On the Second Order Characterization in Traffic Modeling

Tamás Borsos\textsuperscript{1,2}, László Györfi\textsuperscript{1}, András György\textsuperscript{1}

\textsuperscript{1}Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Pázmány P. 1/D, 1117 Budapest, Hungary

\textsuperscript{2}Traffic Analysis and Network Performance Laboratory,
Ericsson Hungary
Laborc 1, 1037 Budapest, Hungary

Abstract

Superposition of ON/OFF processes frequently arise in network traffic modeling and often serves as a basis for performance evaluation of queueing systems. In this paper we consider the tail distribution of a queue, when multiplexing identical ON/OFF sources with keeping the average length of ON and OFF periods constant. The duration of ON and OFF periods follows a distribution that has polynomial tail. We show that for a finite but arbitrarily large number of multiplexed sources the limit distribution of the queue has at least polynomial tail. On the other hand, in the case of the equivalent Gaussian input, that is, a Gaussian process with the same mean and second order characteristics, the queue tail asymptotics is at most sub-exponential. Simulations showed that this significant difference of the queue tail probabilities also appear in case of practical traffic scenarios. These results imply that despite the similarity of multiplexed ON/OFF and Gaussian processes these models should be distinguished from a traffic engineering viewpoint.

Keywords: ON/OFF processes, multiplexing, queue tail asymptotics, Gaussian modeling

1 Introduction

It is well known that traffic on packet data networks and traditional voice traffic show substantially different characteristics. The most significant difference lies in their dissimilar temporal behaviour, i.e., the time dependence structure of a large proportion of data traffic types is not exponential. Packet data appears to be (asymptotically) self-similar, which gives rise to new, long range dependent (LRD) traffic models. Modeling efforts have been mainly focusing on capturing the first and second order characteristics (i.e., marginals and correlation structure) of real data traces. As a parsimonious model, the self-similar fractional Gaussian noise was suggested for this purpose by Leland \textit{et al.} \cite{10}, although they did not provide any physical explanation.

From queueing viewpoint, the heavy correlation of an input process has fundamental impact on traffic engineering, since the tail of the generated queue length distribution is larger than exponential. This phenomenon is also regarded as "buffer ineffectiveness", because large queue build-ups may occur with high probability. For the fractional Brownian model Norros \cite{13} showed that the queue length is at least Weibullian. Using large deviation techniques Duffield and O'Connell \cite{4} proved that this bound is asymptotically tight.

Superimposed ON/OFF sources and the M/G/\infty process are also considered for modeling and trace generation, as these processes can easily be parametrized to take the desired polynomial correlation structure. One type is the "idealized" ON/OFF process, i.e., when an ON period may be followed by another ON period and an OFF period can succeed another OFF period. In this case the ON and OFF period distributions are identical and the process becomes a simple renewal process. The autocorrelation function of such a source is exactly equal to the normalized
integrated tail of the renewal distribution function (see, e.g. [9]). The autocorrelation function of strictly alternating ON/OFF sources are more difficult to calculate. An asymptotic estimate for polynomial ON and OFF distributions is obtained in explicit form in [6]. This result is similar to the idealized case: the tail of the autocorrelation function is asymptotically the same as the integrated tail of the ON/OFF period distribution function. This means that the tail of the autocorrelation function of multiplexed identical ON/OFF sources with power tail ON/OFF durations remains polynomial.

Daniëls and Blondia [3] and Parulekar [14] proved that an M/G/∞ input process which is LRD generates polynomial queue length distribution. Jelenković and Lazar [7] considered the superposition of infinite number of identical ON/OFF sources, while keeping the average rate fixed by increasing the length of OFF periods. In this case the aggregate process converges in distribution to the M/G/∞ process. Similar aggregation is studied in [11]. Another way of multiplexing ON/OFF sources is introduced by Willinger et al. [17]. In that paper the source rate in the ON period is suitably scaled and the limiting process turns out to be fractional Gaussian noise. This model already gives a physical explanation for self-similarity. In [2] it is shown that the queue length has asymptotically Weibullian distribution.

In this paper we consider the discrete time queueing behaviour of the aggregation of a finite but arbitrarily large number of ON/OFF sources with polynomial ON and OFF period distributions. For both long and short range dependent (SRD) cases, we compare these queue lengths with those of the corresponding limiting Gaussian processes (i.e., which have the same autocorrelation functions). While the queue tail asymptotics is always polynomial with the ON/OFF input, it is exponential for SRD and Weibullian for LRD equivalent Gaussian input processes. The results on the tail asymptotics mean that one has to be careful when applying either ON/OFF or Gaussian models. Finally, we present simulation results to demonstrate the difference for practical queue tail probabilities.

2 Multiplexing ON/OFF sources - the queue tail can be large

Consider a discrete time queuing model with constant service rate s, and denote by \(X_n\) the number of arrivals in time slot \(n\). Let the initial length of the queue \(Q_0\) be an arbitrary non-negative integer valued random variable. Then
\[
Q_{n+1} = (Q_n - s + X_{n+1})^+
\]
for \(n \geq 0\). Concerning stability, Loynes [12] proved the following. Let
\[
V_0 = 0,
V_n = \sum_{i=0}^{n-1} X_{i+1} - ns, \quad (n \geq 1),
\]
where the sequence \(\{V_n\}\) is called the workload process. If \(\{X_i\}\) is stationary and ergodic, and \(E\{X_1\} < s\), then there is a stationary and ergodic sequence \(\{Q'_n\}\) and there is an almost surely finite random variable \(N_0\) such that \(Q'_n = Q_n\) for all \(n > N_0\), and \(Q'_0 = \sup_{n>0} V_n\).

In this section it is shown that depending on the duration of the ON periods the queue length \(Q\) may have larger tail than either exponential or Weibullian. The following proposition claims that even when multiplexing a large number of identical ON/OFF sources the queue tail may exceed an arbitrarily long polynomial tail if the ON/OFF distributions are chosen appropriately.

**Proposition 1** For all \(\delta > 0\) there are stationary and ergodic ON/OFF sources \(\{X^{(l)}_n\}, l = 1, 2, \ldots, L\) with \(E\{X^{(l)}_1\} = 1/2\) such that if the service rate \(s\) satisfies \(L > s > L/2\) then for the stationary queue length sequence we have
\[
P\{Q \geq q\} \geq cq^{-\delta},
\]
for any $q \geq 0$ with some constant $c > 0$.

A mathematically detailed and formal proof of the proposition can be found in [1]. Here a more intuitive approach is presented.

**Proof.** Suppose that the length of the ON and OFF periods $\tau_l$ are independent and identically distributed positive random variables with distribution

$$r_j = P\{\tau_1 = j\}, \ j = 1, 2, \ldots, \text{ and } r_0 = 0,$$

and corresponding distribution function $F(z) = \sum_{i=1}^{r_i} r_i$. Assume that $E\{\tau_1\} < \infty$.

Define $\{X_n^{(l)}\}$ to be 1 if the $l^{th}$ source is in ON state in time slot $n$ and 0 otherwise. The processes $\{X_n^{(l)}\}$ are independent, stationary and ergodic. Note that since the length of the ON and OFF periods have the same distribution, $P\{X_n^{(l)} = 1\} = P\{X_n^{(l)} = 0\} = 1/2$. Let $X_n = \sum_{l=1}^{L} X_n^{(l)}$. Because $s > L/2$ the queue length $Q_n$ is stable and has a limit distribution.

Let $W_n^{(l)}$ be the number of packets generated by the $l^{th}$ source in the corresponding ON period up to time $n$. If $n$ is in an OFF period of the source $l$ then $W_n^{(l)} = 0$. If all sources are in ON state for an interval of length $t$, the generated workload process during this time is $t(L-s)$. In this case, the required time to reach queue length $q$ is $t = \lceil \frac{q}{L-s} \rceil$. Therefore, for $q \geq 0$

$$P\{Q_n \geq q\} \geq P\{Q_n \geq q \mid X_n^{(l)} = 1, l = 1, 2, \ldots, L\} P\{X_n^{(l)} = 1, l = 1, 2, \ldots, L\} 2^{-L}$$

$$= \left( P\{W_n^{(l)} \geq \frac{q}{L-s} \mid X_n^{(1)} = 1\} \right)^L 2^{-L}.$$

The second inequality comes from the fact that if $W_n^{(l)} \geq \frac{q}{L-s}$ for all sources, then $Q_n \geq q$. On the other hand, the event $W_n^{(l)} \geq \frac{q}{L-s}$ means that the ON period of source $l$ started at least $\frac{q}{L-s}$ slot before the random time instant $n$ (note that, by condition, all sources are in ON state in the $n^{th}$ slot). Its distribution can then be calculated using the backward recurrence time argument from renewal theory (see, e.g. [8])

$$P\{W_n^{(1)} = j\} = \frac{1 - F(j)}{E\{\tau_1\}} \text{ for } j = 1, 2, \ldots.$$ 

Thus

$$P\{Q_n \geq q\} \geq \left( P\{W_n^{(1)} \geq \frac{q}{L-s} \mid X_n^{(1)} = 1\} \right)^L 2^{-L} = \left( \frac{\sum_{j=\frac{q}{L-s}}^{\infty} (1 - F(j))}{E\{\tau_1\}} \right)^L 2^{-L},$$

therefore

$$P\{Q \geq q\} \geq \left( \frac{\sum_{j=\frac{q}{L-s}}^{\infty} (1 - F(j))}{E\{\tau_1\}} \right)^L 2^{-L}.$$

If the ON and OFF distributions have power tail, i.e.,

$$1 - F(z) = z^{-(1+\gamma)} ,$$

where $\gamma > 0$ and $z \geq 1$, we have that for some constant $c > 0$

$$P\{Q \geq q\} \geq cq^{-L\gamma}.$$ 

The proposition then holds with $\gamma = \delta/L$. \hfill \Box
Remark. Note that the bound of Proposition 1 can be applied for both idealized and strictly alternating ON/OFF sources.

Since in this paper the effects of second order characteristics on queuing performance are considered, the covariance function of the cumulated input process is of primary interest. In the following we calculate it for an aggregated ON/OFF input. For simplicity, idealized ON/OFF sources are considered. Let \( G(j) \) be the distribution function of the backward recurrence time \( W_n^{(1)} \), i.e.,

\[
1 - G(j) = \sum_{i=j}^{\infty} \frac{1 - F(j)}{E\{\tau_i\}}.
\]

Then the covariance function of one ON/OFF source \( \{X_n^{(1)}\} \) (see e.g. [9]) is

\[
R(k) = R(0) (1 - G(k)), \quad k = \pm 1, \pm 2, \ldots
\]

where \( R(0) = \frac{1}{4} \), and the covariance function of \( L \) multiplexed sources \( \{X_n\} \) is

\[
L R(k) = L R(0) (1 - G(k)), \quad k = \pm 1, \pm 2, \ldots
\]

It implies that

\[
\sigma_n^2 = \text{Var} \left\{ \sum_{i=1}^{n} X_i \right\} = 2 \sum_{k=1}^{n-1} L R(k) (n - k) + nL R(0).
\]

\( \{X_n\} \) is called short range dependent (SRD) if the sequence \( \sigma_n^2/n \) is bounded, otherwise long range dependent (LRD). For the choice \( 1 - G(z) = c z^{-\gamma} \) with some constant \( c > 0 \), we have

\[
\sigma_n^2 = 2Lc \sum_{k=1}^{n-1} k^{-\gamma} (n - k) + nL R(0)
\]

After some manipulations we get

\[
\sigma_n^2 \simeq \begin{cases} 
  c_1 n^{2-\gamma} & \text{if } 0 < \gamma < 1 \\
  c_2 n \log n & \text{if } \gamma = 1 \\
  c_3 n & \text{if } \gamma > 1
\end{cases}
\]

(2)

where \( c_1, c_2, c_3 > 0 \) are other constants. It means that for \( 0 < \gamma \leq 1 \), \( \sigma_n^2/n \) is not bounded, hence \( \{X_n\} \) is LRD, and for \( \gamma > 1 \) \( \sigma_n^2/n \) is bounded, so \( \{X_n\} \) is SRD. Note that, although depending on \( \gamma > 0 \), \( \{X_n\} \) may be short range dependent or long range dependent, the queue length distribution always has at least polynomial tail by Proposition 1.

3 Queue tail with Gaussian input

In [16] Willinger et al. proposed to use aggregation of ON/OFF sources for trace generation to model self-similar Gaussian traffic. In that paper they considered idealized ON/OFF processes, while strictly alternating ON/OFF sources are considered in [17]. Taqqu et al. [15] use the Central Limit Theorem to prove the convergence of multiplexed ON/OFF sources to a stationary Gaussian process. On the basis of this result they propose to model a self-similar Gaussian process by multiplexed ON/OFF sources. The ON and OFF period length distributions are set so that the autocorrelation functions of the two processes match. In the following we show that the queue tail for a finite number of sources is larger than that of the limiting Gaussian process.

Consider the equivalent Gaussian process for the aggregate ON/OFF traffic, i.e. the case when the arrival process \( \{X_n\} \) is Gaussian with autocovariance function which has a polynomial tail. If \( \{X_n\} \) is weakly dependent then the tail of the limit distribution is asymptotically
exponential. The exponential tail distribution can be derived using large deviation techniques [5]. The basic tool in this respect is the cumulant generating function of the workload process:

\[ \lambda(\theta) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}\{e^{\theta (\sum_{k=1}^{n} X_k - ns)}\}, \]

assuming that the limit exists. If the set \( \{ \theta; \lambda(\theta) < 0 \} \) is nonempty then let \( \delta = \sup \{ \theta; \lambda(\theta) < 0 \} \). Then for large \( q \), \( \mathbb{P}\{ Q > q \} \simeq e^{-\delta q} \), that is

\[ \lim_{q \to \infty} \frac{1}{q} \log \mathbb{P}\{ Q > q \} = -\delta. \]

For a short range dependent Gaussian process \( \{ X_n \} \) the cumulant generating function of the workload process \( \{ \sum_{i=1}^{n} X_i - ns \} \) is

\[ \lambda(\theta) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E} e^{\theta (\sum_{i=1}^{n} X_i - ns)} = \frac{\theta^2}{2} \lim_{n \to \infty} \frac{\sigma^2_n}{n} + \theta (m - s), \]

where \( m = \mathbb{E}X_i \) is the mean of the process. In our case \( \lim_{n \to \infty} \sigma^2_n/n \) is positive and finite, therefore for \( s > m \) the set \( \{ \lambda(\theta) \leq 0 \} \) is nonempty and the queue length distribution is exponential.

**Proposition 2** Assume that the input \( \{ X_n \} \) is a stationary Gaussian process with variance \( L R(0) = L/4 \) and covariance function \( L R(k) = ck^{-\gamma}, k > 0 \), and mean \( m = L/2 \). Then for \( \gamma > 1 \) and service rate \( s \) such that \( \frac{L}{2} < s < L \),

\[ \mathbb{P}\{ Q > q \} \simeq e^{-\delta q} \]

for some \( \delta > 0 \).

For long range dependent arrivals Duffield and O’Connell [4] proved that the tail may be non-exponential. They introduced the scaled cumulant generating function

\[ \lambda^*(\theta) = \lim_{n \to \infty} \frac{1}{v_n} \log \mathbb{E}\{e^{\frac{\theta}{v_n} (\sum_{k=1}^{n} X_k - ns)}\}, \]

assuming that this limit exists, where \( v_n \) is a monotone increasing function with \( \lim_{n \to \infty} v_n = \infty \). Assume that there exist functions \( g \) and \( h_n \) such that \( h_n \) is monotone increasing with \( \lim_{n \to \infty} h_n = \infty \) and \( \lim_{y \to \infty} \frac{v_n}{h_n} = g(y) \) for all \( y > 0 \). Duffield and O’Connell proved that under certain conditions

\[ \mathbb{P}\{ Q > q \} \simeq e^{-\delta h}, \]

where \( \delta = \inf_{y > 0} g(y) I^*(y) \) and \( I^*(y) \) is the Legendre-Fenchel transform of \( \lambda^*(\theta) \), i.e., \( I^*(y) = \sup_{\theta \in \mathbb{R}} \{ \theta y - \lambda^*(\theta) \} \). They applied this result for the cases where \( \{ \sum_{k=1}^{n} X_k \} \) is a Gaussian process with stationary increments, or Ornstein-Uhlenbeck process, or a squared Bessel process.

In case a of long range dependent Gaussian process the scaled cumulant generating function is

\[ \lambda^*(\theta) = \lim_{n \to \infty} \frac{1}{v_n} \log \mathbb{E} e^{\frac{\theta}{v_n} (\sum_{k=1}^{n} X_k - ns)} = \lim_{n \to \infty} \frac{\theta^2}{2n^2} \frac{v_n}{\sigma^2_n} + \theta (m - s), \]

With scaling functions

\[ v_n = \frac{n^2}{\sigma^2_n} \simeq \frac{n^2}{c_1 n^{2-\gamma}} = \frac{1}{c_1} n^{\gamma}, \quad h_n = n^{\gamma} \]

which are increasing (recall that \( c_1 \) is defined in (2)), we have that

\[ \lambda^*(\theta) = \frac{\theta^2}{2} + \theta (m - s), \]
for which the limit \( g(y) = \frac{1}{c_1 y^\gamma} \) exists, and so
\[
\lim_{q \to \infty} \frac{\sigma_n^2}{q^2} \log P(Q > q) = -\delta,
\]
where
\[
\delta = \inf_{y > 0} \left( \frac{y - m + s)^2}{2c_1 y^\gamma} \right).
\]
These results are summarized in the following proposition.

**Proposition 3** Under the conditions of Proposition 2, for \( 0 < \gamma \leq 1 \) and \( \frac{1}{2} < s < L \),
\[
P\{Q > q\} \sim e^{-\delta q^\gamma}
\]
for some \( \delta > 0 \).

**Remark.** When the input process is a LRD Gaussian process with polynomial covariance function, it is asymptotically second order self-similar. The decay rate of the covariance function is determined by the Hurst parameter \( 1/2 < H < 1 \) : \( \sigma_n^2 \simeq n^{2H} \). Then the scaling function
\[
v_n = \frac{1}{c_1} n^{-2(1-H)}
\]
is polynomial and we get
\[
P(Q > q) \sim e^{-\delta q^{2(1-H)}}.
\]
In our notation
\[
P\{Q > q\} \sim e^{-\delta \sigma_n^2} \simeq e^{-\delta \sigma_1^2 y^{-\gamma}} = e^{-\delta y^{-\gamma} q^\gamma},
\]
which gives the relation \( 2H = 2 - \gamma \).

Combining Propositions 1-3 implies that the widely assumed equivalence of an aggregated ON/OFF traffic and the corresponding Gaussian process may lead to inadequate traffic engineering.

**Proposition 4** The queue tail generated by an aggregated ON/OFF processes with polynomial ON and OFF duration distributions is asymptotically always larger than that of the equivalent Gaussian process.

## 4 Queue tail simulations

The above results on the queue length distributions for the cases of SRD and LRD input processes are asymptotically different for the ON/OFF aggregates and the corresponding equivalent Gaussian sources. However, since all of these bounds refer to the asymptotics, one may be interested in how they perform for a practical number of multiplexed sources and overflow probability values (in the range \( 10^{-2} \) \to \( 10^{-6} \)). To demonstrate the difference between the queue tails simulations were performed and the results are presented graphically in this section.

In the first example the input process is \( L = 50 \) multiplexed ON/OFF sources with service rate \( s = 32 \) to achieve a reasonable 1% zero buffer overflow. The constituent sources are composed of strictly alternating identically distributed ON and OFF periods, resulting an aggregate mean of \( L/2 = 25 \). The decay parameter was chosen to be \( \gamma = 0.4 \), thus the input traffic is long range dependent. Its Gaussian equivalent has a Hurst parameter \( H = \frac{2\gamma + 2}{2\gamma} = 0.8 \).
Figure 1.: Queue tails for 50 ON/OFF sources and its Gaussian equivalent on logarithmic and log-log scale

queue tail of the ON/OFF input was obtained via simulations, while for the Gaussian process the approximation of [4] were used. (The lower and upper bounds of [2] show no significant deviation.)

One can see in Figure 1. that, although both processes have almost the same marginal distributions and autocorrelation functions, their queue tails differ by one or two orders of magnitude even at practical overflow probabilities, which shows the validity of the asymptotic claim of Proposition 4 also for small queue lengths. On the log-log plot the polynomial tail of the queue length for the ON/OFF aggregate input can also be observed even for small values.

In the second example a higher degree of aggregation is considered. The number of ON/OFF sources is $L = 100$, the service rate $s = 60$, $\gamma = 0.4$, resulting the same Hurst parameter $H = 0.8$ for the corresponding Gaussian process. The queuing behaviour is shown in Figure 2.

It can be seen that by increasing the number of ON/OFF sources the queuing performance of the multiplexed traffic somewhat approaches that of the equivalent Gaussian process. However, an order of magnitude difference in the queue tail probabilities remains for as many as 100 aggregated sources. Due to the different asymptotic behaviour of the two queue tails the observed

Figure 2.: Queue tails for 100 ON/OFF sources and its Gaussian equivalent on logarithmic and log-log scale
difference grows further for larger queue lengths. One can also see from the log-log plots that the polynomial tail of the ON/OFF input also appears in case of high degree of multiplexing.

5 Conclusion

In this paper we considered the discrete time queueing behaviour of the superposition of a finite but arbitrarily large number of identical ON/OFF sources. It was proved that if the ON period distribution is polynomial, then the generated queue tail asymptotics will also be polynomial. Comparison with the equivalent Gaussian processes, which have the same correlation structure, showed that while the queue tail asymptotics is polynomial with the ON/OFF input, it is exponential for SRD and Weibullian for LRD equivalent Gaussian input processes. Simulations on the queueing behaviour indicated an order of magnitude difference for high degree of multiplexing and practical overflow probabilities. These results show that multiplexed ON/OFF and Gaussian traffic models are not equivalent from traffic engineering viewpoint, as they significantly differ in their queueing performance.

References


