Performance of Correlated Queues: The Impact of Correlated Service and Inter-arrival Times

Gang Uk Hwang* and Khosrow Sohraby**

*Division of Applied Mathematics
Korea Advanced Institute of Science and Technology
373-1 Kuseong-dong, Yuseong-gu, Taejeon 305-701
South Korea
**Computer Science Telecommunications
University of Missouri - Kansas City
5100 Rockhill Road, Kansas City, MO 64110
U.S.A.

In this paper, we consider correlated queues where the service time of a packet is strongly correlated with its inter-arrival time due to the finite transmission capacity of input and output links. We present an analytical method to derive the LST (Laplace Stieltjes Transform) of the (actual) waiting time distribution and the system time distribution. To investigate the impact of such correlation between service and inter-arrival times on the system performance, we consider a counterpart GI/G/1 queue where the service time and inter-arrival time distributions are the same as in our correlated system, but they are independent. Some numerical examples are provided to show that such correlation gives significant impact on the system performance.

Key words: Correlated Queue, Waiting Time, System Time, Performance Evaluation

1 Introduction

In this paper, we consider correlated queues where service time of a packet and its inter-arrival time are strongly correlated due to the finite transmission capacity of the input and output links. Such correlations naturally arise in a packet switched network with finite speed (i.e., capacity) of links transmitting variable size packets [3]. For instance, consider a finite transmission capacity communication link connecting two nodes, say node 1 and node 2 as in Fig. 1. Packets are assumed to be

* A part of the first author's work was supported in part by KOSEF, and has been done while he was at the University of Missouri - Kansas City as a visiting scholar.
transmitted from node 1 to node 2. In such a system, an inter-arrival time of two consecutive packets at node 2 is a sum of the packet transmission time (the service time at node 1) of the first packet and the inter-occurrence time of two packets. Here, the inter-occurrence time is defined by the time period between the epoch of ending the first packet transmission and the epoch of starting the second packet transmission. Refer to Fig. 1 as well. Since the link connecting two nodes has a finite transmission capacity, say \( \tau \), the link can not send more than \( \tau \times T \) amount of packets from node 1 to node 2 and consequently, larger packets needs longer transmission times at both nodes. Therefore, the inter-arrival times at node 2 are strongly correlated with the service times at node 2. This model gives us a good motivation for our study.

In this paper we analyze a queueing system in node 2 which stores packets coming from a finite transmission capacity input link and forwards them to an output link having the same transmission capacity as the input link has. The packet arrival process to the queue is modelled by an alternating renewal process having two types of periods: say OFF and ON periods. OFF periods capture the inter-occurrence times of consecutive packets at node 1 and the ON periods capture the transmission times of packets at node 1, so that an inter-arrival time between two consecutive packets at the queue consists of an OFF period and an ON period. Since the queue in node 2 can transmit a packet through its output link only after it receives the last bit of the packet, in the analysis we assume that, at the end of each ON period we have a packet arrival of which service time is the same as the ON period that it follows. (Actually, we should further assume that the packet header processing times are negligible). Further, we allow the length of an OFF period to be zero because node 1 can continuously transmit packets to node 2 if it has packets to transmit (see Fig. 1).

In the analysis we assume that the number of consecutive ON periods between two OFF periods of positive length, is geometrically distributed, and OFF periods of positive length are exponentially distributed. Each ON period is assumed to be hyper-exponentially distributed.

Earlier works related to this issue can be found in [3,4] and the references therein. In [3] they considered a system with Poisson arrivals, where the service time \( B_n \) of the \( n \)-th packet is proportional to the inter-arrival time \( A_n \) between the \( n - 1 \)-st
and $n$-th packets, i.e., $B_n = \zeta A_n$ where $\zeta$ is a positive constant. They obtained the LST (Laplace Stieltjes Transform) of the distribution of the system time (defined by the waiting time in the queue plus the service time) by solving a linear functional equation [7] derived from the evolution equation of the system time. However, they only consider the case of $\zeta \neq 1$ due to the technical difficulties associated with the solution of the functional equation. By our assumption that the input link capacity and the output link capacity are the same, we have $\zeta = 1$ in this paper and consequently, our result fills the gap in [3]. Further, the model considered in this paper is a natural generalization of the model in [3] for the case of $\zeta = 1$ in the sense that our hyper-exponential distribution assumption for ON periods is more general than the exponential distribution assumption in [3]. We succeed in obtaining the LST of the (actual) waiting time by taking our embedded points different from those in [3].

There are some other earlier works [1,5,9,11] based on fluid flow models with regard to the correlations considered in this paper. However, as illustrated in [3] fluid flow models do not account for the granularity of the arrival and service processes and hence result in inaccuracy in predicting the system performance. Furthermore, the fluid flow model cannot be used to obtain the (actual) waiting time.

To investigate the impact of the correlations considered in this paper on system performance, we also consider and analyze a counterpart GI/G/1 queue where inter-arrival time and service time distributions are the same as in our correlated system, but they are independent. We provide numerical examples to investigate the impact of the correlations on the system performance. Numerical results illustrate that the uncorrelated system overestimates the system performance, which is in accordance with the results obtained in [3]. Another interesting result is that the change of the system performance for the correlated queue due to the change of statistical characteristics of the ON period (i.e., the packet size distribution) is much less significant than that for the uncorrelated queue. Other interesting observations for the correlated queue are also provided.

The paper is organized as follows: In section 2, we give an exact mathematical model and analyze it to derive the LSTs of the (actual) waiting time distribution and the system time distribution. In section 3, we consider the counterpart GI/G/1 queue and analyze it to derive the LST of the (actual) waiting time. In section 4, we give numerical examples to investigate the impact of the correlations on system performance. In section 5, we give our conclusions.

2 Mathematical Modelling and Analysis

In this section, we consider a correlated queue of a node in a network where the input process to the queue is governed by an alternating ON and OFF process and input link capacity and output link capacity are the same. As mentioned in section
1, OFF periods capture the inter-occurrence times of packets and ON periods capture the transmission times of packets at the previous node in the network. Since the queueing system we consider can serve a packet only after it receives the whole packet, we assume that we have a new packet arrival at the end of each ON period. Since we assume that input link capacity and output link capacity are the same, the service times of packets are equal to the ON periods that they follow. At the end of each ON period either an OFF period or another ON period begins because we allow OFF periods of length 0. Note that consecutive ON periods capture consecutive packet transmissions at the previous node in the network. After each OFF period we assume a new ON period always begins.

In the analysis, ON periods are assumed to be generated according to a hyper-exponential distribution with $M$ phases, i.e. the density function $f(x)$ of the ON periods is given by

$$f(x) = \sum_{i=0}^{M-1} \theta_i \delta_i e^{-\delta_i x}, \quad \sum_{i=0}^{M-1} \theta_i = 1, \theta_i \geq 0.$$ 

We further assume that, at the end of each ON period either an OFF period begins with probability $q = 1 - p$ or another ON period begins with probability $p$. Then the total number of ON periods between two consecutive OFF periods of length greater than 0, is geometrically distributed. For technical difficulty, we assume $p < q \leq 1$. The case of $p \geq q$ is left for further study. The OFF periods of positive length are assumed to be generated according to an exponential distribution with parameter $\beta$.

Let $I_n$ and $J_n$ ($n \geq 1$) denote the lengths of the $n$-th ON and OFF periods, respectively. We take the beginning epochs of ON periods as our embedded points and let $W_n$ denote the unfinished work of the system, defined by the amount of work in the system at the $n$-th embedded point ($n \geq 0$). Time 0 is considered as the 0-th embedded point and we assume that the first ON period starts at time 0. Then $W_n$ satisfies the following evolution equation:

$$W_0 = 0, \quad W_{n+1} = [(W_n - I_{n+1})^+ + I_{n+1} - J_{n+1}]^+, \quad n \geq 0,$$

where $(A)^+ = \max(A, 0)$. Let $S_{n+1}$ denote the type of the exponential distribution that $I_{n+1}$ takes. We then have $S_{n+1} = i$ with probability $\theta_i, 0 \leq i \leq M - 1$ and consequently, $I_{n+1}$ is according to an exponential distribution with parameter $\delta_i$.

Let $W_n(s)$ be the LST (Laplace Stieltjes Transform) of $W_n$ for $\Re(s) \geq 0$. Then, conditioning that $S_{n+1} = i$ we have the following equations:
Substituting (3), (4) and (5) into (2) yields

\[
W_{n+1}(s) = E\left[e^{-s((W_n-I_{n+1})^+ + I_{n+1} - J_{n+1})^+}\right] \\
= E\left[e^{-s((W_n-I_{n+1})^+ + I_{n+1} - J_{n+1})^+}; W_n = 0\right] \\
+ E\left[e^{-s((W_n-I_{n+1})^+ + I_{n+1} - J_{n+1})^+}; W_n > 0\right] \\
= \sum_{i=0}^{M-1} \theta_i P\{W_n = 0\} E\left[e^{-s[I_{n+1} - J_{n+1}]^+}|S_{n+1} = i\right] \\
+ \sum_{i=0}^{M-1} \theta_i \int_{\omega=0}^{\infty} dW_n(\omega) \left\{ \int_{x=0}^{\omega} \delta_i e^{-\delta_i x} E\left[e^{-s(\omega - J_{n+1})^+}\right]dx \right\} \\
+ \int_{x=\omega}^{\infty} \delta_i e^{-\delta_i x} E\left[e^{-s(x - J_{n+1})^+}\right]dx.
\]

Observing that another ON period occurs after the \(n+1\)st ON period with probability \(p\) (in this case we have \(J_{n+1} = 0\)) and that an OFF period occurs after the \(n+1\)st ON period with probability \(q = 1 - p\) (in this case the length of \(J_{n+1}\) is according to an exponential distribution with parameter \(\beta\)), we obtain that

\[
E\left[e^{-s(\omega - J_{n+1})^+}\right] \\
= pe^{-s\omega} + q \left\{ \int_{y=0}^{\omega} \beta e^{-\beta y} e^{-s(\omega - y)} dy + \int_{y=\omega}^{\infty} \beta e^{-\beta y} dy \right\} \\
= pe^{-s\omega} + q \left\{ \frac{\beta}{\beta - s} e^{-s\omega} - \frac{s}{\beta - s} e^{-\beta \omega} \right\} \\
= \frac{\beta - ps}{\beta - s} e^{-s\omega} - \frac{qs}{\beta - s} e^{-\beta \omega},
\]

so it immediately follows that

\[
\int_{x=0}^{\omega} \delta_i e^{-\delta_i x} E\left[e^{-s(\omega - J_{n+1})^+}\right]dx \\
= \frac{\beta - ps}{\beta - s} e^{-s\omega} - \frac{qs}{\beta - s} e^{-\beta \omega} - \frac{\beta - ps}{\beta - s} e^{-(\delta_i+s)\omega} + \frac{qs}{\beta - s} e^{-(\delta_i+s)\omega}, \\
\]

(3)

\[
\int_{x=\omega}^{\infty} \delta_i e^{-\delta_i x} E\left[e^{-s(x - J_{n+1})^+}\right]dx \\
= \frac{\beta - ps}{\beta - s} \frac{\delta_i}{\delta_i + s} e^{-(\delta_i+s)\omega} - \frac{qs}{\beta - s} \frac{\delta_i}{\delta_i + s} e^{-(\delta_i+s)\omega} - \frac{\beta - ps}{\beta - s} \frac{\delta_i}{\beta + \beta} e^{-(\delta_i+s)\omega}.
\]

(4)

In addition, from (4) we obtain

\[
E\left[e^{-s[I_{n+1} - J_{n+1}]^+}|S_{n+1} = i\right] \\
= \int_{x=0}^{\omega} \delta_i e^{-\delta_i x} E\left[e^{-s(x - J_{n+1})^+}\right]dx \\
= \frac{\beta - ps}{\beta - s} \frac{\delta_i}{\beta + s} - \frac{qs}{\beta - s} \frac{\delta_i}{\beta + s}.
\]

(5)

Substituting (3), (4) and (5) into (2) yields
For the analysis, we let
\[ W_n \]
be the unfinished work of the system at an arbitrary embedded epoch in the steady state, and \( W(s) \) be the LST of \( W \), i.e., \( W(s) = \lim_{n \to \infty} W_n(s) \). As we let \( n \to \infty \) in (6) we obtain
\[
W(s) = \sum_{i=0}^{M-1} \theta_i \frac{\beta - ps}{q(\delta_i + s)} W(\delta_i + s) + W(\beta) - \sum_{i=0}^{M-1} \theta_i \frac{\beta}{\delta_i + \beta} W(\delta_i + \beta).
\]
(7)

For the analysis, we let
\[
c = W(\beta) - \sum_{i=1}^{M-1} \theta_i \frac{\beta}{\delta_i + \beta} W(\delta_i + \beta),
\]
\[
a_i(s) = \theta_i \frac{\beta - ps}{q(\delta_i + s)}.
\]
We get that $c$ is a constant since we may assume that all the parameters involved are given. Then we get

$$W(s) = \sum_{i=0}^{M-1} a_i(s)W(\delta_i + s) + c.$$  \hspace{1cm} (8)

By successive substitutions of (8) into itself we get

$$W(s) = \sum_{i_0=0}^{M-1} a_{i_0}(s)W(\delta_{i_0} + s) + c$$

$$= \sum_{i_0=0}^{M-1} a_{i_0}(s) \left\{ \sum_{i_1=0}^{M-1} a_{i_1}(\delta_{i_0} + s)W(\delta_{i_1} + \delta_{i_0} + s) + c \right\} + c$$

$$= \sum_{i_0=0}^{M-1} a_{i_0}(s) \left[ \sum_{i_1=0}^{M-1} a_{i_1}(\delta_{i_0} + s)W(\delta_{i_1} + \delta_{i_0} + s) + \left\{ \sum_{i_0=0}^{M-1} a_{i_0}(s) + 1 \right\} c \right]$$

$$= \ldots$$

$$= \prod_{k=0}^{m-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j})W(s + \sum_{j=0}^{k-1} \delta_{i_j}) + \sum_{l=0}^{m-1} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j})c$$

$$= \ldots$$

$$= \prod_{k=0}^{M-1} \sum_{i_k=0}^{k-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j})W(\infty) + \prod_{l=0}^{M-1} \sum_{i_k=0}^{k-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j})c, \hspace{1cm} (9)$$

where empty sum and product are equal to 0 and 1, respectively, and the above equation holds only when all the quantities involved are finite.

Noting that

$$\lim_{k \to \infty} \left| a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j}) \right| = p\theta_{i_k}/q,$$

we have

$$\left| \sum_{i_k=0}^{M-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j}) \right| < 1$$

for sufficiently large $k$ by our assumption that $q > p$. Hence,

$$\prod_{k=0}^{\infty} \sum_{i_k=0}^{M-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j})$$

is bounded. Since $\sum_{l=0}^{\infty} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j})$ is bounded as well (because $\left| \sum_{i_k=0}^{M-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j}) \right| < 1$ for sufficiently large $k$) and $W(\infty) = 0$, from (9) we get

$$W(s) = \sum_{l=0}^{\infty} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k}(s + \sum_{j=0}^{k-1} \delta_{i_j})c. \hspace{1cm} (10)$$
From the fact that \( W(0) = 1 \) the constant \( c \) is given by
\[
 c = \frac{1}{\sum_{l=0}^{\infty} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k} (\sum_{j=0}^{k-1} \delta_{ij})},
\]
from which we finally have
\[
 W(s) = \frac{\sum_{l=0}^{\infty} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k} (s + \sum_{j=0}^{k-1} \delta_{ij})}{\sum_{l=0}^{\infty} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k} (\sum_{j=0}^{k-1} \delta_{ij})}.
\]

For later use we here compute the expectation \( E[W] \) of \( W \) as well. Observing that
\[
 \frac{d}{ds} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k} (s + \sum_{j=0}^{k-1} \delta_{ij}) = \sum_{l=0}^{\infty} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} \sum_{i_n=0, n \neq k}^{M-1} \sum_{j=0}^{n-1} \sum_{i_k=0}^{M-1} \theta_{i_k} (-p\delta_{ik} - \beta) \quad q \left( \sum_{j=0}^{k} \delta_{ij} \right)^2
\]
we obtain
\[
 E[W] = \frac{\sum_{l=1}^{\infty} \sum_{k=0}^{l-1} \prod_{n=0, n \neq k}^{M-1} \sum_{i_n=0}^{M-1} a_{i_n} (\sum_{j=0}^{n-1} \delta_{ij}) \sum_{i_k=0}^{M-1} \theta_{i_k} (-p\delta_{ik} + \beta) \quad q \left( \sum_{j=0}^{k} \delta_{ij} \right)^2}{\sum_{l=0}^{\infty} \prod_{k=0}^{l-1} \sum_{i_k=0}^{M-1} a_{i_k} (\sum_{j=0}^{k-1} \delta_{ij})}.
\]

We are now ready to derive the LSTs \( A(s) \) and \( Z(s) \) of the (actual) waiting time \( A \) and the system time \( Z \) of a packet in the steady state, respectively. From the equation
\[
 A = (W - I)^+, \quad Z = (W - I)^+ + I
\]
where \( I \) is the generic random variable for \( I_n \), we have
\[
 A(s) = E \left[ e^{-s[(W-I)^+]} \right]
\]
\[
 = E \left[ e^{-s[(W-I)^+]}, W = 0 \right] + E \left[ e^{-s[(W-I)^+]}, W > 0 \right]
\]
\[
 = P\{W = 0\} + \sum_{i=0}^{M-1} \theta_i \int_0^\infty dW(\omega) \left\{ \int_0^x \delta_i e^{-\delta_i x} e^{-s(\omega-x)} dx + \int_x^\infty \delta_i e^{-\delta_i x} dx \right\}
\]
\[
 = P\{W = 0\} + \sum_{i=0}^{M-1} \theta_i \int_0^\infty dW(\omega) \left\{ \frac{\delta_i}{\delta_i - s} e^{-s\omega} - \frac{s}{\delta_i - s} e^{-\delta_i \omega} \right\}
\]
\[
 = P\{W = 0\} + \sum_{i=0}^{M-1} \theta_i \left\{ \frac{\delta_i}{\delta_i - s} W(s) - \frac{s}{\delta_i - s} P\{W = 0\} \right\}
\]
\[
 = \sum_{i=0}^{M-1} \theta_i \left\{ \frac{\delta_i}{\delta_i - s} W(s) - \frac{s}{\delta_i - s} W(\delta_i) \right\}.
\]
From (8) we observe that
\[
\sum_{i=0}^{M-1} \frac{\theta_i}{\delta_i} W(\delta_i) = \frac{q(1-c)}{\beta}.
\]

Consequently, from (14) we find the expectation \(E[A]\) of the (actual) waiting time as follows:
\[
E[A] = E[W] + \frac{q(1-c)}{\beta} - \sum_{i=0}^{M-1} \frac{\theta_i}{\delta_i}.
\]  

(15)

Similarly,
\[
Z(s) = E\left[e^{-s(W-I+I)}\right]
\]
\[
= E\left[e^{-s[(W-I)+I]}; W = 0\right] + E\left[e^{-s[(W-I)+I]; W > 0}\right]
\]
\[
= P\{W = 0\} E[e^{-sI}]
+ \sum_{i=0}^{M-1} \theta_i \int_0^\infty dW(\omega) \left\{\int_{x=0}^\omega \delta_i e^{-\delta_i x} e^{-sx} dx + \int_{x=\omega}^\infty \delta_i e^{-\delta_i x} e^{-sx} dx\right\}
\]
\[
= P\{W = 0\} E[e^{-sI}]
+ \sum_{i=0}^{M-1} \theta_i \int_0^\infty dW(\omega) \left\{e^{-s\omega} - \frac{s}{\delta_i + s} e^{-(s+\delta_i)\omega}\right\}
\]
\[
= \sum_{i=0}^{M-1} \theta_i \{W(s) - \frac{s}{\delta_i + s} W(\delta_i + s)\}
\]
\[
= W(s) - \frac{qs}{\beta - ps} (W(s) - c) \quad \text{(by (8))}
\]
\[
= \frac{\beta - s}{\beta - ps} W(s) + \frac{sq}{\beta - ps} c.
\]

(16)

From (16) we find the expectation \(E[Z]\) of the system time \(Z\) as follows:
\[
E[Z] = E[W] + \frac{q(1-c)}{\beta}.
\]

3 The Counterpart GI/G/1 Queue

In this section we consider the counterpart GI/G/1 queue to the model considered in section 2. In this case, the inter-arrival time and the service time are independent, and the LSTs \(A(s)\) and \(B(s)\) of inter-arrival time distribution and service time distribution, respectively, are given by
\[
A(s) = q \frac{\beta}{s + \beta} \left(\sum_{i=0}^{M-1} \frac{\delta_i}{s + \delta_i}\right) + p \sum_{i=0}^{M-1} \frac{\delta_i}{s + \delta_i}, \quad B(s) = \sum_{i=0}^{M-1} \frac{\theta_i}{s + \delta_i},
\]
where \( p \) and \( q \) are the same as in section 2. Without loss of generality, we may assume that all \( \delta_i \) are distinct. We further assume that \( \beta/p \neq \delta_i, i = 0, \ldots, M - 1 \) for simplicity.

Let \( W^e_n \) be the waiting time at the \( n \)-th arrival. Then the evolution equation for \( \{W^e_n\} \) is give by

\[
W^e_0 = 0, \\
W^e_{n+1} = [W^e_n + B_{n+1} - A_{n+1}]^+, \quad n \geq 0,
\]

where \( A_n \) and \( B_n \) denote the random variables for the \( n \)-th inter-arrival time and the \( n \)-th service time, respectively, and \([x]^+\) denotes \( \max(0, x) \). Let \( W^e(s) \) be the LST of \( W^e = \lim_{n \to \infty} W^e_n \). Observing that

\[
U(s) \triangleq B(s)A(-s) = \left[ \sum_{i=0}^{M-1} \theta_i \frac{\delta_i}{s + \delta_i} \right] \left[ \sum_{j=0}^{M-1} \theta_j \frac{\delta_j}{-s + \delta_j} \right] \frac{(\beta - ps)}{(-s + \beta)},
\]

we have

\[
1 - U(s) = \frac{a(s)}{\prod_{i=0}^{M-1} (s + \delta_i) \prod_{j=0}^{M-1} (-s + \delta_j)(-s + \beta)}
\]

\[
a(s) \triangleq \prod_{i=0}^{M-1} (s + \delta_i) \prod_{j=0}^{M-1} (-s + \delta_j)(-s + \beta)
\]

\[
- \left[ \sum_{i=0}^{M-1} \theta_i \delta_i \prod_{j=0, j \neq i}^{M-1} (s + \delta_j) \right] \left[ \sum_{k=0}^{M-1} \theta_k \delta_k \prod_{l=0, l \neq k}^{M-1} (s + \delta_l) \right] (\beta - ps)
\]

Our next step is to factor \( 1 - U(s) \) as follows:

\[
1 - U(s) = \frac{\psi_+(s)}{\psi_-(s)}
\]

where \( \psi_+(s) \) is analytic and continuous on \( \text{Re}(s) > s_- \) for some \( s_- \leq 0 \), \( \psi_-(s) \) is analytic and continuous on \( \text{Re}(s) < s_+ \) for some \( s_+ \geq 0 \) [2,6,10].

From the definition of \( a(s) \) we know that the polynomial \( a(s) \) should have \( 2M + 1 \) roots. In the following theorem we give the property of \( 2M + 1 \) roots of \( a(s) \).

**Theorem 1** The polynomial \( a(s) \) has \( M \) distinct positive real roots, a single root at \( s = 0 \) and \( M \) distinct negative real roots.

**Proof.** First putting \( s = 0 \) we see \( s = 0 \) is a root of \( a(s) \). Let \( \delta_{-1} = 0 \) and \( \delta_M = \)
We assume \( \delta_{i^*} < \beta/p < \delta_{i^*+1} \) for some \( i^* \in \{-1, 0, \cdots, M - 1\} \). Next from the definition of \( a(s) \) it is easy to show the sign of \( a(s) \) at \( s = \delta_m, 0 \leq m \leq M - 1 \) and \( \beta/p \) as follows:

\[
\text{sign of } a(\delta_m) = \begin{cases} 
(-1)^{m+1} & \text{if } m \leq i^* \\
(-1)^m & \text{if } m \geq i^* + 1
\end{cases}
\]

\[
\text{sign of } a(\beta/p) = (-1)^{i^*}.
\]

In addition we see that \( a(s) \to -\infty \) as \( s \to \infty \) when \( M \) is even and \( a(s) \to \infty \) as \( s \to \infty \) when \( M \) is odd. From the above information when \( i^* = -1 \) we can easily check that the \( M \) distinct roots, say \( r_1, \cdots, r_M \) lie as follows:

\[
\beta/p < r_1 < \cdots < r_M < \delta_{M-1}.
\]

Similarly, when \( 0 \leq i^* \leq M - 2 \) the \( M \) distinct roots lie as follows:

\[
\delta_0 < r_1 < \cdots < \delta_{i^*} < r_{i^*+1} < \beta/p < r_{i^*+2} < \cdots < r_M < \delta_{M-1}
\]

Lastly when \( i^* = M - 1 \) the \( M \) distinct roots lie as follows:

\[
\delta_0 < r_1 < \cdots < r_{M-1} < \delta_{M-1} < r_M < \beta/p.
\]

For the \( M \) distinct negative roots, say \( r_{-1}, \cdots, r_{-M} \) observing that the sign of \( a(-\delta_m) \) is \((-1)^{m+1}\) and that \( a'(0) < 0 \) we see

\[
-\delta_{M-1} < r_{-M} < \cdots < r_{-2} < -\delta_0 < r_{-1},
\]

which completes the proof.

For the analysis we assume, as in the proof of Theorem 1, that the \( 2M + 1 \) roots satisfy

\[
r_{-M} < r_{-M+1} < \cdots < r_{-1} < r_0 = 0 < r_1 < \cdots < r_M.
\]

From the analytic property of \( \psi_+(s) \) and \( \psi_-(s) \), we factor \( 1 - U(s) \) as follows:

\[
\psi_+(s) = \frac{k \prod_{i=-1}^{-M} (s - r_i)}{\prod_{i=0}^{-M} (s + \delta_i)}, \quad \psi_-(s) = \frac{\prod_{i=0}^{-M} (-s + \delta_i) (-s + \beta)}{\prod_{i=0}^{-M} (s - r_i)},
\]

from which we have [2,6,10]

\[
W^{e*}(s) = \frac{k \prod_{i=-1}^{-M} (s + \delta_i)}{\prod_{i=-1}^{-M} (s - r_i)}. \tag{17}
\]

From \( W^{e*}(0) = 1 \) we obtain

\[
k = \frac{\prod_{i=0}^{-M} (\delta_i)}{\prod_{i=-1}^{-M} (-r_i)},
\]
so that we finally get

\[ W^{e*}(s) = \frac{\prod_{i=-1}^{-M}(-r_i) \prod_{i=0}^{M-1}(s + \delta_i)}{\prod_{i=0}^{M-1}(\delta_i) \prod_{i=-1}^{-M}(s - r_i)}. \] (18)

By differentiating (18) and putting \( s = 0 \) we obtain the means \( EW^e \) and \( ES^e \) of the (actual) waiting time and the system time, respectively as follows:

\[
EW^e = \frac{\prod_{i=-1}^{-M}(-r_i)}{\prod_{i=0}^{M-1}(\delta_i)} \sum_{i=0}^{M-1} \frac{r_{-i-1} + \delta_i}{r_{-i-1}^2} \prod_{j=0,j\neq i}^{M-1} \frac{\delta_j}{-r_{-j-1}},
\]

\[
ES^e = EW^e + \sum_{i=0}^{M-1} \frac{\theta_i}{\delta_i}.
\]

### 4 Numerical Results

In the numerical analysis we consider some special examples for our model and give numerical results to investigate the impact of correlations.

In our first example, we consider the correlated queue with \( M = 1 \). In this case the density function \( f(x) \) of an inter-arrival time is given by

\[
f(x) = pf_{\delta}(x) + qf_{\beta} \ast f_{\beta}(x),
\]

where \( f_{\delta}(x) = e^{-\delta x} \) and \( f_{\beta}(x) = e^{-\beta x} \), and \( \ast \) denotes the convolution of two functions. To examine the impact of the correlations between inter-arrival and service times, we consider its counterpart GI/M/1 system [6] whose inter-arrival time density function is given by \( f(x) \) and service time density function is given by \( f_{\delta}(x) \). In the numerical computation, we use \( \delta = 1.2 \) and \( \beta = 1.0 \), and the value of \( p \) is changed from \( p = 0.1 \) to \( p = 0.45 \). Note that as the value of \( p \) increases, the traffic load also increases. The result is given in Fig. 2. As shown in the figure the counterpart GI/M/1 model overestimates the system performance and the overestimation becomes more significant as the traffic load increases, which is in accordance with the previous results [3]. So, the correlations between inter-arrival and service times should be carefully captured to understand the system performance more exactly. In addition, for the correlated queue the mean waiting time increases as the traffic load increases as for its counterpart GI/M/1 queue, but the change of the correlated queue is much less than that of the counterpart (uncorrelated) queue.

Next, to examine the impact of the correlations under the change of the mean of the ON period we use the same model and parameters as in the first example except \( p = 0.45 \). We change the rate \( \delta \) from 0.6 to 1.6. Accordingly, the mean of the ON period is changed from 0.625 to 1.67. Fig. 3 shows that the mean waiting time increases as
the mean of the ON period increases. Further, the change of the expected waiting time for the correlated queue is less significant (due to the change of the ON period distribution) than that for the counterpart GI/M/1 queue as verified in Fig. 2.

In the next two examples, we fix the offered load, denoted by $\rho$, and change the value of $p$ from 0.1 to 0.45 to check the impact of the correlations under a given traffic load. In the numerical computation, we use the same model as in Fig. 2 and set $\beta = 1.0$. Therefore, the mean of the ON period, $1/\delta$, decreases as $p$ increases. Fig. 4 and Fig. 5 show the results for $\rho = 0.7$ and $\rho = 0.85$, respectively. The two figures strongly support the conclusions given in the previous two examples.
Next, we consider the correlated queue with ON periods being hyper-exponentially distributed and change the value of $M$ to compare the performance of our system with that of its counterpart GI/G/1 queue. In the numerical computation, the density function $f(x)$ of an ON period is given by

$$f(x) = \sum_{i=0}^{M-1} \theta_i \delta_i e^{-\delta_i x},$$

where

$$\theta_i = \frac{(1 - \theta)\theta^i}{1 - \theta^M}, \quad \delta_i = \frac{\mu}{\gamma^i}$$
for positive real values $\theta, \mu$ and $\gamma$. Note that $f(x)$ is the density function of a TPT (Truncated Power Tail) distribution, and as $M \to \infty$, the distribution converges to a PT (Power Tail) distribution [8]. We use $\mu = 1.2, \beta = 1.0, \theta = 0.5, \gamma = 3/2$ and $p = 0.1$. The result is given in Fig. 6. As shown in the figure the GI/M/1 model overestimates the system performance as observed before. Further, since the variance of the ON period is increasing as $M$ is increasing for this case, we see that the overestimation becomes more significant as the variance of the ON period becomes large.

The next example is given to check the impact of the variance of the ON period on system performance for the correlated case. Here, for comparison purpose we fix the mean of the ON period and change the variance of the ON period. In the numerical computation, we set $M = 2, \beta = 1.0, p = 0.3$ and $\delta_0 = 2.0$, and change the value of $\theta_0$. Note that, since $M = 2$, we have $\theta_1 = 1 - \theta_0$. The results are given in Fig. 7 where we fix the mean of the ON period distribution at $1/1.5$ and change the value of $\theta_0$ from 0.1 to 0.9. Accordingly, the variance of the ON period increases from 0.447 to 0.694. For comparison purpose, we also plot the mean waiting time in Fig. 7 when the distribution of ON period is exponential with the same mean $1/1.5$. Note that the dotted line of the mean waiting time for the exponential ON period case in Fig. 7 is irrespective of the value of $x$ axis. As shown in the figure, when the variance of the ON period increases with the mean of the distribution fixed, the mean waiting time of a packet increases. Hence, the ON period distribution should have smaller variance to get smaller mean waiting time for the correlated queue.

In the last example, we consider two equivalent systems in the sense that the total workload carried between two OFF periods of length greater than 0, are the same in both systems. We first consider a system with $p = 0$ and the density function of
an ON period is given by
\[ f_1(x) = \alpha e^{-\alpha x}, \]
and its Laplace transform is given by
\[ \frac{\alpha}{s + \alpha}. \]

This system with \( p = 0 \) is referred to as a system of type I for convenience. Next, we construct the other system where \( p > 0 \) and ON periods are exponentially distributed with parameter \( \delta \). Then, for given \( p \) and \( \delta \) the Laplace Transform \( I(s) \) of the distribution for the total periods between two OFF periods of length greater than 0, is given by

\[
I(s) = \sum_{k=1}^{\infty} q p^{k-1} \left( \frac{\delta}{\delta + s} \right)^{k} = \frac{q \delta}{s + q \delta}, \quad q = 1 - p.
\]

So, if we fix the quantity \( \delta q \), then \( I(s) \) is invariant. Furthermore, if we set \( q \delta = \alpha \), then \( I(s) \) is also the Laplace Transform of an ON period for the system of type I. From our observation we can construct a system of type II as follows: The system of type II allows \( p \) to be greater than 0 and the density function of an ON period is given by

\[ f_{II}(x) = \delta e^{-\delta x}, \quad \delta = \frac{\alpha}{q}. \]

We could interpret the behaviors of both systems as follows: The system of type II allows multiple packets between OFF periods (of positive length), while the system of type I allows only a single packet between OFF periods (of positive length). Since both systems have the same amount of workload between two OFF periods (of positive length), the system of type II can be viewed as a variant of the system of type I where a packet can be divided into smaller packets to send. In the numerical
computation we use \( \alpha = 1.2 \) and change the value of \( p \) from 0.1 to 0.45. The results are given in Fig. 8 where we can see that, as \( p \) increases the mean waiting time decreases in the system of type II. Note that the mean waiting time for the system of type I is invariant in the figure because it does not depend on the value of \( p \). We also see that the performances of the system of type I are worse than those of the system of type II, and that the difference in system performance is getting significant as the value of \( p \) increases. This is because more consecutive packet arrivals of smaller size are likely to occur in the system of type II as the value of \( p \) increases, which prevents long delay in the system.

5 Conclusion

In this paper we considered a correlated queue where input and output link transmission capacity are the same and gave an analytic method to derive the LST of the (actual) waiting time and the system time of our system. We used embedded methods to derive the LSTs. From a number of numerical examples, we showed that NOT modelling the correlation between service and inter-arrival times over-estimates system performance and such correlations give a significant impact on system performance. We also showed that the system performance depends largely on the ON period distribution (i.e., the packet size distribution) for the correlated queue. However, the change of the system performance for the correlated queue due to the change of statistical characteristics of the ON period (i.e., the packet size distribution) is much less significant than that for the uncorrelated queue.
Acknowledgment

The authors would like to thank the anonymous referees for their valuable comments.

References


