EURANDOM PREPRINT SERIES
2013-003
January, 2013

Cyclic-type Polling Models with Preparation Times

N. Perel, J.L. Dorsman, M. Vlasiou
ISSN 1389-2355

Cyclic-type Polling Models with Preparation Times

N. Perel[1], J.L. Dorsman[2,3], M. Vlasiou[2,3]

[1]*Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel*

[2]*Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands*

[3]*Probability and Stochastic Networks, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands*

perelnir@post.tau.ac.il, {j.l.dorsman, m.vlasiou}@tue.nl

Keywords: Layered Queuing Networks : Polling Systems : Queuing Theory

Abstract: We consider a system consisting of a server serving in sequence a fixed number of stations. At each station there is an infinite queue of customers that have to undergo a preparation phase before being served. This model is connected to layered queuing networks, to an extension of polling systems, and surprisingly to random graphs. We are interested in the waiting time of the server. The waiting time of the server satisfies a Lindley-type equation of a non-standard form. We give a sufficient condition for the existence of a limiting waiting time distribution in the general case, and assuming preparation times are exponentially distributed, we describe in depth the resulting Markov chain. We provide detailed computations for a special case and extensive numerical results investigating the effect of the system's parameters to the performance of the server.

1 INTRODUCTION

We study a model that involves one server polling multiple stations. The server visits N stations in a cyclic order, serving one customer at a time. At each station there is an infinite queue of customers that needs service. Before being served by the server, a customer must first undergo a preparation phase. Thus the server, after having finished serving a customer at one station, may have to wait for the preparation phase of the customer at the next station to be completed. Immediately after the server concludes his service at some station, another customer from the infinite queue begins his preparation phase there. Our goal is to analyse the transient, as well as the long-run, probabilistic behaviour of this system by quantifying the waiting time of the server, which is directly connected to the system's efficiency and throughput.

This model finds wide applications in enterprise systems, for example when the order of service of the customers is important. A typical operating strategy in healthcare clinics is to have a specialist rotate among several stations. The preparation phase represents the preliminary service a patient typically receives from an assistant or a nurse. The model, however, originates from warehousing. It was introduced in (Park et al., 2003), who consider a storage facility with bi-directional carousels, where a picker serves in turns the carousels. The preparation phase represents the rotation time the carousel needs to bring the item to the origin, while the service time is the actual picking time. The authors study the case of two carousels under specific assumptions. Later on, this special case for two stations has been further analysed under general distributional assumptions in (Vlasiou, 2006). The model we consider in this paper generalises this work from two stations to multiple stations. This extension leads to significant challenges in analysis, but provides valuable managerial insights. Little work has been done on multiple-carousel warehouse systems. Multiple-carousel problems differ intrinsically from single-carousel problems in a number of ways. Such systems tend to be more complicated. The system cannot be viewed as a number of independently operating carousels (McGinnis et al., 1986), since the two separate carousels interact by means of the picker that is assigned to them. Almost all studies involving systems with more than two carousels resort to simulation; see (Litvak and Vlasiou, 2010) for a complete literature review. This paper offers the first analytic results for such systems.

This system can also be viewed as an extension of a 1-limited polling-type system; cf. (Boxma and Groenendijk, 1988; Eisenberg, 1979; van Vuuren and Winands, 2007). In general, polling models have attracted a lot of attention in the literature; see e.g. (Boon et al., 2011; Takagi, 1986; Yechiali, 1993), and the extensive references therein. Limited polling

systems are notoriously difficult to analyse as the k-limited service discipline does not satisfy the so-called branching property; see (Resing, 1993). In our case, we have the added difficulty of an additional preparation phase before service. Traditional approximation methods seem to be of little help. For example, heavy traffic approximations for polling systems seem to be mainly suitable for the study of characteristics of the *customers*, as is typical in the polling literature, rather than the server, as is our case here; cf. (van der Mei, 2007). Here, we assume that there are always waiting customers in front of each station. Therefore, the analysis of the model is parallel to the study of the server of a polling-type system, in which each of the queues is overloaded. As such, heavy traffic diffusion approximations cannot be utilised for this system, since we consider a system that is overloaded, rather than critically loaded, while fluid approximations are equally not straightforward due to the additional preparation phase.

This model is a layered network in which a server, while executing a service, may request a higher-layer service and wait for it to be completed. Layered queuing networks occur naturally in all kinds of information and e-commerce systems, grid systems, and real-time systems such as telecom switches; see (Franks et al., 2009) and references therein for an overview. Layered queues are characterised by simultaneous or separate phases where entities are no longer classified in the traditional roles of "servers" and "customers", but may have also a dual role of being either a server to other entities (of lower layers) or a customer to higher-layer entities. Think of a peer-to-peer network, where users are both customers when downloading a file, but also servers to users who download their files. For our system, one may view the preparation time of a customer as a first phase of service. The service station (lower layer) acts in this case as a *server*. However, the second phase of service (the actual operation) does not necessarily follow immediately. The service station might have to 'wait' for the server to finish working on other stations. At this stage, the service stations act as *customers* waiting to be served by the higher layer, the server. Thus, we see that each service station acts both as a 'server' (preparing the customer) and as a 'customer' (waiting until the server completes his tasks in the previous stations).

This model leads to a Lindley-type equation, which for two stations leads to the equation (in its steady-state form) as $W \stackrel{\mathcal{D}}{=} \left(B - A - W\right)^+$. Here, B denotes the preparation time, A denotes service time and W is the waiting time of the server. The difference from the original (Lindley, 1952) equation is the minus sign in front of W at the right-hand side of the equation, which in Lindley's equation is a plus. Lindley's equation describes the waiting time of a customer in a single-server queue. It is one of the fundamental and best-studied equations in queuing theory. For a detailed study on Lindley's equation we refer to (Asmussen, 2003; Cohen, 1982) and the references therein. The implications of this "minor" difference in sign are rather far-reaching, since even for two stations, in the particular case we study in this paper, Lindley's equation has a simple solution, while for our equation it is probably not possible to derive an explicit expression without making some additional assumptions. In the applied probability literature, there has been considerable interest in the class of Markov chains described by the recursion $W_{n+1} = g(W_n, X_n)$. An important result is the duality theory by (Asmussen and Sigman, 1996), relating the steady-state distribution to a ruin probability associated with a risk process. See also (Borovkov, 1998) and (Kalashnikov, 2002). However, duality does not hold in our case, as our function is non-increasing in its main argument. This fact produces some surprising results when analysing the equation.

We study the waiting time of the server for this model. The waiting time satisfies the Lindley-type recursion (2), which surprisingly emerges when studying maximum weight independent sets in sparse random graphs. Specifically, consider an n-node sparse random (potentially regular) graph and let the nodes of the graph be equipped with nonnegative weights, independently generated according to some common distribution. Rather than only the size of the maximum independent set, consider also the maximum *weight* of an independent set. (Gamarnik et al., 2006) show that for certain weight distributions, a limiting result can be proven both for the maximum independent set and the maximum weight independent set. What is crucial in this computation is recursion (2); cf. (Gamarnik et al., 2006, Eq. (3)). This recursion provides another surprising link between queuing theory and random graphs.

At a glance, other than the analytical results, the major insights we gain for this system are summarised as follows. First, we observe that variability in preparation times has a greater influence on the system than that of service times. In the healthcare setting, one could summarise it as follows: it pays more to have a reliable nurse than a reliable specialist. See Figure 1 for an illustration. Second, a *small* variability of preparation times actually improves the performance of the server, in the sense that he waits less frequently; cf. Figure 2. However, it also decreases the throughput. Thus, the system's designer may wish to consider how to balance these conflicting goals. Next, when

deciding how many stations to assign to a server, the shape of the distribution plays a role. However, in general, when preparation times are smaller than service times and when the preparation times variability is low, only few stations per server (about 5 or 6) already come close to the optimal throughput. The last insight we gain is of mathematical nature. We observe that as the number of stations goes to infinity, the waiting times of the server become uncorrelated. We additionally provide an analytic lower bound on the throughput for the general case and an empirical upper bound. Both bounds are easy to compute, converge exponentially to the true throughput as N goes to infinity, and are tight in some cases. Thus we get quick and accurate estimates on the system's performance. For a discussion, see Section 4.

The rest of the paper is organised as follows: the general model is presented in the next section where we also give a sufficient condition for the existence of a limiting waiting-time distribution. Under the assumption that preparation times are exponential, we study the transient behaviour of the waiting time in Section 3 and provide the transition matrix of the underlying Markov chain. We conclude in Section 4 with insights to the effect of all parameters to the system's performance and give the main conclusions.

2 GENERAL MODEL

We assume that there are $N \geq 2$ identical stations operated by a single server. The system operates as follows. Before being served by the server, a customer must first undergo a preparation phase (not involving the server). Thus the server, after having finished serving a customer at one station, may have to wait for the preparation phase of the customer at the next station to be completed. Immediately after the server concludes his service at some station, another customer from the queue begins his preparation phase there while the server moves to the next station. We are interested in the waiting time of the server. Let B_n denote the preparation time for the n-th customer and let A_n be the time the server spends on this customer. Then the waiting times W_n of the server satisfy

$$W_{n+1} = \big(B_{n+1} - \sum_{i=n-N+2}^{n} A_i - \sum_{i=n-N+2}^{n} W_i\big)^+. \quad (1)$$

We assume that each sequence $\{A_n\}_{n\geq 1}$ and $\{B_n\}_{n\geq 1}$ is comprised of independent, identically distributed (i.i.d) nonnegative random variables with finite means, and the sequences are mutually independent. Moreover, for all $n \geq 1$, A_n (B_n) have a general distribution function F_A (F_B), density function f_A (f_B) and Laplace-Stieltjes Transform (LST) $\alpha(s) = \mathbb{E}\left[e^{-sA}\right]$ ($\beta(s) = \mathbb{E}\left[e^{-sB}\right]$). Eq. (1) can be written as

$$W_{n+1} = \big(X_{n+1} - \sum_{i=n-N+2}^{n} W_i\big)^+, \quad (2)$$

where $X_{n+1} = B_{n+1} - \sum_{i=n-N+2}^{n} A_i$. Note that $\{X_n\}$ are identically distributed for $n \geq N$ according to a random variable X, but are *not* independent. They are only independent with a $N-1$-lag. For example, $\{X_N,\ X_{2N-1},\ X_{3N-2},\ X_{4N-3}, \ldots\}$ are independent. We also define R_n^j to be the residual preparation time in station $(n+j)$ mod N at the moment the server is available for the n-th time, $n \geq 1$, $j = 1, \ldots, N-2$. Clearly, $R_n^{N-1} = B_{n+N-1}$ and $R_n^N = W_n$. Note that we distinguish between stations and visits. Since the server attends the stations in a cyclic order, starting the n-th visit is equivalent to visiting station $j = n$ mod N for the $\lceil \frac{n}{N} \rceil$-th time. The process $(W_n, R_n^1, R_n^2, \ldots, R_n^{N-2})$ is a Markov chain, of which the evolution is given by

$$W_{n+1} = \big(R_n^1 - W_n - A_n\big)^+,$$
$$R_{n+1}^j = \big(R_n^{j+1} - W_n - A_n\big)^+ \quad \text{for } j = 1, 2, \ldots, N-2.$$

Last, we assume that $\mathbb{P}(X \leq 0) > 0$, omitting the subscript when we consider a generic random variable.

2.1 Existence of a Limiting Waiting Time Distribution

Recall that $X_{n+1} = B_{n+1} - \sum_{i=n-N+2}^{n} A_i$, so that $X_N, X_{N+1}, \ldots$ are identically distributed. Note that the stochastic process $\{W_n\}$ is a (possibly delayed) regenerative process with regeneration times $\{n : W_n = W_{n+1} = \ldots = W_{n+N-2} = 0\}$. Moreover, note that this process is aperiodic. Let j be any regeneration time after $t = N-1$. Furthermore, let $\tau = \inf\{n : n > 0, W_j = W_{j+1} = \ldots = W_{j+N-2} = W_{j+n} = W_{j+n+1} = \ldots = W_{j+n+N-2} = 0\}$, so that τ can be interpreted as the time between two regeneration moments. In the Appendix, we show that the mean cycle length $\mathbb{E}[\tau]$ is finite, which implies by standard theory on regenerative processes that the limiting distribution of the waiting time exists and the waiting-time process converges to it (see e.g. (Asmussen, 2003, Cor. VI.1.5 and Thm. VII.3.6)).

3 TRANSIENT ANALYSIS

For the sequel, we assume that preparation times are exponentially distributed with rate μ. Note that the analysis can extend to phase-type preparation times,

but at the cost of more cumbersome expressions. Furthermore, little insight is added by such an extension.

We first show that the waiting time (has an atom at zero and), provided that it is positive, is also exponentially distributed with rate μ. We then compute the atom at zero by computing the transition matrix of the underlying Markov chain. We show that the matrix has a nice structure that can be exploited for numerical computations. Particularly for three stations, we provide further analytic results. We compute the steady-state distribution, and give closed-form expressions for the covariance between two waiting times and for the mean time between two zero waiting times, both for the transient and the steady-state cases.

3.1 The Behaviour of W_{n+1}

We show that $\mathbb{P}(W_{n+1} > x \mid W_{n+1} > 0) = e^{-\mu x}$, for all $n \geq 0$. We prove this claim for $n \geq N-1$ (for $1 \leq n \leq N-2$ it is done in a similar way). In order to calculate $\mathbb{P}(W_{n+1} > x)$, we first calculate it conditioned on the last $N-1$ waiting times. We get, for all $n \geq N-1$,

$$\begin{aligned}
&\mathbb{P}(W_{n+1} > x | W_n = w_n, \ldots, W_{n-N+2} = w_{n-N+2}) \\
&= \mathbb{P}(B_{n+1} > \sum_{i=n-N+2}^{n} A_i + \sum_{i=n-N+2}^{n} w_i + x) \\
&= \int_0^\infty \cdots \int_0^\infty e^{-\mu(\sum_{i=n-N+2}^{n}(y_i+w_i)+x)} \\
&\qquad\qquad dF_{A_{n-N+2}}(y_{n-N+2}) \ldots dF_{A_n}(y_n) \\
&= (\alpha(\mu))^{N-1} \exp\{-\mu(\sum_{i=n-N+2}^{N} w_i + x)\}, \qquad (3)
\end{aligned}$$

where we defined $\alpha(\mu) = \mathbb{E}\left[e^{-\mu A}\right]$. Thus, (3) implies that

$$\mathbb{P}(W_{n+1} > x \mid \sum_{i=n-N+2}^{n} W_i = 0) = (\alpha(\mu))^{N-1} \mathrm{e}^{-\mu x}.$$

From (3) we also get that

$$\begin{aligned}
&\mathbb{P}(W_{n+1} > x | W_{n+1} > 0, W_n = w_n, \ldots, \\
&\qquad\qquad W_{n-N+2} = w_{n-N+2}) \\
&= \frac{\mathbb{P}(W_{n+1} > x \mid W_n = w_n, \ldots, W_{n-N+2} = w_{n-N+2})}{\mathbb{P}(W_{n+1} > 0 \mid W_n = w_n, \ldots, W_{n-N+2} = w_{n-N+2})} \\
&= \frac{(\alpha(\mu))^{N-1} \exp\{-\mu(\sum_{i=n-N+2}^{n} w_i + x)\}}{(\alpha(\mu))^{N-1} \exp\{-\mu(\sum_{i=n-N+2}^{n} w_i)\}} = \mathrm{e}^{-\mu x},
\end{aligned}$$

meaning that, given $W_{n+1} > 0$, W_{n+1} is not affected by the previous $N-1$ waiting times $W_n, W_{n-1}, \ldots, W_{n-N+2}$. A direct conclusion is that $\mathbb{P}(W_{n+1} > x | W_{n+1} > 0) = e^{-\mu x}$ and thus, for $x > 0$,

$$\begin{aligned}
\mathbb{P}(W_{n+1} > x) &= \mathbb{P}(W_{n+1} > x \mid W_{n+1} > 0)\mathbb{P}(W_{n+1} > 0) \\
&\quad + \mathbb{P}(W_{n+1} > x \mid W_{n+1} = 0)\mathbb{P}(W_{n+1} = 0) \\
&= e^{-\mu x}\mathbb{P}(W_{n+1} > 0). \qquad (4)
\end{aligned}$$

That is, the distribution of W_n is a mixture of a mass at zero and the exponential distribution with rate μ. The same argument can be applied in a similar manner for W, the limit of W_n as $n \to \infty$. That is, $\mathbb{P}(W > x) = e^{-\mu x}\mathbb{P}(W > 0)$. We now calculate $\mathbb{P}(W_{n+1} > 0)$ for all n, and $\mathbb{P}(W > 0)$. In order to do that, we will define a Markov chain and calculate its one-step transition probability matrix. A detailed analysis is presented in the next section.

3.1.1 Construction of a Markov Chain

Recall that the process $(W_n, R_n^1, R_n^2, \ldots, R_n^{N-2})$ is a Markov chain and define the auxiliary 0-1 random variables $F_n = I(W_n > 0)$ and $G_n^j = I(R_n^j > 0)$, $j = 1, \ldots, N-2$. That is, $F_n = 0$ if, at the moment the server starts his n-th visit to a station, the customer there completed his preparation process, so that the server does not have to wait. Otherwise, $F_n = 1$. In the same way, $G_n^j = 0$ if, at the moment the server starts his n-th visit to a station, the customer at station $(n+j) \bmod N$ has completed the preparation process, and $G_n^j = 1$ otherwise.

Due to the memoryless property of the preparation times, the process $(F_n, G_n^1, \ldots, G_n^{N-2})$ is a Markov chain on $\{0,1\}^{N-1}$. Thus, the state space consists of 2^{N-1} states, where each state describes the residual preparation time in each station (positive or zero) at the moment the server enters a station for his overall n-th visit. The only station that does not appear in this description is the station the server has just left, since the residual preparation time there is always B (or, in other words, $G_n^{N-1} = 1$ for all n). Let P be the one-step transition probability matrix of the Markov chain $(F_n, G_n^1, \ldots, G_n^{N-2})$, where the states are lexicographically ordered, so that the first state is $(0,0,\ldots,0,0)$, the second one is $(0,0,\ldots,0,1)$, and so on, where the last state is $(1,1,\ldots,1,1)$. For a state $i \in \{0,1\}^{N-1}$ we denote its coordinates by $(i_0, \ldots, i_{N-2})$. Below, we describe how to construct the matrix P.

Theorem 1. *The transition matrix P is described as follows. For a state $i \in \{0,1\}^{N-1}$, define $T(i) = \{r : p_{i,r} > 0\}$ and $k = \sum_{r=0}^{N-2} i_r$. For $j \in T(i)$ also define $m = \sum_{r=0}^{N-2} j_r$ and $d = k - m$. If $F_n = 0$, then*

$$P_{i,j} = \begin{cases} \alpha(m\mu) & \text{if } d = -1, \\ \sum_{l=0}^{d+1} \binom{d+1}{l} (-1)^l \alpha((m+l)\mu) & \text{if } d > -1. \end{cases}$$

For $F_n = 1$, we have

$$P_{i,j} = \begin{cases} \frac{\alpha(m\mu)}{m+1} & \text{if } d = 0, \\ \sum_{l=0}^{d} \binom{d}{l} (-1)^l \frac{\alpha((m+l)\mu)}{m+l+1} & \text{if } d > 0. \end{cases}$$

For all other values of d, $P_{i,j} = 0$.

Proof. For each state $i \in \{0,1\}^{N-1}$, we derive the possible target states, namely $T(i)$, as follows: move all 0's (except the first one if $F_n = 0$) one position to the left and take all possible combinations for the other positions. This sums up to either 2^k or 2^{k+1} possible states, depending on the value of F_n. For example, for $N = 4$, assume the chain is in state $(0,1,0)$, then $T[(0,1,0)] = \{(0,0,0),(0,0,1),(1,0,0),(1,0,1)\}$.

Assume $F_n = 0$, meaning that at the current state, the server starts serving *immediately* upon entering the station.

- If $d = -1$, then all stations that had a positive residual preparation time in state i, will still have a positive residual preparation time in state j. In addition, the preparation time in the station the server has just left did not end as well. As $d = -1$ implies $m = k+1$, this transition occurs when during the service time A, $k+1 = m$ independent exponential preparation times with rate μ were not completed. Therefore,

$$P_{i,j} = \int_0^\infty \mathrm{e}^{-m\mu y} dF_A(y) = \alpha(m\mu).$$

- When $d > -1$, exactly $k+1-m = d+1$ preparation times have ended during a service time A, and the other m did not. Thus,

$$P_{i,j} = \int_{y=0}^\infty (1-\mathrm{e}^{-\mu y})^{d+1} \mathrm{e}^{-m\mu y} dF_A(y) = \sum_{l=0}^{d+1} \binom{d+1}{l} (-1)^l \alpha((m+l)\mu).$$

We now consider the case where $F_n = 1$, implying that the preparation in the station the server had just entered did not finish, and thus an exponential preparation time B remains. Note that when $F_n = 1$, $d = k-m$ is nonnegative.

- The case $d = 0$ means that (i) all stations with a positive residual preparation time in state i, will still have a positive residual preparation time in state j, and (ii) the preparation time in the station that the server has just left also did not end. Namely, this transition occurs when during the preparation time B and service time A, $k = m$ independent exponential preparation times with rate μ were not completed. Therefore,

$$P_{i,j} = \iint_{\mathbb{R}_+^2} \mathrm{e}^{-m\mu(x+y)} \mu \mathrm{e}^{-\mu x} dx dF_A(y) = \frac{\alpha(m\mu)}{m+1}.$$

- When $d > 0$, exactly $k-m = d$ out of k preparations have been completed during the time $A+B$, implying that

$$P_{i,j} = \int_{y=0}^\infty \int_{x=0}^\infty (1-\mathrm{e}^{-\mu(x+y)})^{k-m} \mathrm{e}^{-m\mu(x+y)} \mu \mathrm{e}^{-\mu x} dx dF_A(y) = \sum_{l=0}^{d} \binom{d}{l} (-1)^l \frac{\alpha((m+l)\mu)}{m+l+1}.$$

□

Once the matrix P is computed, we can find the distribution of W_n for all $n \geq 1$, which is given in (4). Therefore, we need $\mathbb{P}(W_n > 0)$.

Let π_n denote the distribution vector on all the 2^{N-1} possible states at the moment the server starts his n-th visit. Recall that the states are lexicographically ordered. We may assume that the Markov chain starts at time 1 from an initial distribution vector $\pi_1 = (0,0,\ldots,0,0,1)$, meaning that at the moment the server enters the first station, in all stations the preparation has not been completed yet, and therefore the Markov chain at that moment is in state $(1,1,\ldots,1,1)$. Now, $\mathbb{P}(W_n > 0)$ is the sum of the last 2^{N-2} elements of the vector $\pi_1 P^{n-1}$. Furthermore, we can also find the distribution of W (which is the limiting distribution of W_n), by solving the equation $\pi P = \pi$, and by summing the last 2^{N-2} elements of π (all the states that start with 1), we get $\mathbb{P}(W > 0)$.

Next we present a detailed analysis for the case with $N = 3$ stations. We again consider exponentially distributed preparation times, although the analysis can evidently extend to phase-type distributions.

3.2 Analysis for $N = 3$ Stations

For $N = 3$, the evolution of the Markov process (W_n, R_n) is given by

$$W_{n+1} = (R_n - W_n - A_n)^+, \; R_{n+1} = (B_{n+2} - W_n - A_n)^+,$$

where W_n is the waiting time of the server at his n-th entrance to a station, and R_n is the residual preparation time in the following station. Clearly, the residual preparation time in the station that was just left by the server is B. We assume that the preparation time in each station is exponentially distributed with parameter μ. We calculate the limiting distribution (W,R) of the Markov chain and the (transient) distribution of W_n. We also derive the covariance $\mathrm{Cov}[W_n, W_{n+k}]$ and the distribution function of the number of visits between two successive zero waiting times of the server. Observe that these are not necessarily two regenerative points. We first need to derive the transition probabilities of the Markov chain.

As described in Section 3.1.1, there are only four (2^{3-1}) relevant states in the corresponding Markov chain. Define $\pi = (\pi_{(0,0)}, \pi_{(0,1)}, \pi_{(1,0)}, \pi_{(1,1)})$ to be

the stationary probabilities vector of the states $\{(0,0), (0,1), (1,0), (1,1)\}$. Let $P_{(i,j),(k,l)}$ be the one-step transition probabilities, i.e.

$$P_{(i,j),(k,l)} = \mathbb{P}(F_{n+1} = k, G_{n+1} = l | F_n = i, G_n = j),$$

for $i, j, k, l = 0, 1$. By Theorem 1 we get that P is given by

$$P = \begin{pmatrix} 1-\alpha(\mu) & \alpha(\mu) & 0 & 0 \\ 1-2\alpha(\mu)+\alpha(2\mu) & \alpha(\mu)-\alpha(2\mu) & \alpha(\mu)-\alpha(2\mu) & \alpha(2\mu) \\ 1-\frac{1}{2}\alpha(\mu) & \frac{1}{2}\alpha(\mu) & 0 & 0 \\ 1-\alpha(\mu)+\frac{1}{3}\alpha(2\mu) & \frac{1}{2}\alpha(\mu)-\frac{1}{3}\alpha(2\mu) & \frac{1}{2}\alpha(\mu)-\frac{1}{3}\alpha(2\mu) & \frac{1}{3}\alpha(2\mu) \end{pmatrix}$$

Thus, we obtain the limiting distribution

$$\pi_{(0,0)} = \frac{12-6\alpha^2(\mu)-12\alpha(\mu)+4\alpha(\mu)\alpha(2\mu)+\alpha^2(\mu)\alpha(2\mu)+8\alpha(2\mu)}{12+6\alpha^2(\mu)+\alpha^2(\mu)\alpha(2\mu)+8\alpha(2\mu)},$$

$$\pi_{(0,1)} = \frac{4\alpha(\mu)(3-\alpha(2\mu))}{12+6\alpha^2(\mu)+\alpha^2(\mu)\alpha(2\mu)+8\alpha(2\mu)},$$

$$\pi_{(1,0)} = \frac{2\alpha(\mu)(\alpha(\mu)\alpha(2\mu)+6\alpha(\mu)-6\alpha(2\mu))}{12+6\alpha^2(\mu)+\alpha^2(\mu)\alpha(2\mu)+8\alpha(2\mu)},$$

$$\pi_{(1,1)} = \frac{12\alpha(\mu)\alpha(2\mu)}{12+6\alpha^2(\mu)+\alpha^2(\mu)\alpha(2\mu)+8\alpha(2\mu)}.$$

To obtain the transient distribution and the covariance function, let $\pi_1 = (0,0,0,1)$ be the initial vector of the Markov chain (F_n, G_n). Then, for all $x \geq 0$

$$\mathbb{P}(W_n > x) = e^{-\mu x}\left(P^{(n-1)}[4,3] + P^{(n-1)}[4,4]\right),$$

with $P[i,j]$ the element in row i and column j. Now,

$$\mathrm{Cov}[W_n, W_{n+k}] = \mathbb{E}[W_n W_{n+k}] - \mathbb{E}[W_n]\mathbb{E}[W_{n+k}].$$

Since $W_n|_{W_n>0} \sim exp(\mu)$, we have for all $k \geq 0$,

$$\begin{aligned}\mathbb{E}[W_{n+k}] &= \mathbb{E}[W_{n+k}|W_{n+k} > 0]\mathbb{P}(W_{n+k} > 0) \\ &= \frac{1}{\mu}\mathbb{P}(W_{n+k} > 0).\end{aligned}$$

To calculate $\mathbb{E}[W_n W_{n+k}]$, we note that

$$\begin{aligned}\mathbb{E}[W_n W_{n+k}] = \mathbb{E}[W_n W_{n+k}|W_n > 0, W_{n+k} > 0] \\ \mathbb{P}(W_n > 0, W_{n+k} > 0),\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}[W_n W_{n+k}|W_n > 0, W_{n+k} > 0] \\ &= \int_{w=0}^{\infty} w\mathbb{E}[W_{n+k}|W_{n+k} > 0, W_n > w]\mu e^{-\mu w} dw \\ &= \int_{w=0}^{\infty} w\frac{1}{\mu}\mu e^{-\mu w} dw = \frac{1}{\mu^2},\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(W_{n+k} > 0, W_n > 0) \\ &= \mathbb{P}(W_{n+k} > 0|W_n > 0)\mathbb{P}(W_n > 0) \\ &= \mathbb{P}(W_n > 0)(\mathbb{P}(W_{n+k} > 0|W_n > 0, R_n = 0) \\ &\qquad \mathbb{P}(R_n = 0|W_n > 0) \\ &\quad + \mathbb{P}(W_{n+k} > 0|W_n > 0, R_n > 0)\mathbb{P}(R_n > 0|W_n > 0)) \\ &= \left(P^{(k)}[3,3] + P^{(k)}[3,4]\right)P^{(n-1)}[4,3] \\ &\quad + \left(P^{(k)}[4,3] + P^{(k)}[4,4]\right)P^{(n-1)}[4,4].\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}[W_n W_{n+k}] = \\ &\frac{1}{\mu^2}\left(\left(P^{(k)}[3,3] + P^{(k)}[3,4]\right)P^{(n-1)}[4,3]\right. \\ &\qquad \left. + \left(P^{(k)}[4,3] + P^{(k)}[4,4]\right)P^{(n-1)}[4,4]\right). \quad (5)\end{aligned}$$

Finally,

$$\begin{aligned}\mathrm{Cov}[W_n, W_{n+k}] = \mathbb{E}[W_n W_{n+k}] \\ - \frac{1}{\mu^2}\mathbb{P}(W_n > 0)\mathbb{P}(W_{n+k} > 0),\end{aligned}$$

with $\mathbb{E}[W_n W_{n+k}]$ given in (5).

Last, we compute the distribution and expectation of visits between two consecutive zero waiting times. Suppose that $W_n = 0$ and define for all $n \geq 1$ the random variable $C_n := C_{|W_n=0}$, describing the length from the moment that $W_n = 0$ until the next time that the server's waiting time is zero. In other words,

$$C_n = \inf\{k : W_{n+k} = 0 \mid W_n = 0\}.$$

The results are summarised in the following theorem.

Theorem 2. *The distribution of C_n is given by*

$$\mathbb{P}(C_n = 1) = 1 - \frac{P^{(n-1)}[4,2]}{P^{(n-1)}[4,1] + P^{(n-1)}[4,2]}\alpha(\mu), \quad (6)$$

$$\begin{aligned}\mathbb{P}(C_n = 2) = \frac{P^{(n-1)}[4,2]}{P^{(n-1)}[4,1] + P^{(n-1)}[4,2]}\alpha(\mu) \\ \left(1 - \frac{1}{2}\alpha(2\mu)\right),\end{aligned}$$

$$\begin{aligned}\mathbb{P}(C_n = k) = \frac{P^{(n-1)}[4,2]}{P^{(n-1)}[4,1] + P^{(n-1)}[4,2]}\alpha(\mu) \\ \left(\frac{1}{2}\alpha(2\mu)\right)\left(\frac{1}{3}\alpha(2\mu)\right)^{k-3}\left(1 - \frac{1}{3}\alpha(2\mu)\right), \ k \geq 3.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}[C_n] = 1 + \frac{P^{(n-1)}[4,2]}{P^{(n-1)}[4,1] + P^{(n-1)}[4,2]} \\ \alpha(\mu)\left(\frac{6+\alpha(2\mu)}{6-2\alpha(2\mu)}\right). \quad (7)\end{aligned}$$

Proof. For $k = 1$, we have that

$$\begin{aligned}
&\mathbb{P}(C_n = 1) = \mathbb{P}(W_{n+1} = 0 | W_n = 0) \cdot \\
&= \mathbb{P}(W_{n+1} = 0 | W_n = 0, R_n = 0)\mathbb{P}(R_n = 0 | W_n = 0) \\
&\quad + \mathbb{P}(W_{n+1} = 0 | W_n = 0, R_n > 0)\mathbb{P}(R_n > 0 | W_n = 0) \\
&= (P[1,1] + P[1,2]) \cdot \frac{P^{(n-1)}[4,1]}{\mathbb{P}(W_n = 0)} \\
&\quad + (P[2,1] + P[2,2]) \cdot \frac{P^{(n-1)}[4,2]}{\mathbb{P}(W_n = 0)} \\
&= \frac{P^{(n-1)}[4,1]}{\mathbb{P}(W_n = 0)} + (1 - \alpha(\mu)) \cdot \frac{P^{(n-1)}[4,2]}{\mathbb{P}(W_n = 0)} \\
&= 1 - \alpha(\mu) \frac{P^{(n-1)}[4,2]}{P^{(n-1)}[4,1] + P^{(n-1)}[4,2]}.
\end{aligned}$$

The results for

$$\begin{aligned}
&\mathbb{P}(C_n = i) \\
&= \mathbb{P}(W_{n+i} = 0, W_{n+i-1} > 0, \ldots, W_{n+1} > 0 | W_n = 0)
\end{aligned}$$

with $i > 1$, follow by expanding this expression into $\mathbb{P}(W_{n+1} > 0 | W_n = 0)$ as well as probabilities of the form $\mathbb{P}(W_{j+2} > 0 | W_{j+1} > 0, W_j = 0)$ and $\mathbb{P}(W_{j+2} > 0 | W_{j+1} > 0, W_j > 0)$. Similar to the derivations above, these probabilities can be computed by using the Markov chain formulation of the previous section. The proof of (7) is straightforward. □

4 INSIGHTS

In the previous sections, we gave closed-form expressions for exponentially distributed preparation times. Here, we obtain general insights into the behaviour of the model by simulation on a larger range of parameter settings. We vary, among other, the number of stations and the distributions of the preparation and service times. We focus on the effect of the first two moments of preparation and service times to the throughput. For their distributions, we choose phase-type distributions based on two-moment-fit approximations commonly used in literature, see e.g. (Tijms, 1994, p. 358–360). We discuss several interesting conclusions based on the simulation results.

Variability of Preparation and Service Times. When controlling the system, the variability of the preparation times seems to play a larger role than the variability of the service time. This is because the server's waiting-time process is much more sensitive to the former than to the latter. See e.g. Figure 1, where the throughput θ is plotted versus the number of queues N. We observe the throughput

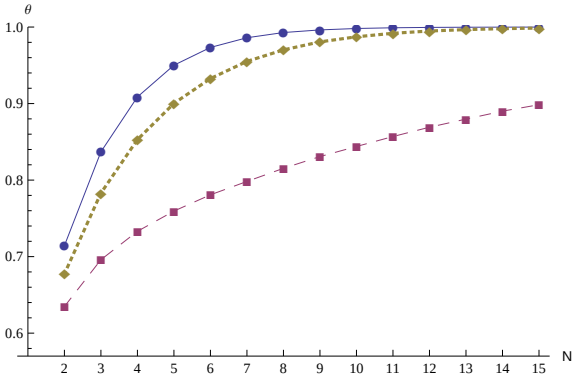

Figure 1: Throughput vs. the number of stations for standard-exponential preparation and service times (solid), highly variable service times (dotted) and highly variable preparation times (dashed).

for various variability settings for both time components. We fix the means at $\mathbb{E}[A] = \mathbb{E}[B] = 1$, and consider the exponential distribution, i.e. $\mathbb{E}[A^2] = \mathbb{E}[B^2] = 2$ (solid curve), and two different phase-type distributions with highly variable service times only, i.e. $\mathbb{E}[A^2] = 10, \mathbb{E}[B^2] = 2$ (dotted curve) and highly variable preparation times only, i.e. $\mathbb{E}[A^2] = 2, \mathbb{E}[B^2] = 10$ (dashed curve). Although the variability of preparation times and service times are varied in similar ways, the dotted curve nears the solid curve as N grows larger much faster than the dashed curve. Therefore, predictability of the preparation times seems to be much more important than that of the service times. This is to be expected; when we observe (1) we see that as the number of stations tends to infinity, the squared coefficient of variation of the sum of service times at the right-hand side of (1) goes to zero, and thus the effect of the sum of the service times is minimal.

In other words, it is more important that one has a reliable assistant than a reliable server, in particular for large systems. In the carousel setting, this is more or less guaranteed; although the preparation times (i.e. rotation times) depend on the picking strategy followed, they are bounded by the length of the carousel and as such exhibit small variability. Whether the picker is robotic (small variability) or human, does influence the system, but not as dramatically as the preparation times do. The influence of variable service times decreases very fast as the number of stations increases, even for highly variable service times. The influence of variable preparation times decreases too but so slowly that it converges to the benchmark case only at infinity. It is natural to expect that as N tends to infinity preparation times become less important. One expects that the preparation will almost surely have expired after serving a very

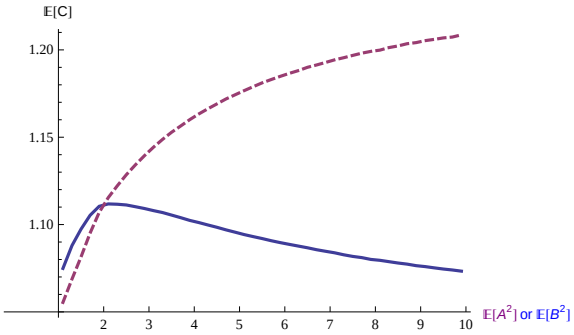

Figure 2: Mean time between two zero waiting times vs. $\mathbb{E}[B^2]$ (solid) and $\mathbb{E}[A^2]$ (dashed).

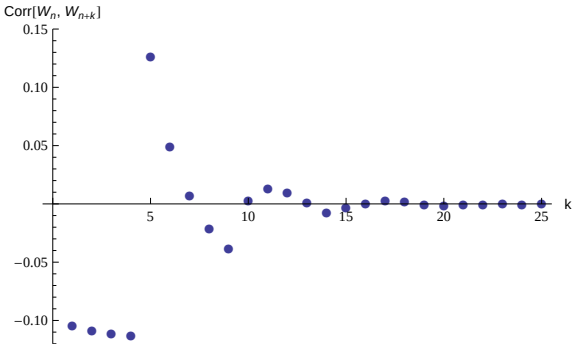

Figure 3: Correlations exhibit periodicity.

large number of stations and that the total throughput will simply equal the rate of service.

This statement is reinforced by Figure 2, where the mean time between two zero waiting times is plotted versus the second moment of the preparation time B (solid curve) or that of the service time A (dashed curve). It is assumed that $N = 4$ and $\mathbb{E}[A] = \mathbb{E}[B] = 1$ throughout for both of these lines. For the first curve, the service times A are taken to be exponentially distributed, while for the second, the preparation times B are taken to be exponentially distributed. From Figure 2, it is apparent that the mean time between two zero waiting times increases (i.e. the frequency of zero waiting times decreases) as the service times becomes more variable. However, mostly the opposite is observed for the preparation times. Although the expected waiting time increases in the variability of the preparation times by Figure 1, apparently the mean time between two zero waiting times now *decreases* anomalously. From this, we conclude that the server's waiting time process behaves more and more erratically as the variability of the preparation times increases and seems to be more resistent against highly variable service times. Again, this effect may be explained by the nature of the waiting time (see (1)), which is expressed in terms of one preparation time, but a *sum* of service times. The squared coefficient of variation of the sum goes to zero.

In summary, we can say that variability of preparation times, as long as it is small, improves the performance of the server, in the sense that he waits less frequently, while variability of service times always improves the performance of the server in the same sense. However, both scenarios decrease the throughput of the system – although waiting times occur less frequently under some variability, when they occur they tend to be longer, thus decreasing the total throughput. Simulation results show about a 10% decrease in throughput under common scenarios when ranging the preparation time variability (i.e. the worst case) from a deterministic to an exponential. Nonetheless, in some service systems this may be an advantage, as it gives the opportunity to perform an additional task (e.g. administration).

Correlations. In general, this system has an interesting correlation structure. In Figure 3 we plot the correlation between two waiting times of lag k against the lag for exponential preparation and service times with rates 1 and 10 respectively. As we see in Figure 3, correlations exhibit a periodic structure, which is natural as it corresponds to a return to the first station. Moreover, as time goes to infinity, the waiting times become uncorrelated, which is again a natural conclusion, as the process is ergodic. As shown in Section 2.1, there exists a unique limiting distribution and the system converges to it, thus as time (or the number of stations for that matter) goes to infinity, the system converges to steady-state regardless of the initial state. Thus, the correlation between waiting times goes to zero because the system loses its memory due to ergodicity. Although the convergence to zero correlations is expected, the way this happens is intriguing. One may expect some form of periodicity, but it is not clear why the first cycle looks different than the rest or why correlations should be forming alternatingly convex and concave loops after the first cycle.

Number of Stations to be Assigned to a Server. One of the important management decisions to be made is the number of stations to be assigned to a server. Think of the warehouse example given earlier. The more carousels assigned to the picker, the better his utilisation. However, the utilisation of each carousel decreases. We wish to understand this interplay. An important measure to be taken into account is the throughput of the system. Note that the throughput is linearly related to the fraction of time the server is operating, since service is completed at rate $\mathbb{E}[B]^{-1}$ whenever the server is not forced to wait.

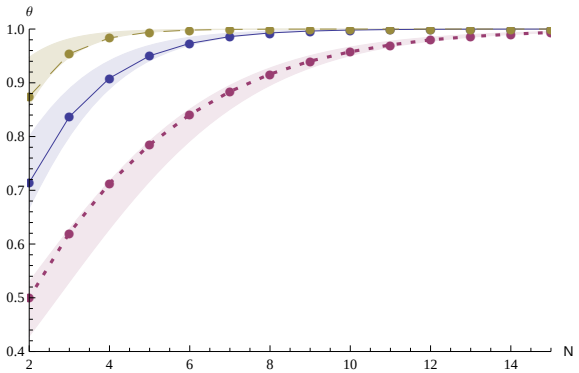

Figure 4: Throughput vs. the number of stations for small (dashed), moderate (solid), and large preparation times (dotted).

The number of stations to be assigned to a server in order to reach near-optimal throughput depends very much on the distributions of the preparation time B and the service time A. This effect is observed in Figure 1, where we see that for highly variable preparation times (dashed line), the throughput increases for every additional station assigned to the server. Evidently, it will converge to the rate of service, but this convergence is *very* slow. On the other hand, variability in the service times influences the system, but the convergence follows more or less the pattern of the exponential case.

When all distributions are exponential, it is evident that the only quantity that matters in the determination of the throughput is $r = \mathbb{E}[A]/\mathbb{E}[B]$. In order to determine the optimal number of stations to assign to a server, we plot in Figure 4 the throughput θ versus the number of stations N for three cases of r, namely for $r = 2$ (top curve), $r = 1$ (solid curve), and $r = 0.5$ (dotted curve). In all three cases, the underlying distributions are exponential. What we observe is that when $r \geq 1$, i.e. the top two curves, the throughput converges fast, and little benefit is added by assigning one more station to the server. This is to be expected, as in this case the mean service time is not smaller than the mean preparation time, and so the server rarely has to wait. In other words, he works at almost full capacity, and thus convergence to the maximum service rate (equal to 1 in all scenarios), is fast. However, when $r < 1$, the convergence is very slow. We conclude that the shape of the distribution plays a role, but in general for $r > 1$ and low variability in preparation times, only few stations per server (say about 5 or 6) are needed to already come close to the maximum throughput.

A Rough Estimate. In Figure 4 we also plot a rough first-order upper bound and an analytic lower bound of the throughput that we derive as follows. The throughput θ of the system is equal to the number of customers N served per cycle over the cycle length, which is $N(\mathbb{E}[W] + \mathbb{E}[A])$. Thus,

$$\theta = (\mathbb{E}[W] + \mathbb{E}[A])^{-1}.$$

A first-order approximation of θ can be produced by estimating $\mathbb{E}[W]$ in the denominator with the mean residual preparation time multiplied with a rough estimate that the server has to wait, i.e. $\mathbb{P}(B > A_1 + \cdots + A_{N-1})$. Then, for exponential preparation times B,

$$\widetilde{\theta}_N = \frac{1}{\frac{\alpha^{N-1}(\mu)}{\mu} + \mathbb{E}[A]}.$$

We observe that this expression is a lower bound of the throughput, since the actual probability a server has to wait is $\mathbb{P}(B > A_1 + W_1 + \cdots + A_{N-1} + W_{N-1})$, and thus smaller. We also observe empirically, that $\widetilde{\theta}_{N+1}$ provides an upper bound of the throughput in the scenarios we examined. Observe that the analytic lower bound becomes tighter as r increases, while the empirical upper bound provides a better estimate for small values of r. As a result, the system's designer can have a quick, easy, and accurate bound on the throughput for all parameter settings.

Our final observation is that since $\widetilde{\theta}_N$ is a lower bound, and Figure 4 suggests that $\widetilde{\theta}_{N+1}$ is an upper bound, we also observe that both θ and $\widetilde{\theta}_N$ converge as N goes to infinity exponentially fast (analytically) to the maximum service rate with the correct (light-tail asymptotics) rate $\log \alpha(\mu)$. As such, these bounds are provably useful.

REFERENCES

Asmussen, S. (2003). *Applied Probability and Queues*. Springer Verlag.

Asmussen, S. and Sigman, K. (1996). Monotone stochastic recursions and their duals. *Probability in the Engineering and Informational Sciences*, 10(1):1–20.

Boon, M. A. A., Van der Mei, R. D., and Winands, E. M. M. (2011). Applications of polling systems. *Surveys in Operations Research and Management Science*, 16(2):67–82.

Borovkov, A. A. (1998). *Ergodicity and Stability of Stochastic Processes*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester.

Boxma, O. J. and Groenendijk, W. P. (1988). Two queues with alternating service and switching times. In Boxma, O. J. and Syski, R., editors, *Queueing Theory and its Applications (Liber Amicorum for J. W. Cohen)*, pages 261–282. Amsterdam: North-Holland.

Cohen, J. W. (1982). *The Single Server Queue*. North-Holland Publishing Co., Amsterdam.

Eisenberg, M. (1979). Two queues with alternating service. *SIAM Journal on Applied Mathematics*, 36:287–303.

Franks, G., Al-Omari, T., Woodside, M., Das, O., and Derisavi, S. (2009). Enhanced modeling and solution of layered queuing networks. *IEEE Transactions on Software Engineering*, 35:148–161.

Gamarnik, D., Nowicki, T., and Swirszcz, G. (2006). Maximum weight independent sets and matchings in sparse random graphs. Exact results using the local weak convergence method. *Random Structures & Algorithms*, 28(1):76–106.

Kalashnikov, V. (2002). Stability bounds for queueing models in terms of weighted metrics. In Suhov, Y., editor, *Analytic Methods in Applied Probability*, volume 207 of *American Mathematical Society Translations Ser. 2*, pages 77–90. American Mathematical Society, Providence, RI.

Lindley, D. V. (1952). The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 277–289.

Litvak, N. and Vlasiou, M. (2010). A survey on performance analysis of warehouse carousel systems. *Statistica Neerlandica*, 64(4):401–447.

McGinnis, L. F., Han, M. H., and White, J. A. (1986). Analysis of rotary rack operations. In White, J., editor, *Proceedings of the 7th International Conference on Automation in Warehousing*, pages 165–171, San Francisco, California. Springer.

Park, B. C., Park, J. Y., and Foley, R. D. (2003). Carousel system performance. *Journal of Applied Probability*, 40(3):602–612.

Resing, J. A. C. (1993). Polling systems and multitype branching processes. *Queueing Systems. Theory and Applications*, 13(4):409–426.

Takagi, H. (1986). *Analysis of polling systems*. MIT press.

Tijms, H. C. (1994). *Stochastic Models: an Algorithmic Approach*. Wiley, Chichester.

van der Mei, R. D. (2007). Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Systems. Theory and Applications*, 57(1):29–46.

van Vuuren, M. and Winands, E. M. M. (2007). Iterative approximation of k-limited polling systems. *Queueing Systems. Theory and Applications*, 55(3):161–178.

Vlasiou, M. (2006). *Lindley-type Recursions*. PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands.

Yechiali, U. (1993). Analysis and control of polling systems. In Donatiello, L. and Nelson, R., editors, *Performance Evaluation of Computer and Communication Systems, Joint Tutorial Papers of Performance '93 and Sigmetrics '93*, volume 729, pages 630–650, London, UK. Springer Berlin / Heidelberg.

APPENDIX

To prove that the mean cycle length $\mathbb{E}[\tau]$ is finite, observe that for any $n \geq N-1$, we have

$$\begin{aligned}\mathbb{P}(\tau > n) &= \mathbb{P}(\bigcap_{i=j+1}^{j+n} \Big\{ \sum_{k=0}^{N-2} W_{i+k} > 0 \Big\}) \\ &\leq \mathbb{P}(\bigcap_{i=j+N-1}^{j+n} \Big\{ \sum_{k=0}^{N-2} W_{i+k} > 0 \Big\}).\end{aligned}$$

Due to (2) and the fact that waiting times are nonnegative, X_n is stochastically larger than W_n. Thus,

$$\begin{aligned}\mathbb{P}(\tau > n) &\leq \mathbb{P}(\bigcap_{i=j+N-1}^{j+n} \Big\{ \sum_{k=0}^{N-2} X_{i+k} > 0 \Big\}) \\ &\leq \mathbb{P}(\bigcap_{i=1}^{\lfloor \frac{n}{N-1} \rfloor} \Big\{ \sum_{k=0}^{N-2} X_{j+i(N-1)+k} > 0 \Big\}) \\ &= \mathbb{P}(\sum_{k=0}^{N-2} X_{j+N-1+k} > 0)^{\lfloor \frac{n}{N-1} \rfloor}, \qquad (8)\end{aligned}$$

where the second equality follows from the fact that the process $\{X_n\}$ exhibits no auto-correlation for lag $N-1$ or more. Additionally, we have that

$$\begin{aligned}&\mathbb{P}(\sum_{k=0}^{N-2} X_{j+N-1+k} > 0) \leq 1 - \mathbb{P}(\bigcap_{k=0}^{N-2} \Big\{ X_{j+N-1+k} \leq 0 \Big\}) \\ &= 1 - \mathbb{P}(X_{j+N-1} \leq 0)\mathbb{P}(X_{j+N} \leq 0 | X_{j+N-1} \leq 0) \cdot \ldots \\ &\qquad \ldots \cdot \mathbb{P}(X_{j+2N-3} \leq 0 | \bigcap_{k=0}^{N-3} \Big\{ X_{j+N-1+k} \leq 0 \Big\}) \\ &\leq 1 - \mathbb{P}(X \leq 0)^{N-1} < 1. \qquad (9)\end{aligned}$$

The second to last inequality holds since the process $\{X_n\}$ exhibits positive auto-correlation with a lag up to $N-2$. This implies that $\mathrm{Cov}[\mathbb{1}_{\{X_n \leq 0\}}, \mathbb{1}_{\{X_{n-k} \leq 0\}}] > 0$ for any $n > 0$ and $0 < k \leq N-2$, so that $\mathbb{P}(X_n \leq 0 | X_{n-k} \leq 0) > \mathbb{P}(X \leq 0)$. The last inequality follows from the assumption that $\mathbb{P}(X \leq 0) > 0$. Finally, from (8) we have that

$$\begin{aligned}\mathbb{E}[\tau] &\leq \sum_{n=0}^{N-2} \mathbb{P}(\tau > n) + \sum_{n=N-1}^{\infty} \mathbb{P}(\tau > n) \\ &\leq N-1 + \sum_{n=0}^{\infty} \mathbb{P}(\sum_{k=0}^{N-2} X_{j+(N-1)+k} > 0)^{\lfloor \frac{n}{N-1} \rfloor} \\ &\leq N-1 + \sum_{n=0}^{\infty} \mathbb{P}(\sum_{k=0}^{N-2} X_{j+(N-1)+k} > 0)^{n} \\ &= N-1 + (1 - \mathbb{P}(\sum_{k=0}^{N-2} X_{j+(N-1)+k} > 0))^{-1} < \infty,\end{aligned}$$

where the last inequality follows from (9).