Even if such Power-Tails in the burst-length distribution are real, for the analysis of the transient behavior, the tail matters at most up to the duration of the observation period. So clearly, truncated tails can be used when finite time-intervals are considered. In our 1-Burst model, we use truncated PT (TPT) distributions, which are introduced in [5] as a family of

2The model can be used for IP packets as well as ATM cells. Due to the history of this research, we use ATM terminology throughout this paper.
TPT Reliability Function: \( m=1.4, \text{Mean}=1 \)

\[ R_T(x) = \frac{1 - \theta}{1 - \theta^T} \sum_{i=0}^{T-1} \theta^i \exp \left[ \frac{-\mu}{\gamma} x \right], \]  

where \( 0 < \theta < 1 \) (here always \( \theta = 0.5 \)), \( \gamma = \theta^{-1/\alpha} \), and \( \mu = (1-\theta)|(1-(\theta\gamma)^T)|/(1-\theta^T)(1-\theta\gamma)^{\text{ON}} \)

The TPT reliability function for truncated power-tail distributions with \( T \) phases is shown in Fig. 1. This function exhibits Power-Law behavior for a limited range of \( x \) until some \( \overline{x_T} \) and drops off exponentially thereafter. That exponential drop-off is called truncation and corresponds to a Maximum Burst Size (MBS) in the 1-Burst model. See [17,18] for a more rigorous definition of the Power-Tail Range, \( \overline{x_T} \), which is a measure of the location of the exponential drop-off of the ON time distribution. Except for demonstration purpose in the right graph of Fig. 3, we always use TPT distribution with large \( T \) in this paper, such that the Power-Tail Range \( \gamma^{T-1}/\mu \) is larger than the finite observation interval.

2.2. Queueing Model

We investigate analytic queueing models that use the 1-Burst process as arrival process to an exponential server of rate \( \nu \). Two different buffer-models are considered in the following: One is the finite-buffer 1-Burst/M/1/B queue, in which a cell loss occurs when an arriving cell hits a full buffer. However, in many scenarios lost cells are retransmitted by higher protocol layers, so they still add to the load at the switch. This scenario can be approximated by an infinite back-up buffer. The model becomes a 1-Burst/M/1 queue, and we talk about buffer-overflow events when a cell arrives at buffer, whose occupancy is at least \( B \). The buffer-size \( B \) is the threshold that separates primary buffer and infinite back-up buffer.

Note that the queueing models in this paper assume the arrival process to be independent of the congestion level. That assumption is of course not true for TCP traffic models, as pointed out in [1]. See [19] for extensions of the ON/OFF approach to incorporate TCP-like behavior. Here we focus on the base model, although the derived results can be generalized.

The 1-Burst/M/1 queue with infinite buffer is stable as long as \( \rho = \kappa/\nu < 1 \). A well-defined steady-state queue-length distribution exists in the stable case, although its moments could well
be infinite. Since we want to compare transient and steady-state results in the following, we will only consider queues with \( \rho < 1 \), or equivalently \( \nu > \kappa \). The following analysis is restricted to the interesting scenario \( \lambda_p > \nu > \kappa \) (during an ON-time, cells arrive faster than the switch can serve them). In addition, we also assume that \( \alpha = 1.4 \), so that the ON-time distribution has a finite mean even for infinite Power-Tails, although its higher moments are infinite in that case.

3. Review of Steady-State Results

It is shown in various papers [9,4,17] that the steady-state queue-length distribution of 1-Burst/M/1 queues (or equivalent fluid-flow models) with \( \lambda_p > \nu \) is Power-Tailed with exponent \( \alpha - 1 \). Thus the Buffer-Overflow Probability (BOP) in the infinite buffer model behaves as

\[
\text{BOP}(B) = \Pr(Q \geq B) \sim \frac{\alpha_{\text{BOP}}}{B^{\alpha-1}},
\]

where \( Q \) is the random variable that describes the queue-length at cell-arrival instants.

The Cell Loss Probability (CLP) in a finite-buffer queue shows the same tail-behavior as the BOP, but with smaller tail-constant \( \alpha_{\text{CLP}} < \alpha_{\text{BOP}} \), see [18]:

\[
\text{CLP}(B) \sim \frac{\alpha_{\text{CLP}}}{B^{\alpha-1}}.
\]

The reason for such behavior is that the long queue-lengths are mainly caused by a single long ON period. The duration of such an over-saturation period is then Power-Tailed with exponent \( \alpha \). See [4,17] for the more general case of multiplexed traffic.

Similar steady-state results are known for the aggregation of ON/OFF sources, see Sect. 5.

4. Transient Analysis

![Figure 2. Simulated Fraction of Overflowed Cells, BOR(t, B), for an 8-Burst/M/1 Queue with \( B = 5000 \): The aggregated traffic of 8 ON/OFF sources is used here for demonstration purpose, see Sect. 5 for more details. In addition, this graph shows the average of all the 100 independent replications (thick line), and the fraction of replications that showed at least one overflow by time \( t \) (dashed line).](image-url)
Although there is no doubt that the steady-state analysis provides useful insights, it is necessary to investigate the behavior that we can expect to observe in actual networks, if the 1-Burst/M/1/B queueing model is an appropriate model. As an illustration, Figure 2 shows the result of a simulation experiment\(^3\), during which the fraction of overflowed cells is observed, called Buffer Overflow Ratio:

\[
\text{BOR}(t, B) := \frac{\text{number of overflowed cells in } [0,t]}{\text{number of arriving cells in } [0,t]} .
\] 

For \(\rho < 1\), the BOR converges stochastically to the steady-state BOP for long observation periods:

\[
\lim_{t \to \infty} \text{BOR}(t, B) = \text{BOP}(B).
\]

However, even though about 40 million arrivals were simulated in each replication in Fig. 2, the random variable \(\text{BOR}(t, B)\) still shows large variance, with the values ranging from 0 to about 35\% after the whole simulated period of \(t = 5\) min.

Furthermore, the results within the individual simulation runs show that the overflow events are highly correlated: 64 of the 100 simulation runs did not show any overflows at all, but if overflows occurred, then a large number of cells overflowed (witness the large jumps in Fig. 2). In this section, we introduce transient performance parameters that are able to describe such fluctuating behavior. Those parameters are based on an investigation of the very first overflow event. The time at which the buffer occupancy of the 1-Burst/M/1 model reaches level \(n\) for the first time is expressed by the random variable

\[
\tau_n := \min_{t \geq 0} \{ t \mid Q_t = n \} .
\]

\(\tau_n\) is called First Passage Time (FPT). Here we always assume an initially empty queue, \(Q_0 = 0\). Note that the FPT is independent of the buffer-model, i.e. there is no difference between the finite-buffer loss model and the infinite-buffer model, as long as \(n \leq B\).

The mean First Passage Time

\[
\text{mFPT}(B) := \mathbb{E} \{ \tau_B \} .
\]

can be computed by an algorithm developed from the martingale properties [2], or by a Linear Algebraic approach [11].

### 4.1. Transient Overflow Probability

The method that is developed in [2] allows the density \(f_{\tau_n}(x)\) of the FPT \(\tau_n\) for 1-Burst/M/1 queues to be obtained via the numerical inversion of its Laplace Transform. The numerical examples in [2] show that for large \(n\), the density function of \(\tau_n\) is very close to an exponential distribution with mean \(\text{mFPT}(n)\) when \(t > B/(\lambda_{p} - \nu)\). In this range for \(t\), an exponential distribution with mean \(\text{mFPT}(n)\) can be used as an approximation for the true \(f_{\tau_n}(t)\). The asymptotic theory that is discussed in [2] confirms the suitability of the exponential approximation and provides rigorous conditions for its asymptotic convergence.

From the knowledge of the distribution of \(\tau_n\), the transient overflow probability

\[
\gamma(t, B) := \text{Pr}(\text{At Least One Overflow in } [0,t]) = \text{Pr}(\tau_{B+1} \leq t) .
\]

(9)

can be derived. Using the aforementioned exponential approximation for the distribution of \(\tau_n\), we obtain

\[
\gamma(t, B) := \text{Pr}(\text{FPT}(B + 1) \leq t) \approx 1 - \exp \left( -\frac{t}{\text{mFPT}(B + 1)} \right) \quad \text{for } t > \frac{B}{\lambda_{p} - \nu} .
\]

\(^3\)Figure 2 shows a scenario with the aggregation of \(N = 8\) sources, which is discussed in Sect. 5.
A discussion of the behavior of \( \gamma(t, B) \) is provided in the next section.

If the ratio \( t / \text{mFPT}(B + 1) \) is very small, the exponential function in Eq. (10) can be approximated by the linear term of its Taylor series:

\[
\gamma(t, B) \approx \frac{t}{\text{mFPT}(B + 1)} \quad \text{for} \quad \frac{B}{\lambda_p - \nu} < t \ll \text{mFPT}(B + 1).
\]  

(11)

In a mathematically rigorous derivation of the asymptotics, two simultaneous limits \( B \to \infty \) and \( t \to \infty \) have to be used in a way that\(^4\)

\[
\lim_{B \to \infty} \frac{t(B)}{B} = \infty \quad \text{and} \quad \lim_{B \to \infty} \frac{t(B)}{B^\alpha} = 0.
\]

(12)


4.2. Conditional Overflow Ratio

With probability \( 1 - \gamma(t, B) \), a 1-Burst /M/1 queue shows no overflow-events at all in an observation period of duration \( t \) for a buffer-size \( B \). However, the simulation results in Fig. 2 indicate that the performance behavior in the other 'overflow' intervals might be extremely poor. In order to investigate the latter, we introduce another transient performance parameter, the (mean) conditional Buffer-Overflow Ratio:

\[
\text{mBOR}_c(t, B) := \mathbb{E} \{ \text{BOR}(t, B) | \text{BOR}(t, B) > 0 \}.
\]

(13)

While the transient overflow probability \( \gamma(t, B) \) is independent of the buffer-model, the fraction of lost cells is smaller in the finite-buffer loss model than the fraction of overflowed cells in the infinite-buffer model. Hence, an analogous definition for the (mean) conditional Cell-Loss Ratio in the finite N-Burst /M/1/B loss model is necessary:

\[
\text{mCLR}_c(t, B) := \mathbb{E} \{ \text{CLR}(t, B) | \text{CLR}(t, B) > 0 \}.
\]

The two transient performance parameters express the expected fraction of overflowed or lost cells, conditioned on at least one observed overflow/loss in a time-interval of length \( t \). Standard stochastic definitions provide a relation between \( \text{mBOR}_c \) and \( \gamma \):

\[
\text{mBOR}_c(t, B) = \frac{\mathbb{E} \{ \text{BOR}(t, B) \}}{\gamma(t, B)}.
\]

An analogous relation holds for the \( \text{mCLR}_c \). Clearly, \( \text{mBOR}_c(t, B) \) can be substantially larger than \( \mathbb{E} \{ \text{BOR}(t, B) \} \) when an overflow event occurs with small probability \( \gamma(t, B) \) within \( [0, t] \), as it happens in high performance networks, or for short time-scale \( t \) (such as the life-time of a connection).

4.2.1. Behavior of \( \text{mBOR}_c(t, B) \)

If \( t > B / (\lambda_p - \nu) \), \( \mathbb{E} \{ \text{BOR}(t, B) \} \) turns out to converge very quickly to \( \text{BOP}(B) \), see the simulation results in Sect. 5. Thus, \( \text{mBOR}_c(B, t) \) is approximately,

\[
\text{mBOR}_c(t, B) = \frac{\mathbb{E} \{ \text{BOR}(t, B) \}}{\gamma(t, B)} \approx \frac{\text{BOP}(B)}{\gamma(t, B)} =: \text{BOP}_c(t, B).
\]

(14)

However, note that for \( t < B / (\lambda_p - \nu) \), \( \text{mBOR}_c \) can be substantially larger than \( \text{BOP}_c \). Even \( \text{BOP}_c > 1 \) is possible for large \( B \) or small \( t \).

\(^4\)Knowing the behavior of the \( \text{mFPT}(B) \), see Sect. 4.2.2, taking the two limits in the described way turns out to meet the requirements on the relation between \( B \) and \( t \) in (11).
Figure 3. Computation of the conditional Buffer Overflow Probability in 1-Burst/M/1 Models: The left graph shows the results for a 1-Burst/M/1 model with sufficiently large (> t) Power-Tail Range of the ON-Time distribution. Surprisingly, BOP_c(t, B) grows with increasing buffer-sizes, see left graph. The asymptote for its growth (dashed straight line) is determined in the text, see Eq. (15). For truncated tails, the right graph shows that the BOP_c converges to a constant instead. In the setting of this graph, an average \( \kappa t = 4.8 \times 10^5 \) cells can be observed in the observation interval of \( t = 30 \) s.

Figure 3 shows in its left graph the numerical results for BOP_c(t, B) computed by the analytic 1-Burst/M/1 model: shown are the steady-state BOP(B), the transient probability \( \gamma(t, B) \) (approximated by Eq. (10)), and finally BOP_c(t, B) as the ratio of the two. We observe first that BOP_c \( \approx \) BOP, below \( B \approx 3000 \). Then, \( \gamma(t, B) \) starts to drop off, and causes BOP_c to grow. For large \( B \), the BOP_c curve looks like a straight line with slope 1 on log-log scale, i.e. it grows linearly with \( B \). The right graph of Fig. 3 makes clear that such a linear growth of the BOP_c is peculiar to the 1-Burst/M/1 queue with Power-tailed ON periods. Truncated tails with smaller PT Range of the ON-time distribution lead to the convergence of BOP_c.

It is shown in Appendix B that the conditional overflow ratio, mBOR_c(B) in a 1-Burst/M/1 model with PT distributed ON times (during which \( \lambda_p > \nu \)) grows linearly for some relevant range of large \( B \):

\[
mBOR_c(t, B) \approx \frac{1}{\alpha - 1} \cdot \frac{1}{i_\Delta} \cdot \frac{1}{1 - \rho} \cdot \frac{B}{\kappa t},
\]

where \( i_\Delta := 1 - \frac{1 - \rho}{\rho} \cdot \frac{1 - b}{b} \).

The derivation of the asymptotic behavior of the mCLR_c in the finite-buffer loss model is slightly less complicated, see Appendix A:

\[
mCLR_c(B, t) \approx \frac{1}{\alpha - 1} \cdot \frac{B}{\kappa t}.
\]

Eq. (17) is not quite mathematically rigorous. A simultaneous limit \( B \to \infty \) and \( t \to \infty \) would have to be considered, see Eq. (12) and Appendix A.

In summary, additional buffer space decreases the number of observation intervals with at least one overflow event, while it increases the conditional Overflow Ratio for LRD ON/OFF traffic.
from the 1-Burst model. For traditional, exponential burst-lengths (or truncated Power-Tails with short Range) the conditional Overflow Ratio converges. However, note that although larger buffers prevent some overflows, they increase the observed fluctuations between the overflow-ratios in different observation intervals. Either no overflows at all, or a huge number of overflows.

### 4.2.2. Corollary: Asymptotic Behavior of the Mean First Passage Time

The knowledge of the asymptotic behavior of the mBORc can be used to derive several corollaries: one of them is a confirmation of the observed asymptotic behavior of the mFPT in [2]:

Since the behavior of the steady-state BOP is known from [4,17] to be \( BOP(B) \sim c_{BOP} / B^\alpha \), and we now found the asymptotic behavior of mBORc(B) to be linearly increasing with \( B \), we can derive the asymptotic behavior of the mFPT. By using the exponential approximation (10) for \( \gamma(t, B) \) and the approximation \( 1 - \exp(-x) \approx x \) for small \( x \), we obtain:

\[
mBORc(t, B) \approx \frac{BOP(B)}{\gamma(t, B)} \approx \frac{BOP(B)}{1 - \exp(-t/mFPT(B))} \approx \frac{BOP(B) \cdot mFPT(B)}{t}.
\]

Using (15), we obtain for the asymptotic behavior of the mean First Passage Time:

\[
mFPT(B) \approx \frac{t \cdot mBORc^{(PT)}(t, B)}{BOP(B)} \sim c_{mFPT} \cdot \frac{B}{B^{1-\alpha}} = c_{mFPT} \cdot B^\alpha.
\]  \hfill (18)

Therefore, for ON/OFF traffic with Power-tailed ON periods with exponent \( \alpha \), the mFPT grows asymptotically by a Power-Law with exponent \( \alpha \). Computations of mFPTs for 1-Burst models in [2] confirm such an asymptotic behavior.

The constant in the asymptotic behavior of mFPT \( (B) \) can be derived in the same manner:

\[
c_{mFPT} = \frac{1}{c_{BOP}} \cdot \frac{1}{\kappa} \cdot \frac{1}{(1 - \rho) (\alpha - 1) \Delta}.
\]  \hfill (19)

Since it is often harder to compute mFPT \( (B) \) for large \( B \) the established relationship is very useful.

A similar relationship between the constant in the tail-behavior of the finite-buffer CLP and the mFPT can be derived from Eq. (17):

\[
c_{CLP} = \frac{1}{c_{mFPT} \kappa (\alpha - 1)}.
\]  \hfill (20)

A combination of Eq. (19 with Eq. (20) yields:

\[
c_{CLP} = (1 - \rho) \Delta c_{BOP}.
\]  \hfill (21)

### 5. Outlook: Multiplexed ON/OFF Traffic

Various papers [4,16,17] are concerned with the steady-state analysis of traffic models whose input is generated by multiple, independent ON/OFF sources. When using \( N \) sources with (truncated) Power-Tail distributions for their ON periods, the queue-length distribution, and thus steady-state performance is to a large extent determined by so-called over-saturation periods, during which time the arrival rate is temporarily higher than the service-rate \( \nu \). It is shown in [17] that at least

\[
\hat{\nu}_0 = \left[ N \cdot \frac{1 - \rho}{\rho} \cdot \frac{1 - b}{b} \right],
\]  \hfill (22)
sources have to be in a long-lasting ON period, in order to create such an over-saturation period, whose duration then shows a Power-Tail with exponent [17]

$$\beta = i_0 (\alpha - 1) + 1.$$  

Note that $\beta = \alpha$ for $i_0 = 1$, which is, in particular, the case for $N = 1$ when $\lambda_p > \nu > \kappa$. The steady-state queue-length distribution is also Power-Tailed with exponent $\beta - 1$, shown in [4,17]. With increasing utilization $\rho$, $\beta$ decreases at the so-called blow-up points, and steady-state performance becomes dramatically worse, see [16]. The steady-state buffer-overflow probability in the infinite-buffer model and the cell-loss probability in the finite-buffer loss system both show a Power-Law decay for increasing buffer-size $B$:

$$\text{BOP}(B) \sim \frac{c_{\text{BOP}}}{B^{\beta - 1}}, \quad \text{CLP}(B) \sim \frac{c_{\text{CLP}}}{B^{\beta - 1}},$$

where $c_{\text{CLP}} < c_{\text{BOP}}$, since $\text{CLP}(B) < \text{BOP}(B)$.

Transient analysis with the same concepts that are discussed in this paper can be applied to that multiplexed scenario, in which the average arrival rate is $\lambda = N \kappa$. However, the analysis of the asymptotic behavior is more complicated and only asymptotic bounds can be obtained, see [18] for the derivation:

$$\frac{1}{\alpha - 1} \frac{1}{i_\Delta} \frac{1}{1 - \rho} \frac{B}{\lambda t} \lesssim m\text{BOR}_e(t, B) \lesssim \frac{1}{i_0 (\alpha - 1)} \frac{1 - b}{b} \frac{\frac{1}{\rho} \frac{1 - b}{b}}{\lambda} \frac{B}{\lambda t},$$

where $i_\Delta := i_0 - N \cdot \frac{1 - \rho}{\rho} \cdot \frac{1 - b}{b}$.

Note that this definition of $i_\Delta$ reduces to (16) for $N = 1$ (which implies $i_0 = 1$). Furthermore, both bounds in (23) reduce to (15) when $N = 1$.

$$m\text{CLR}_e(B,t) \approx \frac{1}{i_0 (\alpha - 1)} \frac{B}{\lambda t}.$$  

As for the single source model, Eq. (24) is not quite mathematically rigorous. A simultaneous limit $B \to \infty$ and $t \to \infty$ would have to be considered, see (12).

The corollary about the asymptotic Power-Law growth of the mFPT still holds, but now with the exponent $\beta$:

$$\text{mFPT}(B) \sim c_{\text{mFPT}} B^\beta.$$  

Furthermore, a relationship between the tail constants in the asymptotic relationship for BOP and CLP can be obtained from results of the transient analysis. See Chapter 6 in [18] for preliminary results.

Validity via Simulation

Finally, we validate the derived asymptotes for the $m\text{BOR}_e$ for a 2-Burst model in a simulation experiment.

The graph in Figure 4 shows that the simulation estimates for $E\{\text{BOR}\}$, $\gamma$ and $m\text{BOR}_e$ correspond well with the computed exact values of the BOP, $\gamma$ and BOP$_e$ in the analytic $N$-Burst/$M$/$1$ model. Note that it is the steady-state BOP that is compared to the simulation estimate for $E\{\text{BOR}\}$. However, since the observation interval $t$ is sufficiently long with respect to the plotted range of $B$, these two values are close.
Figure 4. Validation of the Behavior of the Conditional Overflow Ratio for 2-Burst/M/1 Models via Simulation: The expected overflow ratio $E\{\text{BOR}(t, B)\}$ (marked by ‘x’), the probability $\gamma(t, B)$ (‘+’), and the conditional overflow ratio $\text{mBOR}_c(t, B)$ (bullets) were estimated from averaging about 4000 simulation runs. The analytic computation (solid lines) of the BOP, the approximation for $\gamma$ via the mFPT, and the resulting $\text{BOP}_c$ turned out to be well within the 95\% confidence intervals of the estimators (marked by triangles). The lower bound of the approximation (15) is also shown by the dashed line.
6. Conclusion

In traditional models of telecommunication systems, as for example for queues with ON/OFF arrivals with exponential ON times, steady-state behavior is observed rather quickly. However, in a finite observation period, the behavior of performance models with LRD traffic can be quite different from the steady-state results. Therefore, transient performance parameters can provide a much better description of the system's behavior.

Large cell delays as well as overflows or losses are not independent: Since these events are both due to full buffers, subsequent cells experience large delays or cause buffer overflows. This is true in any queuing model, but it is particularly accentuated in models with LRD arrival processes.

In this paper, two transient performance parameters are introduced that are recommended as a replacement of the steady-state overflow or loss probabilities. The behavior of these transient parameters is discussed for long-range dependent ON/OFF traffic. The observed behavior explains the fluctuations that are frequently observed in measurements of actual network traffic. The original parts of this paper are: the introduction of the mBOR\(_s\), (13), and the derivation of an approximation, (15), for the behavior of mBOR\(_s\)(t, B) for large buffers; results for the corresponding parameter, mCLR\(_s\), in the finite buffer model; the corollary about the behavior of the mFPT(B), (18); the relations (19), (20), and (21) between the tail-constants of the Power-Law behavior of the Buffer-Overflow Probability (in an infinite buffer model), the Cell-Loss Probability (finite buffer), and the mean First Passage Time. Finally, an outlook on the generalization of the results for multiplexed ON/OFF traffic is given in Sect. 5.

The main practical implications of the results in this paper are:

- **Choice of Performance Parameters:**
  Buffer Overflow (Cell Loss) events cannot be considered to be independent for each cell in the traffic, since those events are highly correlated. The substitution of the single parameter, Buffer-Overflow Probability, by the parameter-pair \((\gamma(t, B), \text{mBOR}_s(t, B))\) is recommended. An equivalent replacement is proposed for the Cell Loss Probability in the finite buffer model.

- **Fluctuations between observation intervals:**
  Increasing buffer-space reduces the overall number of overflow-events. However, for ON/OFF traffic with LRD properties, it increases the fluctuations between the individual observation intervals. Many intervals might not show any overflows at all, but if there are any, then a huge fraction of the cells is lost. This discrepancy between individual observation periods increases for larger buffers when LRD properties of the traffic are present.

APPENDIX

A. Asymptotic Approximations for the conditional Cell Loss Ratio

Overflows for N-Burst/M/1 queueing models for large B and \(\rho\) not too close\(^5\) to 1 are caused by long over-saturation periods, during which the mean arrival rate is temporarily raised beyond the service-rate. According to [17], such over-saturation periods for the 1-Burst/M/1 model with \(\lambda_0 > \nu\) have a duration \(X\) that is Power-Tailed with exponent \(\alpha\), i.e.

\[
R(x) \sim c/x^\alpha, \quad f(x) \sim \alpha c/x^{\alpha+1},
\]

when the ON periods are Power-Tail distributed with exponent \(\alpha\).

\(^5\)The server is idle for a substantial fraction \(1 - \rho\) of the time.
During these over-saturation periods, cells accumulate in the queue with average rate \( \lambda_p - \nu > 0 \). Assuming that the queue is empty at the start of the over-saturation period, the length of the over-saturation period has to be at least

\[
X > \frac{B}{\lambda_p - \nu} : x_0
\]

(26)
to be able to fill up the buffer to level \( B \). In the following, we look at over-saturation periods, \( X' \), which are long enough to cause overflows or cell losses, i.e. \( X' \) is the conditional random variable \( X' = (X | X > x_0) \). The density function and reliability function of \( X' \) are obtained by normalization:

\[
f_{X'}(x) = \frac{f(x)}{R(x_0)}, \quad R_{X'}(x) = \frac{R(x)}{R(x_0)}, \quad \text{for} \quad x > x_0.
\]

Hence, assuming a PT distribution (25) with exponent \( \alpha \) for \( X \), the expected value of the overshoot \( X' - x_0 \) is

\[
\mathbb{E} \{ X' - x_0 \} = \frac{1}{\alpha - 1} x_0, \quad \text{for large} \ x_0.
\]

(27)

Note that the expected value of \( X' \) grows linearly with the threshold \( x_0 \). That behavior is a property that is peculiar for Power-Tail distributions.

In the finite-buffer N-Burst/M/1/B model, a long over-saturation period \( X' \gg x_0 \) causes on average \( N_{(d)} \) cell losses with

\[
N_{(d)} = (X' - x_0) (\lambda_p - \nu).
\]

When looking at the asymptotic behavior for large \( B \), the probability that such long over-saturation periods \( X' \) occur during time \( t \) becomes very small. Thus, if overflows occur within time \( t \), then - with high probability - they will be caused by only a single long over-saturation period. Hence, asymptotically for large \( B \) (and large enough \( t \)),

\[
\text{mCLR}_c(t, B) \approx \frac{\mathbb{E} \{ N_{(d)} \}}{\kappa \cdot t} = \frac{\mathbb{E} \{ X' - x_0 \} (\lambda_p - \nu)}{\kappa \cdot t}.
\]

(28)

Together with Eq. (27) and with the definition (26) of \( x_0 \), we obtain for the asymptotic behavior of \( \text{mCLR}_c(t, B) \) for large buffers

\[
\text{mCLR}_c(B, t) \approx \frac{1}{\alpha - 1} \cdot \frac{B}{\kappa t}.
\]

(29)

Note that the derivation of the asymptotic behavior in (29) does not take into account that the over-saturation period \( X' \) could last longer than the observation interval \( t \). In particular for very large buffers \( B \) such a truncation of the over-saturation period by the end of the observation interval is likely to happen.

For a mathematically rigorous derivation of the asymptotic behavior, the limit \( B \rightarrow \infty \) would not be sufficient, but a simultaneous limit \( t \rightarrow \infty \) with some restrictions on the speed of growth of \( t \) in relation to the speed of growth of \( B \) has to be considered. However, for finite \( t \) and large \( B \) (29) provides a good approximation for the \( m\text{CLR}_c \). This is validated for its infinite buffer counterpart \( \text{mBO}_c \) in the simulation experiment in Sect. 5.

B. Conditional Buffer-Overflow Ratio for Infinite Power-Tails

The situation for the infinite-buffer 1-Burst/M/1 model becomes somewhat more complicated. Up to Equation (27) the argumentation is identical: buffer-overflows are caused by long over-saturation periods \( X' > x_0 \). However, in the infinite-buffer model, the queue-length can grow
to level $Q_1 > B$ during the over-saturation period, where
\[ Q_1 = X' (\lambda_p - \nu), \]
when assuming that the queue is empty at the beginning of the burst that caused the overflow. As soon as buffer-occupancy $B$ is reached, all arriving cells cause overflow events. Thus, the number of overflow events during the over-saturation period is on average
\[ N_{\text{ov1}} = (X' - x_0) \lambda_p. \]
The critical difference to the finite-buffer model is that even after the over-saturation period $X'$ ends, additional buffer-overflow events occur until the buffer has drained below the occupancy of $B$ cells. The draining in the 1-Burst $/M/1$ model occurs with average rate $(\nu - \kappa) > 0$. Thus the duration of the drain period is approximately
\[ T_{\text{dr}} = \frac{Q_1 - B}{\nu - \kappa}. \]

During the drain period of duration $T_{\text{dr}}$, on average another $N_{\text{dr}}$ buffer-overflow events occur:
\[ N_{\text{dr}} = T_{\text{dr}} \kappa = (Q_1 - B) \cdot \frac{\kappa}{\nu - \kappa} = (Q_1 - B) \cdot \frac{\rho}{1 - \rho}. \]
Hence, during a single large over-saturation period $X'$, the expected number of overflow events is
\[ \mathbb{E} \left\{ N_{\text{ov1}} + N_{\text{dr}} \right\} = \mathbb{E} \left\{ X' - x_0 \right\} \left[ \lambda_p + (\lambda_p - \nu) \frac{\rho}{1 - \rho} \right] = \mathbb{E} \left\{ X' - x_0 \right\} \frac{\lambda_p - \lambda}{1 - \rho}. \tag{30} \]
If $\gamma(t, B)$ is small and overflows occur in an observation period of duration $t$, it is very likely that those overflows are caused by only a single over-saturation period. Therefore, if we neglect the unlikely overflow events that might occur after the drain period has passed by, approximately for large $B$:
\[ \text{mBOR}_r(t, B) \approx \frac{\mathbb{E} \left\{ N_{\text{ov1}} + N_{\text{dr}} \right\}}{\kappa \cdot t}. \tag{31} \]
Putting the formulas (30), (27), and (26) together and using the burstiness $b = 1 - \kappa / \lambda_p$ of the sources, we get:
\[ \text{mBOR}_r(t, B) \approx \frac{B}{\alpha - 1} \cdot \frac{1}{i_\Delta} \cdot \frac{1}{\kappa \cdot \frac{1}{\kappa t}} \cdot \frac{1}{1 - \rho}. \tag{32} \]

Note that for a mathematically rigorous derivation of the asymptotic behavior again a simultaneous limit $B, t \to \infty$ has to be considered, see the end of the previous section.

REFERENCES

6. Michael Greiner, Manfred Jobmann, and Claudia Klippelberg: Telecommunication traffic, queuing models and subexponential distributions, Queueing Systems 33, pp. 125-152, 1999