Daniel Zaragoza
IPS/EST/DEE Rua Vale de Chaves, Estefanilha 2914-508 - Setúbal Portugal.
Carlos Belo
IT/IST Av. Rovisco Pais, 11049-001 Lisboa Tel.: 21 8418459 e-mail: carlos.belo@lx.it.pt

Abstract

Experimental results with standard ON-OFF sources, as input to a one server queue, indicate that the M/D/1 model is adequate below a certain utilization, \( \rho^* \) (say), and plain wrong above. Moreover, as the output capacity increases, \( \rho^* \) increases towards 1. We provide simple engineering formulae, together with computable conditions of their applicability, to physically explain these facts. We also propose two rules/methods to determine quite accurately \( \rho^* \). One of these approaches is adequate to approximately predict the average queue length above \( \rho^* \).

I INTRODUCTION

Table I below best summarizes the motivation for the present study. \( W_{\text{sim}} \) (in ms) is the average waiting time in queue taken from \[1\] and \[2\]. \( \rho \) is the utilization and \( N \) is the number of sources. \( Q_{\text{sim}} = \lambda W_{\text{sim}} \) by Little’s law. \( Q_{\text{M/D/1}} = \rho^2/(1-\rho) \) is the well-known formula for the M/D/1 queue.

<table>
<thead>
<tr>
<th>( N ) [1]</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>125</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>.658</td>
<td>.731</td>
<td>.804</td>
<td>.878</td>
<td>.914</td>
</tr>
<tr>
<td>( W_{\text{sim}} )</td>
<td>.31</td>
<td>.45</td>
<td>.89</td>
<td>4.07</td>
<td>10.4</td>
</tr>
<tr>
<td>( Q_{\text{sim}} )</td>
<td>.61</td>
<td>1.01</td>
<td>2.15</td>
<td>10.71</td>
<td>28.52</td>
</tr>
<tr>
<td>( Q_{\text{M/D/1}} )</td>
<td>.63</td>
<td>99</td>
<td>1.65</td>
<td>3.14</td>
<td>4.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( N ) [2]</th>
<th>52</th>
<th>66</th>
<th>79</th>
<th>92</th>
<th>105</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>.395</td>
<td>.501</td>
<td>.600</td>
<td>.699</td>
<td>.797</td>
</tr>
<tr>
<td>( W_{\text{sim}} )</td>
<td>.0054</td>
<td>.0083</td>
<td>.04</td>
<td>.519</td>
<td>3.901</td>
</tr>
<tr>
<td>( Q_{\text{sim}} )</td>
<td>.13</td>
<td>25</td>
<td>1.41</td>
<td>21.38</td>
<td>183.37</td>
</tr>
<tr>
<td>( Q_{\text{M/D/1}} )</td>
<td>.13</td>
<td>25</td>
<td>.45</td>
<td>.81</td>
<td>1.57</td>
</tr>
</tbody>
</table>

These results were obtained by simulating, on a packet per packet basis, a multiplexer fed by a superposition of periodic ON-OFF sources. Details will be given in section II.

As can be seen from these results, the M/D/1 queuing model is adequate below a certain value of \( N \) and \( \rho \) (\( N^* \) and \( \rho^* \) say) but above it is not. Moreover, some authors point out that as the capacity increases (and with it, \( N \)), \( \rho^* \) increases towards 1. This is also visible in simulation results of \[2\] not shown here.

Because of the dramatic differences in the queue behavior, these results have raised the following questions.
  - a) Is the arrival process seen by the multiplexer truly Poisson, below \( \rho^* \)? Under which conditions?
  - b) Because \( \rho^* \) depends strongly on the characteristics of the ON-OFF sources and not only on the capacity, how to determine simply and accurately \( \rho^* \)?

The now “classical”, underload/overload (UL/OL) \[3\] or cell-scale/burst-scale \[4\] approaches have led to Engineering ‘rules of thumb’ (the \( r \) sigma rules) to determine the number of admissible sources which implicitly corresponds to \( N^* \). Examples of these rules can be found in \[5\] or \[6\]. However, they are used to avoid the problems raised by overbooking rather to explain it. On a case by case basis and at the expense of computer time, it is also possible to (in principle) solve numerically the equations or to run simulations.

The organization of this paper and our procedure are as follows. In section II we mainly define our models and notations. A multiplexer, with output capacity \( C \) b/s, is fed by multiple sources. Each individual source is modeled as a periodic ON-OFF process. The aggregation of \( N \) such sources is a Birth and Death (BD) process where the state, \( k \), is the number of active sources and for which statistics can be calculated both in continuous and discrete time. We assume that the sources are not synchronized. The various time scales of interest are mentioned and discussed. Then, provided the holding time of the BD process in state \( k \) is long enough, it is shown that for large \( k \) the packet arrival can be well approximated by a Poisson process. Moreover, a fluid flow approximation can well be used. The OL and UL zones appear whenever \( k \) is above or below a certain value \( M \) that depends on \( C \). In section III, further insight is gained by “simplifying” the input process. This simplification is here called the Averaged Input Process (AIP) and it gives physical interpretations to explain the observed simulated results. Other published interpretations are commented. Section IV presents two methods for determining \( \rho^* \). The first one is empirical and is based on the well-known 3-sigma rule that is used as a first order approximation. It is very simple and does not use the AIP. It also explains why as \( C \) increases, \( \rho^* \) also increases towards 1. The second method is based on the AIP and uses a fluid flow approach. It appears to be accurate for the computation of \( \rho^* \) and usable to determine the queue behavior above this value. Finally, in section V, we give our conclusions and indicate our future work.

II MODELS.

This section mainly serves the purpose of defining our models and notations. Many of the formulae indicated below can be found in standard textbooks, e.g. \[3\], \[7\].

A) Periodic ON-OFF source.

There exist several equivalent representations both in continuous and discrete time for this type of source. We use the same as \[1\] and \[2\]. During an ON period the source
emits one packet every $T$ seconds. The packets are of fixed size, $P$ bits, the peak rate being $h = PT$ in b/s or 1/P in pkt/s. The ON period is exponentially distributed with parameter $\mu = 1/T_1$. The OFF periods are also exponentially distributed with parameter $\lambda = 1/T_2$; $T_2$ being the mean OFF duration. During the OFF period no packet is sent. The ON periods are independent and identically distributed (iid), so are the OFF periods. Moreover, ON and OFF durations are independent. The activity factor of the source defined as $a = T_1/(T_1 + T_2)$ is also the probability that the source is ON at an arbitrary time. The average rate of this source is $m = ah$.

**B) Superposition of $N$ independent ON-OFF sources**

The $N$ sources are assumed statistically identical and independent. Let $n_N(t)$ be the number of active sources at time $t$. It is a specialized version of a continuous time Markov chain (CTMC) called a BD process. Embedded at the transition instants, in discrete time, there exists a discrete time chain (DTMC). Note that, more generally, the superposition of independent Markov renewal processes (MRP) is again a MRP [8].

The state space of $n_N(t)$ is $\{0,...,N\}$ and the transition rates are $\lambda_k = (N-k)\lambda$ for transitions $k\to k+1$ and $\mu_k = k\mu$ for transitions $k\to k-1$. The holding time in state $k$ is exponentially distributed with average

$$\tau_k = 1/(\lambda_k + \mu_k) = 1/\lambda_k.$$  (1)

The probability of transition from $k$ to $k+1$ is $P_{k,k+1} = \lambda_k / (\lambda_k + \mu_k)$. In the reverse direction, the transition probability is

$$P_{k,k-1} = \mu_k / (\lambda_k + \mu_k) = 1 - P_{k,k+1}$$  (2)

Because the sources are assumed periodic, they generate $n_N(t)$ packets in the interval $[t, t+T]$. For the BD process, the stationary distribution, $\{P_N(k)\}$, of $n_N(t)$ at an arbitrary time, is given by a binomial distribution. Let $P_N(k) = \Pr[k$ active sources, among $N$, at any time]

$$P_N(k) = \binom{N}{k} a^k (1-a)^{N-k}$$  (3)

The second order statistics are respectively $m_N = aN$ and $\sigma_N = \sqrt{a(1-a)N}$.

At this point we can make a remark. Write

$$k = m_N + r \sigma_N = m_N (1 + r \sqrt{(1-a)/a} \sqrt{N})$$  (4)

Then for “reasonable” values of $P_N(k)$, $r$ can be determined and fixed (e.g. between 3 and 6). From (4) it can be seen that as $N$ gets large, $k$ does not deviates much from the average. However, the rate of decrease is in $1/\sqrt{N}$, i.e., very slow.

So far we have considered the CTMC. For the DTMC, we denote $\{\pi_N(k)\}$ as its stationary distribution. It can be computed [9] by

$$\pi_k = a_k P_N(k) / \sum_j a_j P_N(j)$$  (5)

**C) Multiplexer and superposition.**

The multiplexer is modeled as a server with capacity $C$ b/s (i.e. capacity of the output link) and infinite buffer to hold the packets. It has $N$ input links corresponding to the $N$ sources. We assume that the input lines have peak rate $h$, so that when a source is ON, the packets form a “string” of contiguous bits. The multiplexer, however, deals with packets not bits. We assume no synchronization in the packet emission of the sources. We note $b_o = 1/C$ the output bit duration and $b_i = T/P$ the input bit duration. Let $\Delta = T/C$, the time necessary to transmit a packet on the output link. The utilization of the multiplexer is noted $\rho$, as usual, and is given by

$$\rho = (aNP/T) / C = aN \Delta T = aN/M, M = T/\Delta = b_i / b_o.$$  (5)

Note that $M$ (integer part of $T/\Delta$ and >1) is the number of packets that can be served “at once” during $T$. Because we assume (also as usual) that the sources are uniformly distributed, it may happens that one or more packet arrive while one is being transmitted. This is cell scale congestion. If $N \leq M$, then there is only cell-scale congestion. However, the utilization is low. If, to increase the utilization, one allows $N > M$ then, burst-scale congestion enters into play. From now on we assume $N > M$.

Table II summarizes the parameters used in [1] and [2].

**Table II**

<table>
<thead>
<tr>
<th></th>
<th>T1 (ms)</th>
<th>T2 (ms)</th>
<th>T (ms)</th>
<th>C (Mb/s)</th>
<th>P (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>352</td>
<td>650</td>
<td>16</td>
<td>1.356</td>
<td>64</td>
</tr>
<tr>
<td>[2]</td>
<td>321</td>
<td>1273</td>
<td>0.449</td>
<td>25</td>
<td>53</td>
</tr>
</tbody>
</table>

**D) Time scales of interest.**

From the previous discussion various time scales of interest can be identified.
- At the bit level: $b_o, b_i$. Output, input bit durations.
- At the packet level: $T$ periodicity of a source, $\Delta$ transmission time of a packet.
- At the aggregated sources level: i) $\tau_k$ holding time of the CTMC in state $k$. ii) Sequence of $\tau_k$. OL or UL regions modeled as the AIP below.

The sources are assumed uniformly distributed over T. Define $p = P[1$ bit arrival during $b_o] = P[1$ pkt arrival during $\Delta]$. Then,

$$p = 1/M = \Delta T = b_i / b_o = h/C$$  (6)

Suppose $k$ sources are active. Let $P_0(j)$ = $P[j$ input bits (among $k$) in $b_o] = P[j$ pkt (among $k$) in $\Delta]$. Then from standard textbooks, such as [7], we have

$$P_0(j) = \binom{k}{j} p^j (1-p)^{k-j}.$$  (7)
Which is again a binomial distribution - this time with parameter $p$. The second order statistics are respectively $m_k = k p$, $\sigma_k = \sqrt{kp(1-p)k}$. We note that $kp$ is the average number of arrival per service time. We would like to have $kp = p_{M/D/I}(k)$ i.e., for each state $k$, the queue would behave like a M/D/1 queue.

To that end it is necessary to evaluate the probability of events in disjoint intervals. Again from [7], and after some algebraic manipulation, we have

$$P_k(r,s) / P_k(s)L_k(s) = q(r,s) = (k-r)!(k-s)!/(1 - 2p)^{k-r-s} / k!(k-r-s)!/(1 - p)^{2k-r-s} \quad (8)$$

For $q = 1$ the events are independent. This can only be attained in the limit. However, for practical purposes, if $1-q$ is very small then one can consider the events as independent.

We have now established all the useful formulae for our argumentation, which also apply to the cases [1] and [2] at hand. Given the parameters of individual sources, it is not difficult to evaluate numerically the following.

i) From (3) evaluate the range of “reasonable” $k$. Then from (1) calculate the holding time in state $k$. The ratio $\tau_k/b_0$ should be large i.e. in the range of 1000 times or more.

ii) From (7) evaluate the discrepancy with a Poisson distribution of rate $kp$ as a function of $kp$ and $k$ (note that $kp$ is about 1), again for reasonable $j$.

iii) From (8), evaluate the difference $1-q$, which should be small in order that events in disjoint intervals be almost independent.

Once these calculations are done, provided the assumptions are verified and because the packet are of fixed length (more generally, one can work with a minimum packet length), one can assert that

i) During a (long) sojourn time in state $k$, the packet arrival is well approximated by a Poisson process of rate $kp = p_{M/D/I}(k)$. For $kp < 1$, in the UL regime, the M/D/1 queue is stable with a short transient regime. This can be checked by calculating the transient regime of the M/M/1 queue and removing the variability of the service time. However, when $kp > 1$, in the OL regime, the queue is not stable anymore, yet, the arrival process may still be very close to a Poisson process.

ii) In either the UL or OL regimes, provided the holding times are large, one can disregard the statistical fluctuations and use a fluid flow approach. Note we do not need the Poisson approximation here, only a long holding time is required.

For large $N$, it is notoriously difficult to numerically solve the complete queuing system either because of numerical instabilities or because of the sheer size of the system to solve. As we are also interested in a physical understanding of the OL region we “simplify” the input process. This is the purpose of the next section.

III THE AVERAGED INPUT PROCESS (AIP)

For the OL-UL regimes, further insight can be gained by approximating the input process by a two-state process. One state corresponds to the case where $n_B(t) > M$ i.e. the OL state, the other to $n_B(t) \leq M$ i.e. the UL state. We note that the idea is not new and a pretty large number of publications have considered it in different ways. However, our approach is more physical (i.e. “intuitive”) and we contribute here a simple result that seems new to our best knowledge.

It is well known that for a BD process, statistics of interest can be obtained by matrix geometric methods because the distributions are of phase-type. However, simpler methods can be used to obtain them. We are interested in:

- Average (and standard deviation) for the holding times and number of transition in the OL and UL states.
- Average level in the OL and UL states.

As for the first item, let $\tau_i (\tau_2)$ be the average absolute-time spent above M (below and including M). Let $U_1 (U_2)$ be the average number of transitions above M (below and including M). It can be shown that

$$\tau_i = P_M(n>M) / P_M(M), \quad (9)$$
$$\tau_2 = P_M(n=\infty) / P_M(M), \quad (10)$$
$$U_1 = \pi_M(n>M) / \pi_M(M), \quad (11)$$
$$U_2 = \pi_M(n=\infty) / \pi_M(M). \quad (12)$$

With the symbols $X(>M)$ meaning $\sum_{j\in M} X(j)$, $M < j \leq N$ and $X(\leq M)$ meaning $\sum_{j\in M} X(j)$, $0 \leq j \leq M$.

Also, the mean time between overloads is $1/P_M(M)\lambda_M = 1/P_M(M+1)\lambda_{M+1}$ in continuous time and $1/\pi_M(M)\lambda_{M+1}$ in discrete time.

Using the work of [10] it is also possible to calculate the standard deviation, $\sigma_M$, for the quantities of interest. To spare space and because we will not use them here, we omit the formulae. We only note that the holding times are not exponentially distributed and that the ratio $\sigma_M/\text{mean}$ depends on M and tend (very slowly) towards 1 from above as M increases.

For the calculation of the average level above M, the following algorithm serves as a proof. Let $M^*$ denote the average level above M in the OL state. During an OL period the number of transitions is odd, and $n_B(t)$ is never below M+1. The OL period starts at the level M+1 and ends when a transition M+1->M occurs.

i) Because $U_1$ is not necessarily integer, choose $U$ as the odd “ceiling” integer (e.g. $3 < U_1 < 5$). Let $U_1$ be the discrete time in state $U_1$.

ii) But $U_1$ and $U_2$ are given and $U_1 \leq U \leq U_2$ be the discrete time in the OL state. Let $v = U_1 / \pi_M(M)\lambda_{M+1}$, then $v = \text{true level} - M$, i.e., $v$ is an offset. Let $i(j)$ denote the transition probability from $i$ to $j$. Because transitions are only possible from adjacent states we have that

$$v = (v-1)_{a_i} + (v+1)_{a_i} + (v-1)_{a_j} + (v+1)_{a_j} \quad (13)$$

iii) The following conditions must apply:

- $v = 0$ if $v > v_{\text{max}} = (U+1)/2$ OR $v < 1$ OR $v > U - u + 1$.
- $v = 1$ if $v = 2$.
- $v = 1$ if $U \geq (U+1)/2$ AND $v = U - u + 1$.
we note that given M with N increasing (or fixing N and shown in Table III. Focusing on physical interpretation, first, agreement was found with these results. However, our method appear in that report. Using our method, a pretty good
interval
The second equation is obtained by considering that in the congestion period (i.e. OL period), the volume of information lost and the number of “bursts” arriving during that period in the OL it is M*hτ1 and in the UL rULτ2. Note that rUL is less than aNh but can be very close if τ1/τ2 is very small.
At this point several comments can be made.
- Independently of mathematical proofs we have verified by simulation that the above formulae (and those derived from [10]) give indeed the correct values.
- As a crosscheck, we have applied our method to the case described in [11]. In that report, the concern is about bufferless connection admission control and the occupation time above the link capacity for the superposition of a large number (e.g. 1400) of generally distributed ON-OFF sources. Both theoretical and simulation results about the duration of a congestion period (i.e. OL period), the volume of information lost and the number of “bursts” arriving during that period appear in that report. Using our method, a pretty good agreement was found with these results. However, our method is not asymptotic.
- Returning to our initial problem, numerical results are shown in Table III. Focusing on physical interpretation, first, we note that given M with N increasing (or fixing N and decreasing M), the frequency of overload, P(>M), increases and consequently the probability of underload decreases. Moreover, as N increases, τ1 (U1) increases somewhat but not much, i.e. the overload period lasts a little longer. However, the underload period is dramatically reduced. This means that, very often, time is too short in the UL period to “evacuate” the previously accumulated work. Second, we note that the average delay (or queue length) is the weighted sum of two terms. These terms are the averages over the OL and UL periods, the weights are the probabilities of OL and UL respectively. In the UL region, the delays are that of an M/D/1 queue with an averaged input rate which can be close to aNh. In the OL zone, the average delay is larger, however the weight is P(>M) which can be small enough as to cancel this contribution. As a further argument, in fig. 7 of [1] appears also the standard deviation, στ, of the delay. We note that for the average, the M/D/1 formula is accurate up to N = 100 but for στ, N* is about 95- 97! This means that the delay (or queue length) in the OL state has great influence even with small frequency of occurrence.

\[ M** = \left[ \sum_{n=0}^{\infty} \sigma(v)_n \right] / U \]  \hspace{1cm} (14)

\[ M* = M + M** U_1 / U \]  \hspace{1cm} (15)

Note that τ1 can also be calculated with this algorithm.

In the OL and UL regimes, the rates, rDL and rUL can be computed by rDL = M* h and rUL = aNh - (M* - aN)h (τ1/τ2). The second equation is obtained by considering that in the interval τ1+τ2 the number of units emitted is (τ1+τ2)aNh, wherever in the OL it is M*hτ1 and in the UL rULτ2. Note that rUL is less than aNh but can be very close if τ1/τ2 is very small.

We mention briefly other published methods. All of these works propose approximation methods to determine the packet losses in a finite buffer multiplexer. In [12] several approximation methods are compared (renewal, MMPP - Markov Modulated Deterministic Process - fluid flow). Three fitting procedures of a 2-MMPP are proposed. In [13], another fitting procedure for a 2-MMPP is proposed. The equivalent of τ1 is obtained by matrix geometric methods, the OL and UL rates are computed with the equivalent of our Pn(k) and Pn(>M). However, τ2 is deduced from the previous calculations i.e. not directly as we do. In [14], a Markov Modulated Deterministic Process (MMDP) is proposed and the fitting procedure for a 2-MMDP consist in calculating the OL and UL rates as in [13], however τ1 and τ2 are recursively evaluated by a method which is different both from [13] and ours.

To complete our study, we need now to find a way to determine ρ*.

**IV CRITERIA FOR DETERMINING ρ*.

We propose here two criteria that allow to calculating ρ*. The first one is the “good old” 3σ rule. The second uses a fluid flow approach.

**A) The 3σ rule.

It can be seen as an empirical rule very often used especially when a Gaussian variable is involved. In (4) let k = M, and call N* and ρ* = aN*/h/C, the associated values for r = 3 (or a fixed r). Then, with A = r^2(1-a)/2, one obtain a quadratic equation in ρ* which solution is

\[ \rho* = 1 + A/M - \sqrt{[(1+A/M)^2 - 1]} \]  \hspace{1cm} (16)

Applying this formula to the cases at hand, we find for case [1], ρ* = 71 and N* = 97, which is pretty close to the experimental value of about 100. For case [2], we find ρ* = 0.59 and N* = 77. This is 17% above the experimental value of about 66. This rule is simple and may be too simple; yet, it is not very far off. As can be seen from Table III, the holding time in the OL can be quite large and so would the delay; the rule does not take this into account. Nevertheless, it can be seen from (16) that, as the capacity and consequently M increases, ρ* tends to 1 from below. This is compatible with our remark about equation (4).
B) A 2-state fluid flow approach.

Using the parameters \((\tau_1, \tau_2, M^*)\) of the AIP, we model the system as a 2-state Markov fluid of packets. The rate in the OL state is \((1/\Delta)(M^* - M)\) and in the UL state \((1/\Delta)(M - M_0)\) (letting \(M_0 = 0\) in the end). Using the standard method (i.e. eigenvalues and eigenvectors), one obtain in units of packets

\[
P(Q \leq x) = 1 - \eta e^{[(1 + \gamma)(1 - \eta)(T/M\tau_1)/(M^* - M)]x}
\]  

(17)

with \(\eta = (M^*/\Delta)\tau_1 / (\tau_1 + \tau_2) = (M^*/M) P_M(>M)\) and \(\gamma = \tau_1/\tau_2\). Then, the average fluid queue length can be calculated as

\[
Q_{FL} = (M^* \tau_1 / T) [\eta/(1 - \eta)] (M^* - M)/(1 + \gamma)
\]  

(18)

Table IV illustrates the fact that (18) gives fair approximations to the simulation results.

<table>
<thead>
<tr>
<th>(N)</th>
<th>110</th>
<th>120</th>
<th>125</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qsim</td>
<td>2.15</td>
<td>10.71</td>
<td>28.52</td>
</tr>
<tr>
<td>Observation</td>
<td>2.14</td>
<td>13.78</td>
<td>30.30</td>
</tr>
</tbody>
</table>

From this, we deduce our second rule as follows. Calculate \(Q_{FL}\) with (18). Calculate \(Q_{MD/1} = \rho^{1/2}(1-\rho)\). Determine \(N^*\) such that the ratio \(Q_{FL}/Q_{MD/1} = 0.1\).

Applying this rule to the cases at hand, we find for case [1], \(N^* = 97\), with a ratio of 0.101. For case [2], \(N^* = 62\), the ratio is 0.9939. Other values are: \(N = 63\), ratio=0.139; \(N = 64\), ratio=0.191; \(N = 66\), ratio=0.373; and, \(N = 70\) ratio=1.078. Clearly, the increase in queue length is very fast. Applying our method to other parameters in [2], we have found that it works pretty well even if the Poisson approximation does not hold because the number of sources is too small.

V CONCLUSIONS AND FUTURE WORK.

Experimental facts indicate that the M/D/1 model is adequate below a certain utilization, \(\rho^*\), and plain wrong above. Moreover, as the capacity increases, \(\rho^*\) increases towards 1. We have provided physical interpretation together with engineering formulae to explain these facts. We also have given computable conditions under which they can happen. We have proposed two rules/methods to determine quite accurately \(\rho^*\).

Our future work include:

- Further experimental and theoretical evaluation of (17), (18) under various distributions including heavy tailed. This is because in [11], it is shown that the asymptotic behavior is independent of the distributions of the ON-OFF periods.
- Further theoretical and experimental evaluation under conditions for which the Poisson approximation does not hold but the fluid approximation can still be used.

VI REFERENCES.


