Probability in the Engineering and Informational Sciences, **30**, 2016, 153–184.
doi:10.1017/S0269964815000339

SERVER WAITING TIMES IN INFINITE SUPPLY POLLING SYSTEMS WITH PREPARATION TIMES

JAN-PIETER L. DORSMAN
Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands
E-mail: j.l.dorsman@math.leidenuniv.nl

NIR PEREL
Department of Industrial Engineering and Management, Shenkar - Engineering, Design and Art, Ramat Gan, Israel
E-mail: perelnir@shenkar.ac.il

MARIA VLASIOU
EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

Stochastics, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
E-mail: m.vlasiou@tue.nl

We consider a system consisting of a single server serving 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 queueing networks, to an extension of polling systems and surprisingly to random graphs. We are interested in the waiting time of the server. For the case where the server polls the stations cyclically, we give a sufficient condition for the existence of a limiting waiting-time distribution and we study the tail behavior of the stationary waiting time. Furthermore, assuming that preparation times are exponentially distributed, we describe in depth the resulting Markov chain. We also investigate a model variation where the server does not necessarily poll the stations in a cyclic order, but always serves the customer with the earliest completed preparation phase. We show that the mean waiting time under this dynamic allocation never exceeds that of the cyclic case, but that the waiting-time distributions corresponding to both cases are not necessarily stochastically ordered. Finally, we provide extensive numerical results investigating and comparing the effect of the system's parameters to the performance of the server for both models.

1. INTRODUCTION

We study a model that involves one server polling multiple stations. We initially assume that the server visits N stations in a cyclic order, serving exactly one customer per visit to a station. 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

 0269-9648/15 $25.00

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 analyze the transient, as well as the long-run, probabilistic behavior 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 [14], who considered 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 analyzed under general distributional assumptions in [22]. The model we consider in this paper generalizes 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 [13], since the 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 [12] 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 one-limited polling-type system; cf. [5,7,21]. In general, polling models have attracted a lot of attention in the literature; see, for example, [4,19,26], and the extensive references therein. Limited polling systems are notoriously difficult to analyze as the k-limited service discipline does not satisfy the so-called branching property; see [16]. In our case, we have the added difficulty of an additional preparation phase before service. We assume that when the service of a customer at a station ends, there is always a new customer waiting in front of the same station. In the carousel setting, this means that there is always an ample supply of items to pick. Furthermore, in many service systems, appointments with customers occur on a scheduled basis, so that this assumption is also a natural one in that setting. As a result of this assumption, the analysis of the model is parallel to the study of the server in a one-limited polling-type system, where each of the queues is critically loaded. Note that our main interest in this paper is in the waiting time of the *server*, rather than that of the *customers*.

Yet another way to view the system 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 queueing networks occur naturally in all kinds of information and e-commerce systems, grid systems, and real-time systems such as telecom switches; see [8] and references therein for an overview. Layered queues are characterized by simultaneous or separate phases where entities are no longer classified in the traditional roles of "servers" and "customers", but may also have a dual role of being a server to other entities (of lower layers) and a customer to higher-layer entities. An example of this is found in 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).

We study the waiting time of the server for this model. Under cyclic routing assumptions, the waiting time satisfies the 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 non-negative weights, independently generated according to some common distribution. It is shown in [9] that for certain weight distributions, a limiting result can be proven not only for the maximum independent set, but also for the maximum *weight* of an independent set. What is crucial in these computations is the recursion in Eq. (2); cf. [9, Eq. (3)]. This recursion provides another surprising link between queueing theory and random graphs.

This paper is an extended version of our conference paper [15]. We extend the conference paper mainly in two directions. First, we analyze the system more rigorously and provide several additional limiting results for the waiting-time distribution, such as its tail asymptotics, under cyclic routing assumptions of the server. Second, we extensively study the question of how the waiting-time distribution of the server is affected when we drop the assumption that the server is forced to visit the stations cyclically. Then, the server will always visit the service station which has its preparation phase completed first in an effort to reduce his overall waiting time. Thus, the order in which stations are served will become dynamic. The removal of the cyclic condition has a significant impact on the analysis, since the waiting time in the new dynamic model does not satisfy a recursion such as (2) anymore. Several results comparing the two models were already derived in [24] for the special case of two service stations, but these results generally either do not hold for a larger number of service stations or their derivation is not trivially extended to a general number of service points.

The dynamic model which arises after removal of the cyclic condition turns out to be equivalent to the classical machine-repair problem. This problem, also known as the *computer-terminal model* (cf. [2]) or as the *time-sharing system* (cf. [11, Section 4.11]), is well studied in the literature. In the machine-repair problem, there is a number of machines working in parallel, and a single repairman. As soon as a machine fails, it joins a repair queue in order to be repaired by the repairman. This model is one of the key models to describe problems with a finite input population. A fairly extensive analysis of the machine-repair model can be found in Takács [18, Chapter 5]. The extensive literature available on the machine-repair problem mainly focuses on the waiting times of the machines, but ignores the idle times of the repairman. The latter question has not been treated extensively in the classical literature, perhaps because in the machine-repair problem, the operating time of the machine is usually more valuable than the utilization of the repairman. In our setting, however, we are concerned with the idle times of the repairman.

At a glance, other than the analytical results, the major insights we gain for both models are summarized as follows. First, we observe that, in the cyclic model, variability in preparation times has a greater influence than the variability in service times. Intuitively, this can be explained by the fact that the waiting time depends on the *single* (remaining) preparation time of the queue ahead, but also on *all* service times encountered since the last service at the same queue. Any variance in the service distribution is thus mitigated by the multitude of service times, but this mitigation effect does not arise for the preparation times. In the healthcare setting, one could summarize it as follows: it pays more to have a reliable nurse than a reliable specialist. In the dynamic model, however, it appears that the

waiting time of the server is almost insensitive to the variability of the preparation times. This phenomenon can be understood by drawing an analogy with an Erlang loss model and its well-known insensitivity property. Second, a *small* variability of preparation times actually improves the performance of the server under cyclic routing assumptions in the sense that he waits less frequently; cf. Figure 2. However, it also decreases the throughput, as the server will need to wait much longer in case preparation times are highly variable. Thus, the system's designer may wish to consider how to balance these conflicting goals. Again, this effect does not occur in the dynamic model, since the server selects the queue with the least remaining preparation time when all queues are still in preparation. Next, when deciding how many stations to assign to a server in the cyclic model, the shapes of both the preparation time distribution and the service time distribution play a role, since they affect the throughput of the server. 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. When dropping the cyclic assumption, the expected waiting time of the server decreases, so that even fewer stations are required per server to guarantee a high utilization rate of the server. The last major insight that we gain is of a mathematical nature. We observe that as the number of stations goes to infinity, the waiting times of the server become uncorrelated. The correlation structures of the waiting times, however, turn out to be very surprising. We additionally provide an analytic lower bound on the throughput for the cyclic case and an empirical upper bound. Both of these bounds are easy to compute, converge exponentially to the true throughput as the number of stations goes to infinity and are tight in some cases. Thus, we get quick and accurate estimates on the system's performance. We provide additional intuition and a more in-depth discussion on these and other observations in Sections 4 and 5.3.

The rest of the paper is organized as follows. The general model is presented in Section 2 along with detailed descriptions of the cyclic and dynamic model variations. In Section 3, we provide analytical results for the cyclic model. More specifically, we give a sufficient condition for the existence of a limiting waiting-time distribution and investigate the tail behavior of the waiting time. Under the assumption that preparation times are exponential, we also study the transient behavior of the waiting time and provide the transition matrix of the underlying Markov chain. Section 4 provides insights into the effect of all parameters on the system's performance for the cyclic model. We compare the server's waiting-time distribution of the cyclic model with that of the dynamic model in Section 5. We show that these distributions are not necessarily stochastically ordered. Nevertheless, the mean waiting time in the cyclic case turns out to always be at least as large as the mean waiting time in the dynamic case. Finally, we investigate how the insights obtained for the cyclic model compare to the behavior of the waiting-time distribution in the dynamic model.

2. MODEL DESCRIPTION

We assume that there are $N \geq 2$ identical service stations, $Q_1, \ldots, Q_N$, operated by a single server. Each of these service stations has an infinite supply of customers. Before being served by the server for a duration A, a customer must first undergo a preparation phase with duration B (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. Consequently, at each point in time, there is exactly

one customer at a service station who is either in service, waiting for service or undergoing preparation. Unless otherwise stated, we assume that A and B are continuous random variables with finite means, general distribution functions F_A (F_B) and Laplace–Stieltjes transforms $\alpha(s) = \mathbb{E}[e^{-sA}]$ and $\beta(s) = \mathbb{E}[e^{-sB}]$.

Initially, we are interested in the waiting time of the server when assuming he serves the stations in a cyclic order. Thus, after having served a customer at service station Q_i, the server will move to service station Q_{i+1} to serve a customer there. Note that indices of service stations are to be understood modulo N, so that service station Q_i actually refers to service station $Q_{((i-1) \mod N)+1}$. We will refer to this as the *cyclic model*, or equivalently, the cyclic case. Later on, we compare the performance of this model to the *dynamic model*, or the dynamic case. In this model, the server no longer moves through the service stations in a cyclic manner after completing a service, but visits the service station corresponding to the customer that finishes or has finished its preparation phase the earliest. Thus, in the dynamic model, it is also possible that he visits the same service station twice in a row. We comment on both scenarios in more detail below.

The cyclic model. Let B_n denotes the preparation time of the nth customer served and let A_n be the time the server spends on this customer. We assume that $\{B_n\}_{n\geq 1}$ and $\{A_n\}_{n\geq 1}$ are comprised of independent and identically distributed (i.i.d.) realizations of the random variables B and A. The waiting time W_n^{C} incurred by the server just before serving the nth customer then satisfies the equation

$$W_{n+1}^{\mathrm{C}} = \left(B_{n+1} - \sum_{i=n-N+2}^{n} A_i - \sum_{i=n-N+2}^{n} W_i^{\mathrm{C}} \right)^+. \tag{1}$$

This equation can be rewritten as

$$W_{n+1}^{\mathrm{C}} = \left(X_{n+1} - \sum_{i=n-N+2}^{n} W_i^{\mathrm{C}} \right)^+, \tag{2}$$

where $X_{n+1} = B_{n+1} - \sum_{i=n-N+2}^{n} A_i$. Note that $\{X_n, n \geq 0\}$ is comprised of identically distributed realizations of a random variable X. However, these realizations are not necessarily independent. They are only independent with an $(N-1)$-lag. For example, $\{X_N,\ X_{2N-1},\ X_{3N-2},\ X_{4N-3}, \ldots\}$ are independent. Furthermore, we assume without loss of generality that in the cyclic case, the server first visits Q_1 after time zero. Define $R_{j,n}^{\mathrm{C}}$ to be the residual preparation time at $Q_{(n+j) \mod N}$ just after the completion of the $(n-1)$st service in the cyclic case, $n \geq 1$, $j = 1, \ldots, N-2$. Clearly, $R_{N-1,n}^{\mathrm{C}} = B_{n+N-1}$ and $R_{N,n}^{\mathrm{C}} = W_n^{\mathrm{C}}$. Then, the process $\{(W_n^{\mathrm{C}}, R_{1,n}^{\mathrm{C}}, R_{2,n}^{\mathrm{C}}, \ldots, R_{N-2,n}^{\mathrm{C}}), n \geq 1\}$ is a Markov chain, of which the evolution is given by $W_{n+1}^{\mathrm{C}} = \left(R_{1,n}^{\mathrm{C}} - W_n^{\mathrm{C}} - A_n\right)^+$ and $R_{j,n+1}^{\mathrm{C}} = \left(R_{j+1,n}^{\mathrm{C}} - W_n^{\mathrm{C}} - A_n\right)^+$ for $j = 1, 2, \ldots, N-2$.

The dynamic model. As the number of station visits between two visits of the same station is now stochastic, there is no simple equivalent of (1) available for the waiting times $\{W_n^{\mathrm{D}}, n \geq 0\}$ of the server in the dynamic case. When defining $R_{j,n}^{\mathrm{D}}$ to be the residual preparation time at Q_j just after the $(n-1)$st service, the process $\{(R_{1,n}^{\mathrm{D}}, \ldots, R_{N,n}^{\mathrm{D}}), n \geq 1\}$ also forms a Markov chain. Evidently, we have that $W_n^{\mathrm{D}} = \min_{j\in\{1,\ldots,N\}}\{R_{j,n}^{\mathrm{D}}\}$. Furthermore, we have that $R_{j,n}^{\mathrm{D}}$ is an independent copy of B if the $(n-1)$st customer was served at Q_j. Otherwise, we have that $R_{j,n+1}^{\mathrm{D}} = \left(R_{j,n}^{\mathrm{D}} - W_n^{\mathrm{D}} - A_n\right)^+$.

3. WAITING-TIME ANALYSIS OF THE CYCLIC MODEL

In this section, we study the waiting-time distribution of the server in the cyclic model. First, we investigate the existence of a unique limiting waiting-time distribution in Section 3.1. Then, we study the tail behavior of the stationary waiting time in Section 3.2 for several classes of preparation time distributions. Finally, Section 3.3 shows how to compute the distribution of W_n^{C} for any $n \geq 1$ under the assumption of exponential preparation times. The analysis presented in this section can conceptually be extended easily to allow for phase-type preparation times.

3.1. Existence of a Limiting Waiting-Time Distribution

We will argue in this section that a unique limiting waiting-time distribution exists under the natural assumption that $\mathbb{P}(X \leq 0) > 0$. Note that the stochastic process $\{W_n^{\mathrm{C}}, n \geq 1\}$ is an aperiodic (possibly delayed) regenerative process with regeneration times $\{n : W_n^{\mathrm{C}} = W_{n-1}^{\mathrm{C}} = \cdots = W_{n-2N+4}^{\mathrm{C}} = 0\}$. Colloquially speaking, this is due to the fact that the server's waiting time is independent of past waiting times in case the server did not have to wait in the past two polling cycles. Let j be any regeneration time after $t = 2N-4$. Furthermore, let $\tau = \min\{n : n > 0, W_j^{\mathrm{C}} = W_{j-1}^{\mathrm{C}} = \cdots = W_{j-2N+4}^{\mathrm{C}} = W_{j+n}^{\mathrm{C}} = W_{j+n-1}^{\mathrm{C}} = \cdots = W_{j+n-2N+4}^{\mathrm{C}} = 0\}$, so that τ can be interpreted as the time between two regeneration moments.

We will now show that $\mathbb{E}[\tau]$ is finite, which implies by the standard theory on regenerative processes that the limiting distribution of the waiting time exists and that the waiting-time process converges to it (see, e.g., [1, Corollary VI.1.5 and Theorem VII.3.6]). To this end, observe that for any $n \geq 2N-3$,

$$\mathbb{P}(\tau > n) = \mathbb{P}\left(\bigcap_{i=j+1}^{j+n} \left\{\sum_{k=0}^{2N-4} W_{i-k}^{\mathrm{C}} > 0\right\}\right) \leq \mathbb{P}\left(\bigcap_{i=j+2N-3}^{j+n} \left\{\sum_{k=0}^{2N-4} W_{i-k}^{\mathrm{C}} > 0\right\}\right).$$

Due to (2) and the fact that waiting times are non-negative, X_n is stochastically not smaller than W_n^{C}. In other words, we have that

$$\mathbb{P}\left(W_n^{\mathrm{C}} > 0 \mid W_{n-1}^{\mathrm{C}}, W_{n-2}^{\mathrm{C}}, \ldots\right) \leq \mathbb{P}\left(X_n > 0 \mid W_{n-1}^{\mathrm{C}}, W_{n-2}^{\mathrm{C}}, \ldots\right).$$

We also obviously have that $\mathbb{P}\left(X_n > 0 \mid W_{n-k}^{\mathrm{C}} = 0\right) \leq \mathbb{P}(X_n > 0)$ for any $k \in \{1, 2, \ldots\}$. As a result, we have for any $n \geq 2N-3$ that

$$\begin{aligned}\mathbb{P}(\tau > n) &\leq \mathbb{P}\left(\bigcap_{i=j+2N-3}^{j+n} \left\{\sum_{k=0}^{2N-4} X_{i-k} > 0\right\}\right) \leq \mathbb{P}\left(\bigcap_{i=1}^{\lfloor \frac{n}{2N-3} \rfloor} \left\{\sum_{k=0}^{2N-4} X_{j+i(2N-3)-k} > 0\right\}\right) \\ &= \mathbb{P}\left(\sum_{k=0}^{2N-4} X_{j+2N-3-k} > 0\right)^{\lfloor \frac{n}{2N-3} \rfloor} < \mathbb{P}\left(\sum_{k=1}^{2N-3} X_{j+k} > 0\right)^{\frac{n}{2N-3}-1}, \end{aligned} \tag{3}$$

where the equality follows from the fact that the process $\{X_n, n \geq 0\}$ exhibits no autocorrelation for lag $N-1$ or more. The last inequality holds since $\sum_{k=0}^{2N-4} X_{j+2N-3-k} =$

$\sum_{k=1}^{2N-3} X_{j+k}$ and $\lfloor n/(2N-3) \rfloor > n/(2N-3) - 1$. Additionally, we have that

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{2N-3} X_{j+k} > 0\right) &\leq 1 - \mathbb{P}\left(\bigcap_{k=1}^{2N-3} \left\{X_{j+k} \leq 0\right\}\right) \\ &\quad \times \cdots \times \mathbb{P}\left(X_{j+2N-3} \leq 0 \mid \bigcap_{k=1}^{2N-4} \left\{X_{j+k} \leq 0\right\}\right) \\ &\leq 1 - \mathbb{P}(X \leq 0)^{2N-3}. \end{aligned} \tag{4}$$

The last inequality holds since the process $\{X_n, n \geq 0\}$ exhibits positive autocorrelation with a lag up to $N-2$, but no autocorrelation for lag $N-1$ or more. Thus, we have that $\mathrm{Cov}[\mathbb{1}_{\{X_{n+k} \leq 0\}}, \mathbb{1}_{\{X_n \leq 0\}}] \geq 0$ for any $n > N-1$ and $0 < k \leq N-2$, so that $\mathbb{P}(X_{n+k} \leq 0 \mid X_n \leq 0) \geq \mathbb{P}(X \leq 0)$. For $k > N-2$, however, we have that $\mathbb{P}(X_{n+k} \leq 0 \mid X_n \leq 0) = \mathbb{P}(X \leq 0)$. Finally, from (3), we infer that

$$\begin{aligned}\mathbb{E}[\tau] &= \sum_{n=0}^{2N-4} \mathbb{P}(\tau > n) + \sum_{n=2N-3}^{\infty} \mathbb{P}(\tau > n) \leq 2N - 3 + \sum_{n=0}^{\infty} \mathbb{P}\left(\sum_{k=1}^{2N-3} X_{j+k} > 0\right)^{\frac{n}{2N-3}-1} \\ &\leq 2N - 3 + \sum_{n=0}^{\infty} (1 - \mathbb{P}(X \leq 0)^{2N-3})^{\frac{n}{2N-3}-1} \\ &= 2N - 3 + \frac{1}{1 - \mathbb{P}(X \leq 0)^{2N-3}} \frac{1}{1 - (1 - \mathbb{P}(X \leq 0)^{2N-3})^{\frac{1}{2N-3}}} < \infty, \end{aligned}$$

where the second inequality follows from (4). The last inequality holds true under the assumption that $\mathbb{P}(X \leq 0) \in (0,1)$. Observe that in the trivial case of $\mathbb{P}(X \leq 0) = 1$, the server never waits, resulting in zero waiting times. Therefore, we conclude that a unique limiting distribution exists for the waiting time when $\mathbb{P}(X \leq 0) > 0$. The existence of such a distribution in the theoretical case $\mathbb{P}(X < 0) = 0$ is proved in [23, Section 2.2] for $N = 2$, but this result seems hard to extend to a general value of N.

3.2. Tail Behavior

We now study the tail behavior of W^{C}, the stationary waiting time. For two classes of preparation time distributions, we derive the asymptotic behavior of the probability that the waiting time W^{C} exceeds some large value x. The tail behavior may be useful when, for example, the distribution of W^{C} cannot be computed exactly or when knowledge on the full distribution of W^{C} is not necessary. In the remainder of this section, we write $f \sim g$ for two functions $f(x)$ and $g(x)$ when $\lim_{x\to\infty} f(x)/g(x) = 1$. We also require the notion of regularly varying and rapidly varying functions.

A measurable function $f : (0,\infty) \to (0,\infty)$ is called regularly varying of a finite index κ if

$$\lim_{x\to\infty} \frac{f(lx)}{f(x)} = l^{\kappa}$$

for any $l > 0$. Observe that this definition demands that the index κ is finite. The definition can be extended to include cases for which κ is not finite, leading to the notion of rapid variation. A measurable function $f : (0,\infty) \to (0,\infty)$ is rapidly varying of index $-\infty$ if it

satisfies

$$\lim_{x\to\infty}\frac{f(lx)}{f(x)}=\begin{cases}0 & \text{if } l>1,\\ 1 & \text{if } l=1,\\ \infty & \text{otherwise.}\end{cases}$$

A comprehensive account of the theory and applications of regular variation is given in [3]. By convention, we will call a random variable regularly varying or rapidly varying if its complementary cumulative distribution function has the corresponding property.

We start with the class of preparation time distributions that satisfies

$$\lim_{x\to\infty}\frac{\mathbb{P}\left(B>x+y\right)}{\mathbb{P}\left(B>x\right)}=e^{-\kappa y}$$

for some finite constant $\kappa\geq 0$, or equivalently

$$\lim_{x\to\infty}\frac{\mathbb{P}\left(e^{B}>e^{x}e^{y}\right)}{\mathbb{P}\left(e^{B}>e^{x}\right)}=(e^{y})^{-\kappa}.$$

Thus, we regard the class of distributions of B for which e^B is a regularly varying random variable with index $-\kappa\leq 0$. For $\kappa=0$, this means that the random variable B is long-tailed and thus, in particular, heavy-tailed. If $\kappa>0$, then B is light-tailed, but not lighter than the tail of an exponential distribution.

In order to study the tail behavior of W^{C} for this class of preparation time distributions, we will use the following proposition obtained in [6, Corollary 3.6].

Proposition 3.1: *If $Y>0$ is a regularly varying random variable with index $-\kappa$, $\kappa\geq 0$ and $Z>0$ is a random variable independent of Y satisfying $\mathbb{E}[Z^{\kappa+\epsilon}]<\infty$ for some $\epsilon>0$, then YZ is also regularly varying with index $-\kappa$. In particular, we have that*

$$\mathbb{P}\left(YZ>x\right)\sim\mathbb{E}[Z^{\kappa}]\mathbb{P}\left(Y>x\right).$$

Now, let $\overline{Y}=B-A$, and let $\overline{Z}$ be a random variable with a distribution equal to the limiting distribution of $W_n^{\mathrm{C}}+\sum_{i=n-N+2}^{n-1}(A_i+W_i^{\mathrm{C}})$ as $n\to\infty$ under the conditions of Section 3.1. Then we have, due to the recursion in (1), that $W^{\mathrm{C}}\overset{d}{=}\overline{Y}-\overline{Z}$. The following theorem states that the tail of W behaves asymptotically as the tail of B or the tail of $\overline{Y}$ multiplied by a constant.

Theorem 3.2: *Let e^B be regularly varying with index $-\kappa$, $\kappa>0$. Then, we have for the tail of W^{C} that*

$$\mathbb{P}\left(W^{\mathrm{C}}>x\right)\sim\mathbb{E}[e^{-\kappa(A+\overline{Z})}]\mathbb{P}\left(B>x\right)\ \textit{and}\ \mathbb{P}\left(W^{\mathrm{C}}>x\right)\sim\mathbb{E}[e^{-\kappa\overline{Z}}]\mathbb{P}\left(\overline{Y}>x\right).$$

Proof: We have from (1) that $\mathbb{P}\left(W^{\mathrm{C}}>x\right)=\mathbb{P}\left(B-A-\overline{Z}>x\right)$, or equivalently, that $\mathbb{P}\left(e^{W^{\mathrm{C}}}>e^{x}\right)=\mathbb{P}\left(e^{B}e^{-(A+\overline{Z})}>e^{x}\right)$. Note that $e^{-(A+\overline{Z})}$ is a positive random variable,

which for any $\epsilon > 0$ satisfies

$$\mathbb{E}[e^{-(\kappa+\epsilon)(A+\overline{Z})}] \leq 1 < \infty,$$

as $A + \overline{Z}$ cannot take negative values. Therefore, we obtain by applying Proposition 3.1 with $Y = e^B$ and $Z = e^{-(A+\overline{Z})}$ that

$$\mathbb{P}\left(e^{W^{\mathrm{C}}} > e^x\right) \sim \mathbb{E}[e^{-\kappa(A+\overline{Z})}]\mathbb{P}\left(e^B > e^x\right) = \mathbb{E}[e^{-\kappa(A+\overline{Z})}]\mathbb{P}\left(B > x\right).$$

For the second part of the theorem, note that $\mathbb{E}[e^{-(\kappa+\epsilon)A}] \leq 1 < \infty$ for any $\epsilon > 0$ as A only takes non-negative values. Therefore, since e^B is regularly varying with index $-\kappa$, $e^{\overline{Y}}$ is too by Proposition 3.1. The expression for the tail of W^{C} in terms of the tail of $\overline{Y}$ now follows from an analysis similar to the one above using Proposition 3.1 with $Y = e^{\overline{Y}}$ and $Z = e^{-\overline{Z}}$. ■

An example of a random variable B that satisfies the conditions of this theorem is the one asymptotically having the tail distribution $\mathbb{P}\left(B > x\right) \sim c_0 x^{c_1} e^{-c_2 x}$ for some real-valued constants c_i, $i = 0, 1, 2$, where $c_0, c_2 > 0$.

We now consider the class of preparation time distributions for which e^B is rapidly varying with index $-\infty$, that is,

$$\lim_{x\to\infty} \frac{\mathbb{P}\left(e^B > e^x e^y\right)}{\mathbb{P}\left(e^B > e^x\right)} = \lim_{x\to\infty} \frac{\mathbb{P}\left(B > x + y\right)}{\mathbb{P}\left(B > x\right)} = \begin{cases} 0 & \text{if } y > 0, \\ 1 & \text{if } y = 0, \\ \infty & \text{if } y < 0. \end{cases}$$

This is equivalent to letting the index κ that was given previously go to infinity. For the random variable B, this means that it is extremely light-tailed. As an example, one can think of a distribution for which the tail is given by $\mathbb{P}\left(B > x\right) = e^{-x^p}$, where $p > 1$.

For this class of preparation time distributions, we derive the asymptotic behavior of the tail of W^{C} under the assumption that $\mathbb{P}\left(\overline{Z} = 0\right) > 0$. Thus, we assume among other things that the distribution of A has an atom at zero. The following theorem states that, as before, the tail of W^{C} then behaves asymptotically as the tail of $\overline{Y}$ multiplied by a constant. A similar result under more general assumptions on the distribution of A and B seems hard to obtain, unless $N = 2$ (cf. [23]).

THEOREM 3.3: *Let e^B be rapidly varying with index $-\infty$. If $\mathbb{P}\left(\overline{Z} = 0\right) > 0$, the tail of W^{C} satisfies*

$$\mathbb{P}\left(W^{\mathrm{C}} > x\right) \sim \mathbb{P}\left(\overline{Y} > x\right)\mathbb{P}\left(\overline{Z} = 0\right).$$

PROOF: Note that according to (1),

$$\begin{aligned} \mathbb{P}\left(W^{\mathrm{C}} > x\right) &= \lim_{n\to\infty} \mathbb{P}\left(B_n - \sum_{i=n-N+2}^{n} A_i - \sum_{i=n-N+2}^{n} W_i^{\mathrm{C}} > x\right) = \mathbb{P}\left(\overline{Y} - \overline{Z} > x\right) \\ &= \mathbb{P}\left(\overline{Y} > x\right)\mathbb{P}\left(\overline{Z} = 0\right) + \mathbb{P}\left(\overline{Y} - \overline{Z} > x \mid 0 < \overline{Z} < \epsilon\right)\mathbb{P}\left(0 < \overline{Z} < \epsilon\right) \\ &\quad + \mathbb{P}\left(\overline{Y} - \overline{Z} > x \mid \overline{Z} \geq \epsilon\right)\mathbb{P}\left(\overline{Z} \geq \epsilon\right) \end{aligned} \tag{5}$$

for some $\epsilon > 0$. Since the last two terms of the right-hand side of (5) are non-negative, we conclude immediately that

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left(W^{\mathrm{C}} > x\right)}{\mathbb{P}\left(\overline{Y} > x\right)\mathbb{P}\left(\overline{Z} = 0\right)} \geq 1. \tag{6}$$

Concerning the upper limit, observe that $\mathbb{P}\left(\overline{Y} - \overline{Z} > x \mid 0 < \overline{Z} < \epsilon\right) \leq \mathbb{P}\left(\overline{Y} > x\right)$ and that $\mathbb{P}\left(\overline{Y} - \overline{Z} > x \mid \overline{Z} \geq \epsilon\right) \leq \mathbb{P}\left(\overline{Y} > x + \epsilon\right)$. As e^B is rapidly varying, $e^{\overline{Y}}$ is too; see, for example, [23, Lemma 1]. Therefore, we have for $\epsilon > 0$ that

$$\lim_{x\to\infty} \frac{\mathbb{P}\left(\overline{Y} > x + \epsilon\right)}{\mathbb{P}\left(\overline{Y} > x\right)} = 0.$$

Combining the above arguments, we obtain from (5) that

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left(W^{\mathrm{C}} > x\right)}{\mathbb{P}\left(\overline{Y} > x\right)\mathbb{P}\left(\overline{Z} = 0\right)} \leq 1 + \frac{\mathbb{P}\left(0 < \overline{Z} < \epsilon\right)}{\mathbb{P}\left(\overline{Z} = 0\right)}. \tag{7}$$

By taking the limit $\epsilon \to 0$, we therefore have that

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left(W^{\mathrm{C}} > x\right)}{\mathbb{P}\left(\overline{Y} > x\right)\mathbb{P}\left(\overline{Z} = 0\right)} = 1,$$

since the inequalities in $\mathbb{P}\left(0 < \overline{Z} < \epsilon\right)$ are strict, $\mathbb{P}\left(\overline{Z} = 0\right)$ is positive and the left-hand side of (7) does not depend on ϵ. Combining (6) with this expression now leads to the theorem. ■

3.3. Transient Analysis

In this section, 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 calculate 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 as well as for the mean time between two zero waiting times, both for the transient and the steady-state cases.

3.3.1. The behavior of W^{C}_{n+1}. We show that the waiting time, given that it is positive, is exponentially (μ) distributed. For $n \geq N-1$, we have that

$$\begin{aligned}
&\mathbb{P}\left(W^{\mathrm{C}}_{n+1} > x \mid W^{\mathrm{C}}_{n} = w_n, \ldots, W^{\mathrm{C}}_{n-N+2} = w_{n-N+2}\right) \\
&\quad = \mathbb{P}\left(B_{n+1} > \sum_{i=n-N+2}^{n} A_i + \sum_{i=n-N+2}^{n} w_i + x\right) \\
&\quad = \int_{y_{n-N+2}=0}^{\infty} \cdots \int_{y_n=0}^{\infty} e^{-\mu\left(\sum_{i=n-N+2}^{n}(y_i+w_i)+x\right)} dF_{A_n}(y_n) \ldots dF_{A_{n-N+2}}(y_{n-N+2}) \\
&\quad = (\alpha(\mu))^{N-1} e^{-\mu\left(\sum_{i=n-N+2}^{N} w_i + x\right)},
\end{aligned} \tag{8}$$

where we defined $\alpha(\mu) = \mathbb{E}[e^{-\mu A}]$. From this equation, we conclude that

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

meaning that W_{n+1}^{C}, provided that it is positive, is not affected by the previous $N-1$ waiting times. A direct conclusion is that $\mathbb{P}\left(W_{n+1}^{\mathrm{C}} > x \mid W_{n+1}^{\mathrm{C}} > 0\right) = e^{-\mu x}$ so that

$$\begin{aligned}
\mathbb{P}\left(W_{n+1}^{\mathrm{C}} > x\right) &= \mathbb{P}\left(W_{n+1}^{\mathrm{C}} > x \mid W_{n+1}^{\mathrm{C}} > 0\right)\mathbb{P}\left(W_{n+1}^{\mathrm{C}} > 0\right) \\
&\quad + \mathbb{P}\left(W_{n+1}^{\mathrm{C}} > x \mid W_{n+1}^{\mathrm{C}} = 0\right)\mathbb{P}\left(W_{n+1}^{\mathrm{C}} = 0\right) = e^{-\mu x}\mathbb{P}\left(W_{n+1}^{\mathrm{C}} > 0\right). \qquad \textbf{(9)}
\end{aligned}$$

That is, the distribution of W_n^{C} is a mixture of a mass at zero and the exponential distribution with rate μ, in case $n \geq N-1$. The same result for $1 \leq n < N-1$ follows by performing a similar analysis. The argument can also be applied for W^{C}, the limit of W_n^{C} as $n \to \infty$, so that $\mathbb{P}\left(W^{\mathrm{C}} > x\right) = e^{-\mu x}\mathbb{P}\left(W^{\mathrm{C}} > 0\right)$. We now calculate $\mathbb{P}\left(W_{n+1}^{\mathrm{C}} > 0\right)$ for all n, and $\mathbb{P}\left(W^{\mathrm{C}} > 0\right)$. To this end, we will define a Markov chain and calculate its one-step transition probability matrix.

3.3.2. Construction of a Markov chain. Recall that the process $\{(W_n^{\mathrm{C}}, R_{1,n}^{\mathrm{C}}, R_{2,n}^{\mathrm{C}}, \ldots, R_{N-2,n}^{\mathrm{C}}), n \geq 1\}$ is a Markov chain. We have just showed that W_n^{C}, provided that it is positive, is distributed according to B irrespective of the previous waiting times when B follows an exponential distribution. It is also trivial to see that a residual preparation time $R_{j,N}^{\mathrm{C}}$, given that it is positive, has the same distribution as B, because of the memoryless property of the exponential distribution. Due to these observations, the process $\{(F_n^{\mathrm{C}}, G_{1,n}^{\mathrm{C}}, \ldots, G_{N-2,n}^{\mathrm{C}}), n \geq 1\}$ is a Markov chain on the state space $\mathcal{S}^{\mathrm{C}} = \{0,1\}^{N-1}$, where $F_n^{\mathrm{C}} = \mathbb{1}_{\{W_n^{\mathrm{C}} > 0\}}$ and $G_{j,n}^{\mathrm{C}} = \mathbb{1}_{\{R_{j,n}^{\mathrm{C}} > 0\}}$. A state $i = (i_1, \ldots, i_{N-1}) \in \mathcal{S}^{\mathrm{C}}$ describes the residual preparation time at each station (positive or zero) at the start of the nth waiting time of the server (including zero waiting times). The only station that does not appear in this description is the station the server has just served before this instant, since the residual preparation time there is always larger than zero (or, in other words, $G_{N-1,n}^{\mathrm{C}} = 1$ for all n).

Before we derive the one-step transition probabilities of this Markov chain, we first observe that the Markov chain, provided that it is in state $i \in \mathcal{S}^{\mathrm{C}}$, may not be able to transition directly to any state $j \in \mathcal{S}^{\mathrm{C}}$. This is a result of the fact that a preparation phase that is already completed when transitioning to state i, obviously remains completed until after the following transition unless its corresponding service station is served in between the two transitions. In that case, a new preparation phase starts at the next transition. In other words, the Markov chain can only move from a state i to a state j when $j_{k-1} = 0$ for each $k \in \{2, \ldots, N-1\}$ for which $i_k = 0$. Therefore, we define the set $T(i) = \{j : j_{k-1} \leq i_k \forall k \in \{2, \ldots, N-1\}\}$ to be the set of possible states the Markov chain can transition to after a visit to state i. For any state i, we also define $k_i = \sum_{r=1}^{N-1} i_r$ to be the number of preparation phases that is in progress just before the system moves to state i. Finally, we define $d_{i,j} = k_i - k_j$ to be the difference between these numbers corresponding to states i and j.

Using these definitions, we can now derive the one-step transition probabilities $P_{i,j}$ from any state $i \in \mathcal{S}^{\mathrm{C}}$ to any state $j \in \mathcal{S}^{\mathrm{C}}$. These results are summarized in the following proposition.

PROPOSITION 3.4: *The one-step transition probabilities of the Markov chain* $\{(F_n^{\mathrm{C}}, G_{1,n}^{\mathrm{C}}, \ldots, G_{N-2,n}^{\mathrm{C}}), n \geq 1\}$ *are given by*

$$P_{i,j} = \begin{cases} \displaystyle\sum_{l=0}^{d_{i,j}+1} \binom{d_{i,j}+1}{l} (-1)^l \alpha((k_j+l)\mu) & \text{if } i_1 = 0 \text{ and } j \in T(i), \\ \displaystyle\sum_{l=0}^{d_{i,j}} \binom{d_{i,j}}{l} (-1)^l \frac{\alpha((k_j+l)\mu)}{k_j+l+1} & \text{if } i_1 = 1 \text{ and } j \in T(i), \\ 0 & \text{otherwise} \end{cases}$$

for any $i, j \in \mathcal{S}^{\mathrm{C}}$.

PROOF: When $i_1 = 0$ and $j \in T(i)$, a service phase starts immediately when the Markov chain enters state i. Therefore, the time between the transition to state i and the next transition to state j amounts exactly to the duration of this service phase. As the transition to state i marks the start of a new preparation phase at the service station served just before this transition, the number of preparation phases in progress just after this transition equals $k_i + 1$. If the chain then transitions to j, it means that exactly k_j of these preparation phases should still be in progress after the transition to state j. The other $(k_i + 1) - k_j = d_{i,j} + 1$ preparation phases, however, must finish over the course of a service time A. Therefore, we have in this case that

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

When $i_1 = 1$ and $j \in T(i)$, the time until the transition to state j does not only consist of a service time A, but also of some waiting time needed for the preparation phase at the server's location to finish. We have seen in Section 3.3.1 that the distribution of this waiting time equals that of B, independently of other waiting times. Of the $k_i + 1$ preparation phases just after the transition to state i, the preparation phase at the server's location finishes at any rate before the next transition. Consequently, for the Markov chain to transition from state i to state j, exactly k_j of the remaining k_i preparation phases must still be in progress after the transition to state j, and the other $k_i - k_j = d_{i,j}$ should not. Thus, for this case, we have that

$$\begin{aligned} P_{i,j} &= \int_{x=0}^{\infty} \int_{y=0}^{\infty} (1 - e^{-\mu(x+y)})^{d_{i,j}} e^{-k_j \mu(x+y)} \mu e^{-\mu x} dF_A(y) dx \\ &= \sum_{l=0}^{d_{i,j}} \binom{d_{i,j}}{l} (-1)^l \frac{\alpha((k_j+l)\mu)}{k_j+l+1}. \end{aligned}$$

Finally, it is obvious by definition of $T(i)$ that $P_{i,j} = 0$ if $j \notin T(i)$. This completes the derivation of the one-step transition probability matrix. ■

Now that the one-step transition probabilities are derived, the one-step transition probability matrix $P = (P_{i,j})_{i,j \in \mathcal{S}^{\mathrm{C}}}$ can be constructed, for example, by arranging all states in

lexicographic order. Using this matrix, one can compute the unknown $\mathbb{P}\left(W_n^{\mathrm{C}} > 0\right)$ needed to obtain the transient distribution of W_n^{C} for any n (cf. (9)) or, in case $n \to \infty$, the stationary distribution of W^{C}. Without loss of generality, we can assume that the system starts in an arbitrary state $k \in \mathcal{S}^{\mathrm{C}}$. Let e_k be the unit vector of which the entry at the index which corresponds to state k equals one (and all other elements equal zero). Then, by standard theory on Markov chains, $\mathbb{P}\left(W_n^{\mathrm{C}} > 0\right)$ equals the sum of the entries of the vector $e_k P^{n-1}$ that, according to the ordering of states chosen, correspond to states for which the first element equals one (i.e., a non-zero waiting time). Likewise, the steady-state probability $\mathbb{P}\left(W^{\mathrm{C}} > 0\right)$ can be found by computing the unique vector π satisfying $\pi = \pi P$ and $\sum_{i \in \mathcal{S}^{\mathrm{C}}} \pi_i = 1$. The probability $\mathbb{P}\left(W^{\mathrm{C}} > 0\right)$ is then again given by the sum of the entries of π that correspond to states of which the first element equals one. For illustratory purposes, we present a detailed analysis for the case with $N = 3$ stations in the next section. We again consider exponentially distributed preparation times, although the analysis can evidently extend to phase-type distributions.

3.3.3. Analysis for $N = 3$ stations. In this section, we calculate the limiting distribution (W, R) of the Markov chain and the (transient) distribution of W_n^{C} when $N = 3$. We also derive the covariance $\mathrm{Cov}[W_n^{\mathrm{C}}, W_{n+k}^{\mathrm{C}}]$ and the distribution function of the number of visits between two successive zero waiting times of the server. Observe that these are not necessarily regenerative points. We first derive the one-step transition probability matrix of the Markov chain.

For the case $N = 3$, there are only four relevant states in the corresponding Markov chain so that $\mathcal{S}^{\mathrm{C}} = \{(0,0),(0,1),(1,0),(1,1)\}$. We arrange the states lexicographically, so that the columns and rows of the matrix P correspond to $(0,0),(0,1),(1,0)$, and $(1,1)$ respectively. Then, using Proposition 3.4, we have that the one-step transition probability matrix is given by

$$P = \begin{pmatrix} 1-\alpha(\mu) & \alpha(\mu) & 0 & 0 \\ 1-2\alpha(\mu)+\alpha(2\mu) & \alpha(\mu)-\alpha(2\mu) & \alpha(\mu)-\alpha(2\mu) & \alpha(2\mu) \\ 1-\frac{1}{2}\alpha(\mu) & \frac{1}{2}\alpha(\mu) & 0 & 0 \\ 1-\alpha(\mu)+\frac{1}{3}\alpha(2\mu) & \frac{1}{2}\alpha(\mu)-\frac{1}{3}\alpha(2\mu) & \frac{1}{2}\alpha(\mu)-\frac{1}{3}\alpha(2\mu) & \frac{1}{3}\alpha(2\mu) \end{pmatrix}. \tag{10}$$

The unique limiting distribution of the Markov chain is given by the vector π satisfying $\pi = \pi P$ and $\sum_{i \in \mathcal{S}^{\mathrm{C}}} \pi_i = 1$. After some computations, we obtain

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

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

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

and

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

When the system starts, we assume that a preparation phase is initiated at each service point. Thus, the Markov chain starts in the state $(1,1)$ at $n=1$. The event $W_n^{\mathrm{C}}>0$ coincides with the event that the Markov chain finds itself in the state $(1,0)$ or $(1,1)$ after $n-1$ transitions. The probability of the latter event equals $P^{(n-1)}_{(1,1),(1,0)}+P^{(n-1)}_{(1,1),(1,1)}$, so that a combination with (9) yields the following expression for the transient waiting-time distribution:

$$\mathbb{P}\left(W_n^{\mathrm{C}}>x\right)=e^{-\mu x}\left(P^{(n-1)}_{(1,1),(1,0)}+P^{(n-1)}_{(1,1),(1,1)}\right)$$

for all $x>0$. Similarly, an expression for the steady-state waiting-time distribution is given by

$$\mathbb{P}\left(W^{\mathrm{C}}>x\right)=e^{-\mu x}\left(\pi_{(1,0)}+\pi_{(1,1)}\right).$$

Next, to calculate the auto-covariance $\mathrm{Cov}[W_n^{\mathrm{C}},W_{n+k}^{\mathrm{C}}]=\mathbb{E}[W_n^{\mathrm{C}}W_{n+k}^{\mathrm{C}}]-\mathbb{E}[W_n^{\mathrm{C}}]\mathbb{E}[W_{n+k}^{\mathrm{C}}]$, observe that for all $k\geq 0$,

$$\begin{aligned}\mathbb{E}[W_{n+k}^{\mathrm{C}}]&=\mathbb{E}[W_{n+k}^{\mathrm{C}}\mid W_{n+k}^{\mathrm{C}}>0]\mathbb{P}\left(W_{n+k}^{\mathrm{C}}>0\right)=\frac{1}{\mu}\mathbb{P}\left(W_{n+k}^{\mathrm{C}}>0\right)\\&=\frac{1}{\mu}\left(P^{(n+k-1)}_{(1,1),(1,0)}+P^{(n+k-1)}_{(1,1),(1,1)}\right).\end{aligned}$$

Furthermore, we have that $\mathbb{E}[W_n^{\mathrm{C}}W_{n+k}^{\mathrm{C}}]=\mathbb{E}[W_n^{\mathrm{C}}W_{n+k}^{\mathrm{C}}\mid W_n^{\mathrm{C}}>0,W_{n+k}^{\mathrm{C}}>0]\mathbb{P}\left(W_n^{\mathrm{C}}>0,\right.$ $\left.W_{n+k}^{\mathrm{C}}>0\right)$, where

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

and

$$\begin{aligned}\mathbb{P}\left(W_n^{\mathrm{C}}>0,W_{n_k}^{\mathrm{C}}>0\right)&=\mathbb{P}\left(W_{n+k}^{\mathrm{C}}>0\mid W_n^{\mathrm{C}}>0\right)\mathbb{P}\left(W_n^{\mathrm{C}}>0\right)\\&=\mathbb{P}\left(W_{n+k}^{\mathrm{C}}>0\mid W_n^{\mathrm{C}}>0,R_{1,n}^{\mathrm{C}}=0\right)\mathbb{P}\left(W_n^{\mathrm{C}}>0,R_{1,n}^{\mathrm{C}}=0\right)\\&\quad+\mathbb{P}\left(W_{n+k}^{\mathrm{C}}>0\mid W_n^{\mathrm{C}}>0,R_{1,n}^{\mathrm{C}}>0\right)\mathbb{P}\left(W_n^{\mathrm{C}}>0,R_{1,n}^{\mathrm{C}}>0\right)\\&=\left(P^{(k)}_{(1,0),(1,0)}+P^{(k)}_{(1,0),(1,1)}\right)P^{(n-1)}_{(1,1),(1,0)}\\&\quad+\left(P^{(k)}_{(1,1),(1,0)}+P^{(k)}_{(1,1),(1,1)}\right)P^{(n-1)}_{(1,1),(1,1)}.\end{aligned}$$

Therefore, we obtain by combining the expressions above that

$$\begin{aligned}\mathrm{Cov}[W_n^{\mathrm{C}},W_{n+k}^{\mathrm{C}}]&=\frac{1}{\mu^2}\left(\left(P^{(k)}_{(1,0),(1,0)}+P^{(k)}_{(1,0),(1,1)}\right)P^{(n-1)}_{(1,1),(1,0)}\right.\\&\quad\left.+\left(P^{(k)}_{(1,1),(1,0)}+P^{(k)}_{(1,1),(1,1)}\right)P^{(n-1)}_{(1,1),(1,1)}\right)\\&\quad-\frac{1}{\mu^2}\left(P^{(n-1)}_{(1,1),(1,0)}+P^{(n-1)}_{(1,1),(1,1)}\right)\left(P^{(n+k-1)}_{(1,1),(1,0)}+P^{(n+k-1)}_{(1,1),(1,1)}\right).\end{aligned}\tag{11}$$

Last, we compute the distribution and expectation of visits between two consecutive zero waiting times. Suppose that $W_n^{\mathrm{C}}=0$ and define C_n^{C} to be the length from the moment

that $W_n^{\mathrm{C}} = 0$ until the next time that the server's waiting time is zero. In other words,

$$C_n^{\mathrm{C}} = \inf_{k\geq 1}\{k : W_{n+k}^{\mathrm{C}} = 0 \mid W_n^{\mathrm{C}} = 0\}.$$

Observe that for $N = 2$, the points $\{C_n^{\mathrm{C}}, n \geq 1\}$ constitute regenerative times for the waiting-time process $\{W_n^c, n \geq 0\}$, as a zero waiting marks a regenerative point in that case. However, for larger N, this is not necessarily the case. The results are summarized in the following proposition.

PROPOSITION 3.5: *The distribution of* C_n^{C} *is given by*

$$\mathbb{P}\left(C_n^{\mathrm{C}} = 1\right) = 1 - \frac{P_{(1,1),(0,1)}^{(n-1)}}{P_{(1,1),(0,0)}^{(n-1)} + P_{(1,1),(0,1)}^{(n-1)}}\alpha(\mu),$$

$$\mathbb{P}\left(C_n^{\mathrm{C}} = 2\right) = \frac{P_{(1,1),(0,1)}^{(n-1)}}{P_{(1,1),(0,0)}^{(n-1)} + P_{(1,1),(0,1)}^{(n-1)}}\alpha(\mu)\left(1 - \frac{1}{2}\alpha(2\mu)\right)$$

and

$$\mathbb{P}\left(C_n^{\mathrm{C}} = k\right) = \frac{P_{(1,1),(0,1)}^{(n-1)}}{P_{(1,1),(0,0)}^{(n-1)} + P_{(1,1),(0,1)}^{(n-1)}}\alpha(\mu)\left(\frac{1}{2}\alpha(2\mu)\right)\left(\frac{1}{3}\alpha(2\mu)\right)^{k-3} \times \left(1 - \frac{1}{3}\alpha(2\mu)\right) \quad \textit{for } k \geq 3.$$

Moreover,

$$\mathbb{E}[C_n^{\mathrm{C}}] = 1 + \frac{P_{(1,1),(0,1)}^{(n-1)}}{P_{(1,1),(0,0)}^{(n-1)} + P_{(1,1),(0,1)}^{(n-1)}}\alpha(\mu)\left(\frac{6 + \alpha(2\mu)}{6 - 2\alpha(2\mu)}\right). \tag{12}$$

PROOF: For $k = 1$, we have that

$$\begin{aligned}
\mathbb{P}\left(C_n^{\mathrm{C}} = 1\right) &= \mathbb{P}\left(W_{n+1}^{\mathrm{C}} = 0 \mid W_n^{\mathrm{C}} = 0\right) \\
&= \mathbb{P}\left(W_{n+1}^{\mathrm{C}} = 0 \mid W_n^{\mathrm{C}} = 0, R_{1,n}^{\mathrm{C}} = 0\right)\mathbb{P}\left(R_{1,n}^{\mathrm{C}} = 0 \mid W_n^{\mathrm{C}} = 0\right) \\
&\quad + \mathbb{P}\left(W_{n+1}^{\mathrm{C}} = 0 \mid W_n^{\mathrm{C}} = 0, R_{1,n}^{\mathrm{C}} > 0\right)\mathbb{P}\left(R_{1,n}^{\mathrm{C}} > 0 \mid W_n^{\mathrm{C}} = 0\right) \\
&= \left(P_{(0,0),(0,0)} + P_{(0,0),(0,1)}\right)\frac{P_{(1,1),(0,0)}^{(n-1)}}{\mathbb{P}\left(W_n^{\mathrm{C}} = 0\right)} + \left(P_{(0,1),(0,0)} + P_{(0,1),(0,1)}\right)\frac{P_{(1,1),(0,1)}^{(n-1)}}{\mathbb{P}\left(W_n^{\mathrm{C}} = 0\right)} \\
&= \frac{P_{(1,1),(0,0)}^{(n-1)}}{\mathbb{P}\left(W_n^{\mathrm{C}} = 0\right)} + (1 - \alpha(\mu))\frac{P_{(1,1),(0,1)}^{(n-1)}}{\mathbb{P}\left(W_n^{\mathrm{C}} = 0\right)} = 1 - \frac{P_{(1,1),(0,1)}^{(n-1)}}{P_{(1,1),(0,0)}^{(n-1)} + P_{(1,1),(0,1)}^{(n-1)}}\alpha(\mu).
\end{aligned}$$

The results for

$$\mathbb{P}\left(C_n^{\mathrm{C}} = i\right) = \mathbb{P}\left(W_{n+i}^{\mathrm{C}} = 0, W_{n+i-1}^{\mathrm{C}} > 0, \ldots, W_{n+1}^{\mathrm{C}} > 0 \mid W_n^{\mathrm{C}} = 0\right)$$

with $i > 1$ follow by expanding this expression into $\mathbb{P}\left(W_{n+1}^{\mathrm{C}} > 0 \mid W_n^{\mathrm{C}} = 0\right)$ as well as probabilities of the form $\mathbb{P}\left(W_{j+2}^{\mathrm{C}} > 0 \mid W_{j+1}^{\mathrm{C}} > 0, W_j^{\mathrm{C}} = 0\right)$ and $\mathbb{P}(W_{j+2}^{\mathrm{C}} > 0 \mid W_{j+1}^{\mathrm{C}} > 0,$

$W_j^{\mathrm{C}} > 0$); see also [25, Section 3.2]. Similar to the derivations above, these probabilities can be computed using the Markov chain formulation of the previous section. The expectation in (12) then follows by computing $\mathbb{E}[C_n^{\mathrm{C}}] = \sum_{k=0}^{\infty} k\mathbb{P}\left(C_n^{\mathrm{C}} = k\right)$. ■

Remark 3.1: Throughout this section, we assumed that preparation times are equally distributed at each of the service stations. One might also be interested in the case where the duration of a customer's preparation phase at service station Q_i is exponential with a station-specific rate μ_i. Then, it follows immediately from the analysis of Section 3.3.1 that the server's waiting time at Q_i, provided that it is positive, is also exponentially (μ_i) distributed. Furthermore, the size of the mass at zero can still be computed by constructing a Markov chain using the same conceptual methods. However, in this case, the current position of the server needs to be included in the state space to retain the Markov property, and the residual preparation times in the system are not necessarily identically distributed anymore. Therefore, the expressions will become more cumbersome, providing little additional insight into the behavior of the system.

Remark 3.2: In this section, we mainly studied the waiting time W^{C} of the server as a performance measure. Another important performance measure pertaining to the system is the throughput θ^{C}, that is, the mean number of customers that finish their service per unit of time. Observe that θ^{C} is equal to the number of customers N served per cycle over the expected cycle length, which has duration $N(\mathbb{E}[W^{\mathrm{C}}] + \mathbb{E}[A])$. Thus,

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

see also [14]. As such, the results of this section can be readily applied to analyze the throughput of the system, since $\mathbb{E}[A]$ is a known constant. In Section 4, we will focus on the impact of the parameter settings on the throughput of the system.

4. INSIGHTS

In the previous sections, we gave closed-form expressions for exponentially distributed preparation times. Here, we obtain general insights into the behavior of the cyclic model by simulation on a larger range of parameter settings. We vary, among other things, the number of stations and the distributions of the preparation and service times. We focus on the effect of the first two moments of the preparation and service times to the throughput. For their distributions, we choose phase-type distributions based on two-moment-fit approximations commonly used in the literature; see, for example, [20, pp. 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, for example, Figure 1, where the throughput θ^{C} is plotted versus the number of queues N. We observe the throughput for various variability settings for both the time components. We fix the means at $\mathbb{E}[A] = \mathbb{E}[B] = 1$, and first consider the same phase-type distributions with low variability for both the preparation and service time, that is, $\mathbb{E}[A^2] = \mathbb{E}[B^2] = 1.5$ (solid curve). We also consider the case with highly variable service times only, that is, $\mathbb{E}[A^2] = 10, \mathbb{E}[B^2] = 1.5$ (dotted curve) and highly variable preparation times only, that is, $\mathbb{E}[A^2] = 1.5, \mathbb{E}[B^2] = 10$ (dashed curve). Although the variability of preparation

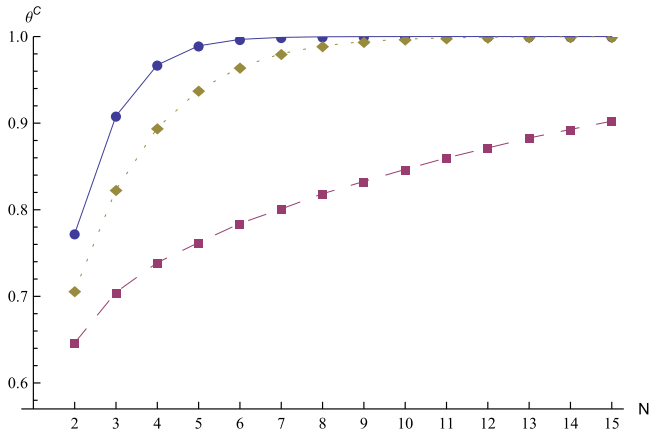

FIGURE **1.** (Color online) Throughput versus the number of stations for moderately variable preparation and service times (solid curve), highly variable service times (dotted curve), and highly variable preparation times (dashed curve).

times and service times is 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.

Intuitively, this can be understood as follows. 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. Therefore, the effect of variability in service times is less far-reaching, as the consequences of a large variance of the service time distribution are mitigated by the fact that the waiting time only depends on a sum of service times. In other words, in service systems, 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 (larger variability) does influence the system, but not as dramatically as the preparation times do.

A similar effect is observed in Figure 2, where the mean number of positive waiting times 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 become 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 resistant 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.

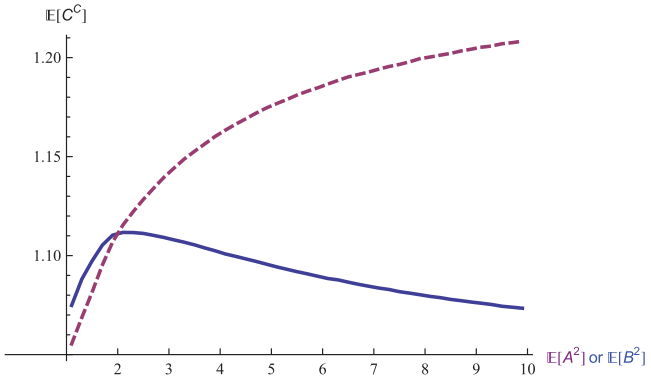

FIGURE **2.** (Color online) Mean time between two zero waiting times versus $\mathbb{E}[B^2]$ (solid curve) and $\mathbb{E}[A^2]$ (dashed curve).

In summary, we can say that an increase in the 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 for exponential preparation and service times with rates 1 and 10, respectively. As we can see in Figure 3, correlations exhibit a periodic structure, which is natural as it corresponds to a return to the first station. Moreover, as the lag increases, the waiting times become uncorrelated, which is again a natural conclusion. As shown in Section 3.1, there exists a unique limiting waiting-time distribution and the system converges to it, thus as time goes to infinity, the system

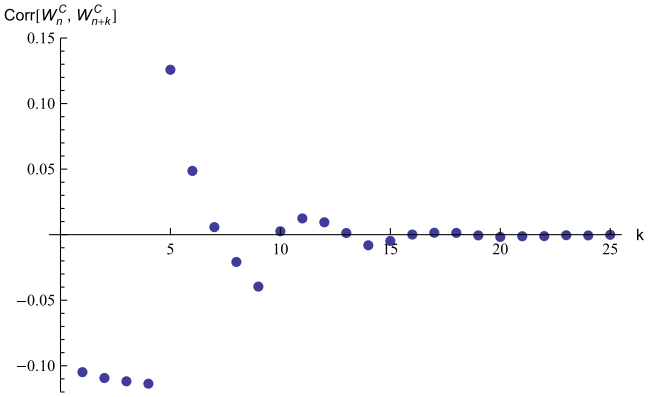

FIGURE **3.** (Color online) Correlations exhibit periodicity.

converges to a steady state regardless of the initial state. Thus, the correlation between waiting times goes to zero. Although the convergence to zero correlations is expected, the way this happens is intriguing. One may expect some form of periodicity, but intuitively, 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. For instance, in the warehouse example given earlier, the more carousels assigned to the picker, the better his utilization. However, the utilization 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 $\mathbb{E}[A]^{-1}$ whenever the server is not forced to wait. 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 does not converge very fast to the optimal throughput when assigning additional stations to the server. Variability in the service times influences the system, but the convergence follows more or less the pattern of the exponential case.

Intuitively, the effect of the variability of the preparation time and service time distributions on the rate of convergence is easily understood, as highly variable (and thus potentially very large) preparation times may sometimes force the server to suffer from very long waiting times. This evidently has a negative impact on the throughput.

When all distributions are exponential, it is evident that the only quantity that matters in the determination of the throughput is $r = \mathbb{E}[B]/\mathbb{E}[A]$. In order to determine the optimal number of stations to assign to a server, we plot in Figure 4 the throughput θ^{C} versus the number of stations N for three cases of r, namely for $r = 0.5$ (top curve), $r = 1$ (solid curve) and $r = 2$ (dotted curve). In all three cases, the underlying distributions are exponential. What we observe is that when $r \leq 1$, that is, 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,

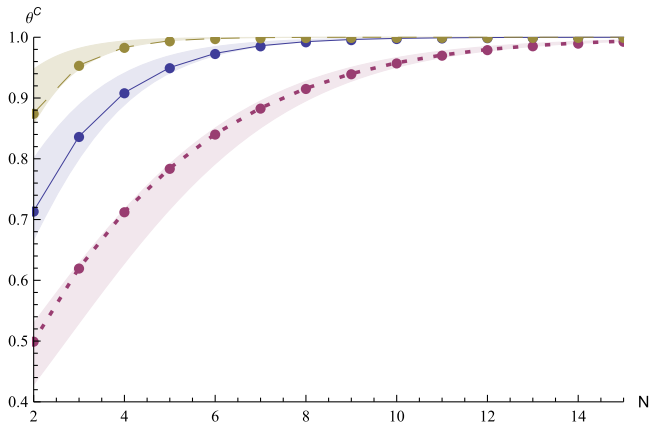

FIGURE 4. (Color online) Throughput versus the number of stations for small (dashed curve), moderate (solid curve), and large preparation times (dotted curve).

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 $r \leq 1$ 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. Recall that the throughput θ^{C} satisfies

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

An approximation $\widetilde{\theta}_N^{\mathrm{C}}$ of θ^{C} can be produced by replacing $\mathbb{E}[W]$ in the denominator by the mean residual preparation time multiplied with a rough estimate that the server has to wait, for example, $\mathbb{P}(B > A_1 + \cdots + A_{N-1})$. Then, for exponentially (μ) distributed preparation times,

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

We observe that this expression is a lower bound of the throughput, since the actual (stationary) probability that a server has to wait equals $\lim_{n\to\infty} \mathbb{P}(B > A_{n-N+2} + \cdots + A_n + W_{n-N+2} + \cdots + W_n)$ and is thus smaller. We also observe empirically that $\widetilde{\theta}_{N+1}^{\mathrm{C}}$ provides an upper bound of the throughput in the scenarios we examined. 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.

5. COMPARISON WITH THE DYNAMIC MODEL

In this section, we compare the performance of the cyclic model with that of the dynamic model described in Section 2, which in the literature is known as the machine-repair problem. In the classical machine-repair problem, there is a number of machines that are served by a unique repairman when they fail. The machines are working independently and as soon as a machine fails, it joins a queue formed in front of the repairman where it is served in order of arrival. A machine that is repaired is assumed to be as good as new. The machine-repair problem with N machines is thus completely equivalent to the model studied in the previous sections, after we drop the assumption that the server visits the service stations cyclically. After a service, the server will instead visit the service station corresponding to the customer who completes or has had its preparation completed earlier than all of the other customers that are first in line at the other service stations. It is thus also possible for the server to serve two customers in a row at the same service station, in case all other service stations still have a preparation phase in progress when the preparation phase following the service of the first customer completes. In the machine-repair model, there are N machines that work in parallel (the service stations), the preparation time of the customer is equivalent to the life time of the machine until it fails and the service time of the customer is the time the repairman needs to repair the machine. Thus, the waiting time of the server in the dynamic model is equivalent to the idle time of the repairman between the repair of one machine and the breakdown of the next. Although the machine-repair problem is thoroughly treated in the literature, relatively little attention has been given to the idle time of the repairman. In the following we will refer to the *server* or *customers*

instead of the *repairman* or *machines* in order to illustrate the analogies between the two models.

Although the waiting time of the server has received little attention, this quantity of interest may be analyzed using a Markov chain approach when assuming phase-type preparation times. For exponentially (μ) distributed preparation times, the waiting-time distribution is obtained as follows. Evidently, a non-zero waiting time occurs in the system only if just after the end of a service, a preparation is in progress at every service station. The waiting time then lasts until one of these N preparation times finishes. Analogously to Section 3.3.1, due to the memoryless distribution of the exponential distribution, the waiting time, provided that it is positive, is thus exponentially $(N\mu)$ distributed

$$\mathbb{P}\left(W_n^{\mathrm{D}} > x\right) = e^{-N\mu x}\mathbb{P}\left(W_n^{\mathrm{D}} > 0\right).$$

To compute $\mathbb{P}\left(W_n^{\mathrm{D}} > 0\right)$, we formulate a Markov chain similar to Section 3.3. Let Z_n^{D} be the number of preparation phases in progress in the complete system just after the service of the nth customer. Then, again due to the memoryless process of the exponential distribution, $\{Z_n^{\mathrm{D}}, n \geq 0\}$ constitutes a Markov chain on the state space $\{1, \ldots, N\}$. Observe that zero is not included in the state space, as the end of a service always marks the start of a preparation phase. The one-step transition probability from state i to state j is then given by

$$P_{i,j} = \begin{cases} \binom{i}{j-1} \sum_{k=0}^{i-j+1} \binom{i-j+1}{k} (-1)^k \alpha((k+j-1)\mu), & \text{if } i \in \{1, \ldots, N-1\}, \\ & j \in \{1, \ldots, i+1\}, \\ \binom{N-1}{j-1} \sum_{k=0}^{N-j} \binom{N-j}{k} (-1)^k \alpha((k+j-1)\mu), & \text{if } i = N, j \in \{1, \ldots, N\}, \\ 0 & \text{otherwise.} \end{cases}$$

The expression for $i \in \{1, \ldots, N-1\}$ and $j \in \{1, \ldots, i+1\}$ follows by noting that in such case $i - j + 1$ preparation phases have been completed during the service time that marks the transition and $j - 1$ preparation phases have not. The distribution of the number of phases completed during this service time A is obviously binomially distributed with parameters $i - 1$ and $1 - e^{-\mu A}$. Therefore, we have that

$$\begin{aligned} P_{i,j} &= \int_{x=0}^{\infty} \binom{i}{j-1} (1 - e^{-\mu x})^{i-j+1} (e^{-\mu x})^{j-1} dF_A(x) \\ &= \binom{i}{j-1} \sum_{k=0}^{i-j+1} \binom{i-j+1}{k} (-1)^k \alpha((k+j-1)\mu) \end{aligned}$$

for $i \in \{1, \ldots, N-1\}$, $j \in \{1, \ldots, i+1\}$. The one-step transition probability for $i = N$ and $j \in \{1, \ldots, N\}$ follows by noting that in that case first one preparation phase has to finish before service can start. Therefore, $P_{N,j} = P_{N-1,j}$ for all $j \in \{1, \ldots, N-1\}$. Finally, transitions corresponding to any other combination of states are not possible, leading to a transition probability of zero. Now that the Markov chain is constructed, we have that

$$\mathbb{P}\left(W_n^{\mathrm{D}} > 0\right) = \mathbb{P}\left(Z_{n-1}^{\mathrm{D}} = N\right).$$

Thus, $\mathbb{P}\left(W_n^{\mathrm{D}} > 0\right)$, as well as its steady-state version $\lim_{n\to\infty} \mathbb{P}\left(W_n^{\mathrm{D}} > 0\right)$, can be computed using standard Markov chain techniques. Note that the latter limiting probability indeed

exists, since we have an aperiodic and irreducible Markov chain due to the fact that the distributions of A and B are continuous. Similar to computations in Section 3.3.3, also expressions for the auto-covariance and expected number of transitions between two zero waiting times can be computed by analyzing the constructed Markov chain. This concludes the analysis for exponential preparation times. Conceptually, this analysis can easily be extended to phase-type distribution times, at the cost of more cumbersome expressions.

Now that we know how to compute the waiting-time distribution of the dynamic model for phase-type preparation times, we investigate whether there is any connection between the waiting-time distributions of both models. In Section 5.1, we will observe that the waiting times of both models are not necessarily stochastically ordered. Nevertheless, Section 5.2 shows by means of a sample-path argument that the mean waiting time in the cyclic case is never shorter than that of the dynamic case for any distribution of the preparation time and the service time. Finally, in Section 5.3, we study how the insights obtained in Section 4 for the cyclic model compare to the dynamic model using numerical results.

5.1. Stochastic Ordering of the Waiting Times

Intuitively, one might argue that the waiting time W^{C} of the cyclic system is stochastically larger than or equal to the waiting time W^{D} of the dynamic system, since one expects that large waiting times occur with higher probability in the cyclic system. In other words, one may conjecture that $\mathbb{P}\left(W^{\mathrm{C}} > x\right) \geq \mathbb{P}\left(W^{\mathrm{D}} > x\right)$ for all $x \geq 0$. However, this is not necessarily true. One may think of a theoretical setting where the duration of a service time always equals zero. Then, we have for the cyclic case that the nth waiting time is zero if the preparation time B_n preceding the service of the nth customer is already completed when the server arrives at the service station. This happens, for example, with positive probability when preparation times are exponentially (μ) distributed, leading to $\mathbb{P}\left(W^{\mathrm{C}} > 0\right) < 1$. In the dynamic case, a zero waiting time could only occur if two preparation phases of different service stations finish at exactly the same point in time. This is, however, not possible, since preparation times are continuously distributed. Hence, we have that $\mathbb{P}\left(W^{\mathrm{D}} > 0\right) = 1$, providing a counterexample to the conjecture mentioned above.

This theoretical setting is not the only possible counterexample. Figure 5(a) depicts the waiting-time distributions for both the cyclic and the dynamic cases in a system with $N = 3$ service stations, standard-exponential preparation times and exponential (10) service times. This figure shows that a lack of stochastic ordering can occur in a realistic setting, as

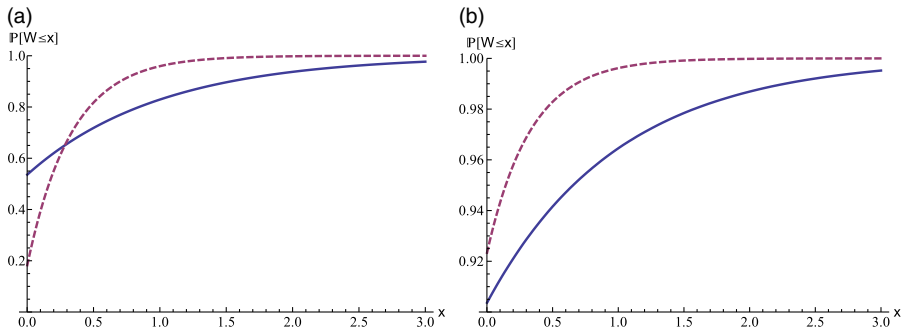

FIGURE **5.** (Color online) Waiting-time distribution for the cyclic (solid curve) and dynamic model (dashed curve) for $N = 3$ and standard-exponential preparation times. Service times are exponentially distributed with $\mathbb{E}[A] = 0.1$ (a) and $\mathbb{E}[A] = 2$(b).

there clearly exist values of x in this case for which $\mathbb{P}\left(W^{\mathrm{C}} > x\right) < \mathbb{P}\left(W^{\mathrm{D}} > x\right)$. Of course, there also exist systems for which the waiting times are actually stochastically ordered. For instance, Figure 5(b) shows the waiting-time distributions for the same example, except that the service times are now exponentially (0.5) distributed instead. The figure suggests that the waiting-time distributions now never intersect, which implies that they are indeed stochastically ordered. Observe though that a stochastic ordering is not possible in case $N = 2$. It was shown in [24, Theorem 4] that for that case $\mathbb{P}\left(W^{\mathrm{C}} > 0\right) \leq \mathbb{P}\left(W^{\mathrm{D}} > 0\right)$ for all distributions of A and B and that there does not exist a stochastic ordering for the waiting-time distributions in case preparation times are non-deterministic.

As it is now clear that the waiting-time distributions are not necessarily stochastically ordered, one may still argue that there must at least exist an increasing convex ordering. In other words, one might expect that $\mathbb{E}[\phi(W^{\mathrm{C}})] \geq \mathbb{E}[\phi(W^{\mathrm{D}})]$ for any increasing convex function ϕ. If the waiting-time distributions intersect exactly once like in Figure 5(a), the Karlin–Novikoff cut-criterion (cf. [17]) implies that a convex ordering indeed exists. However, the second example in Figure 5(b) shows that there is not always such an intersection, so that the existence of an increasing convex ordering for the general case is hard to prove. Therefore, we focus on the expected waiting times instead in the next section.

5.2. Mean Waiting Times

Although the waiting-time distributions of the cyclic case and the dynamic case are not necessarily stochastically ordered, one may still reasonably expect that $\mathbb{E}[W^{\mathrm{C}}] \geq \mathbb{E}[W^{\mathrm{D}}]$. In this section, we prove that this weaker conjecture, contrary to the ones in the previous section, holds true for *any* non-negative distribution for A and B by using a sample path argument. We assume the sequences of realizations $\{b_i, i \geq 1\}$ and $\{a_i, i \geq 1\}$ for the preparation and service times, respectively, to be the same for both scenarios. More specifically, we assume that in both cases the ith customer that leaves the system does so after having received a service with duration a_i, after which a new customer at the same service station initiates a preparation phase with duration b_i. Furthermore, we assume that when both systems start up, the remaining preparation time of the customer at Q_j at time zero equals ζ_j, $j = 1, \ldots, N$.

To prove that the mean waiting time of the server in the dynamic case does not exceed that of the cyclic case, we require some additional notation. We will denote by $\zeta_{(j)}$ the jth order statistic of $\zeta_1, \ldots, \zeta_N$, that is, the jth smallest value among $\zeta_1, \ldots, \zeta_N$. Let d_i^{C} be the departure time of the ith customer after time zero in the cyclic case. The index of the service station at which the server completes a service at time d_i^{C} in the cyclic case is denoted by q_i^{C}. Note that $q_i^{\mathrm{C}} = ((i-1) \bmod N) + 1$ for $i > 0$. Furthermore, let $h_{i,j}^{\mathrm{C}}$ be the first moment after d_i^{C} that a customer at service station $((q_i^{\mathrm{C}} + j - 1) \bmod N) + 1$ has its preparation phase completed and is ready to be served by the server in the cyclic case, $j = 1, \ldots, N-1$.

With these definitions, we obviously have for the first departure that $d_1^{\mathrm{C}} = \zeta_1 + a_1$. Subsequent departures, which are marked by d_i^{C}, also occur exactly a_i time units after the server starts serving the ith customer. For $1 < i \leq N-1$ (thus, during the remainder of the first cycle), the start of the ith service occurs at time $\max\{d_{i-1}^{\mathrm{C}}, \zeta_i\}$, whereas for $i \geq N$ (corresponding to later cycles) the ith service is initiated at time $\max\{d_{i-1}^{\mathrm{C}}, h_{i-1,1}^{\mathrm{C}}\} = h_{i-1,1}^{\mathrm{C}}$. Therefore,

$$d_i^{\mathrm{C}} = \begin{cases} \zeta_1 + a_1 & \text{if } i = 1, \\ \max\{d_{i-1}^{\mathrm{C}}, \zeta_i\} + a_i & \text{if } 1 < i \leq N-1, \\ h_{i-1,1}^{\mathrm{C}} + a_i & \text{otherwise.} \end{cases} \tag{13}$$

As for the h values, we have for $i \leq N-1$ that the first point in time $h^{\mathrm{C}}_{i,1}$ after d^{C}_i that a customer at Q_{i+1} has its preparation phase completed obviously equals either d^{C}_i or ζ_{i+1} (whichever happens last). Hence, for $1 \leq i \leq N-1$,

$$h^{\mathrm{C}}_{i,1} = \max\{d^{\mathrm{C}}_i, \zeta_{i+1}\}. \tag{14}$$

For $i \geq N$, this expression is more involved. When the server has finished his $(i-1)$st service, a new preparation phase starts at the corresponding service station while the server moves to the next station. The newly started preparation phase ends at $d^{\mathrm{C}}_{i-1} + b_{i-1}$. It takes $N-1$ additional switches of the server before the customer corresponding to this preparation phase can be served. Hence, $h^{\mathrm{C}}_{i,N-1}$ takes the maximum value of this number and d^{C}_i. For other values of j, $h^{\mathrm{C}}_{i,j}$ retains the value $h^{\mathrm{C}}_{i-1,j+1}$ corresponding to the situation after the $(i-1)$st service, in case this value exceeds d^{C}_i. The shift in the second index is caused because the server has moved one position in the cycle to the next service station between the $(i-1)$st and the ith service. To summarize, we thus have for $i \geq N$ that

$$h^{\mathrm{C}}_{i,j} = \begin{cases} \max\{d^{\mathrm{C}}_i, h^{\mathrm{C}}_{i-1,j+1}\} & \text{if } j \neq N-1, \\ \max\{d^{\mathrm{C}}_i, d^{\mathrm{C}}_{i-1} + b_{i-1}\} & \text{if } j = N-1. \end{cases} \tag{15}$$

To finalize the notation, let d^{D}_i, q^{D}_i, and $h^{\mathrm{D}}_{i,j}$ be defined similarly to d^{C}_i, q^{C}_i, and $h^{\mathrm{C}}_{i,j}$ for the dynamic model, respectively. In the dynamic case, the server always moves to the service station with the earliest completed preparation phase. Evidently, we have that $d^{\mathrm{D}}_1 = \zeta_{(1)} + a_1$. For $1 < i \leq N-1$, the preparation phase of the ith served customer finishes before or at time $\zeta_{(i)}$. Therefore, we have for $1 < i \leq N-1$ that

$$d^{\mathrm{D}}_i \leq \max\{d^{\mathrm{D}}_{i-1}, \zeta_{(i)}\} + a_i. \tag{16}$$

For values of i larger than $N-1$, we have that the preparation phase the ith customer goes through has already finished or finishes exactly at time $\min_{j\in\{1,\ldots,N-1\}}\{h^{\mathrm{D}}_{i-1,j}\}$, provided that the $(i-1)$st customer was served at another station. Otherwise, it obviously finishes at time $d^{\mathrm{D}}_{i-1} + b_i$. Thus, for $i \geq N$, we have

$$d^{\mathrm{D}}_i = \min\left\{ \min_{j\in\{1,\ldots,N-1\}}\{h^{\mathrm{D}}_{i-1,j}\}, d^{\mathrm{D}}_{i-1} + b_i \right\} + a_i. \tag{17}$$

By the definition of $h^{\mathrm{D}}_{i,j}$, it is now not hard to see that for $1 \leq i \leq N-1$,

$$\min_{j\in\{1,\ldots,N-1\}}\{h^{\mathrm{D}}_{i,j}\} \leq \max\{d^{\mathrm{D}}_i, \zeta_{(i+1)}\}. \tag{18}$$

For values of i larger than $N-1$, one needs to keep careful track of the position of the server, but otherwise $h^{\mathrm{D}}_{i,j}$ is expressed similarly to (15). Namely, for $i \geq N$, we have that

$$h^{\mathrm{D}}_{i,j} = \begin{cases} \max\{d^{\mathrm{D}}_i, h_{i-1,j+((q^{\mathrm{D}}_i - q^{\mathrm{D}}_{i-1}) \bmod N)}\} & \text{if } j \neq N - ((q^{\mathrm{D}}_i - q^{\mathrm{D}}_{i-1}) \bmod N), \\ \max\{d^{\mathrm{D}}_i, d^{\mathrm{D}}_{i-1} + b_{i-1}\} & \text{if } j = N - ((q^{\mathrm{D}}_i - q^{\mathrm{D}}_{i-1}) \bmod N), \end{cases} \tag{19}$$

where $(q^{\mathrm{D}}_i - q^{\mathrm{D}}_{i-1}) \bmod N$ represents the shift in position of the server between time d^{D}_{i-1} and time d^{D}_i in the dynamic case.

Now that we have introduced all notation required, we perform two preliminary steps before proving the desired result. First, we show in Lemma 5.1 that $d^{\mathrm{C}}_i \geq d^{\mathrm{D}}_i$ for $i = 1, \ldots, N-1$. Thus, we first establish that $d^{\mathrm{C}}_i \geq d^{\mathrm{D}}_i$ for the special case of the first

cycle, at the start of which a preparation phase commences at each service point. Then, Lemma 5.2 shows that this inequality in fact also holds for $i \geq N$. In other words, the result $d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}$ persists after the first cycle. Based on these lemmas, Theorem 5.3 finally states that $\mathbb{E}[W^{\mathrm{C}}] \geq \mathbb{E}[W^{\mathrm{D}}]$ without any assumption on the distributions of the preparation and service times other than that both distributions have a non-negative support.

LEMMA 5.1: *For the first cycle, namely for $i = 1, \ldots, N-1$, we have that*

$$d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}} \text{ and } h_{i,1}^{\mathrm{C}} \geq \min_{j \in \{1,\ldots,N-1\}} \{h_{i,j}^{\mathrm{D}}\}.$$

PROOF: We first focus on the first part of the lemma and prove by induction that $d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}$ for $i = 1, \ldots, N-1$. We obviously have that

$$d_1^{\mathrm{C}} = \zeta_1 + a_1 \geq \zeta_{(1)} + a_1 = d_1^{\mathrm{D}},$$

which acts as a first step of the induction argument. We now show that $d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}$ for any $1 < i \leq N-1$ under the assumption that $d_k^{\mathrm{C}} \geq d_k^{\mathrm{D}}$ for all $k < i$. More specifically, we conclude based on (13) and (16) that

$$d_i^{\mathrm{C}} = \max\{d_{i-1}^{\mathrm{C}}, \zeta_i\} + a_i \geq \max\{d_{i-1}^{\mathrm{D}}, \zeta_{(i)}\} + a_i \geq d_i^{\mathrm{D}}$$

for any $1 < i \leq N-1$ by showing that each of the arguments of the second maximum operator does not exceed $\max\{d_{i-1}^{\mathrm{C}}, \zeta_i\}$. To see this for the first argument, note that

$$\max\{d_{i-1}^{\mathrm{C}}, \zeta_i\} \geq d_{i-1}^{\mathrm{C}} \geq d_{i-1}^{\mathrm{D}}$$

by the induction assumption. A similar observation for the second argument follows by noting that

$$\max\{d_{i-1}^{\mathrm{C}}, \zeta_i\} \geq \max\left\{\max_{j \in \{1,\ldots,i-1\}} \{\zeta_j\}, \zeta_i\right\} = \max_{j \in \{1,\ldots,i\}} \{\zeta_j\} \geq \zeta_{(i)}.$$

The first inequality holds since d_{i-1}^{C} must be larger than any of the times $\zeta_1, \ldots, \zeta_{i-1}$, as by time d_{i-1}^{C} the server has served one customer at the service stations $1, \ldots, i-1$ already in the cyclic case.

For the second part of the lemma, we observe based on (14) and (18) that for $i = 1, \ldots, N-1$,

$$h_{i,1}^{\mathrm{C}} = \max\{d_i^{\mathrm{C}}, \zeta_{i+1}\} \geq \max\{d_i^{\mathrm{D}}, \zeta_{(i+1)}\} \geq \min_{j \in \{1,\ldots,N-1\}} \{h_{i,j}^{\mathrm{D}}\}. \tag{20}$$

The first inequality follows by similar steps to those above. Namely, we obviously have that $\max\{d_i^{\mathrm{C}}, \zeta_{i+1}\} \geq d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}$ by the first part of the lemma already proved and that

$$\max\{d_i^{\mathrm{C}}, \zeta_{i+1}\} \geq \max\left\{\max_{j \in \{1,\ldots,i\}} \{\zeta_j\}, \zeta_{i+1}\right\} = \max_{j \in \{1,\ldots,i+1\}} \{\zeta_j\} \geq \zeta_{(i+1)}.$$

This concludes the proof. ■

We now generalize the result obtained in Lemma 5.1 and show that $d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}$ for all $i \geq 1$ in the following lemma.

LEMMA 5.2: *At every point in time, namely for every $i \geq 1$, we have that*

$$d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}} \text{ and } h_{i,1}^{\mathrm{C}} \geq \min_{j\in\{1,\ldots,N-1\}} \{h_{i,j}^{\mathrm{D}}\}.$$

PROOF: We have proved this statement already in Lemma 5.1 for $i = 1, \ldots, N-1$. To prove the result for larger i, we again apply induction, where Lemma 5.1 acts as a first step.

For the induction step, we now prove that $d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}$ and $h_{i,1}^{\mathrm{C}} \geq \min_{j\in\{1,\ldots,N-1\}}\{h_{i,j}^{\mathrm{D}}\}$ for all $i \geq N$ under the assumption that $d_k^{\mathrm{C}} \geq d_k^{\mathrm{D}}$ and $h_{k,1}^{\mathrm{C}} \geq \min_{j\in\{1,\ldots,N-1\}}\{h_{k,j}^{\mathrm{D}}\}$ for all $k < i$. The former statement $d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}$ is easily seen to hold true by observing based on (13) and (17) that

$$d_i^{\mathrm{C}} = h_{i-1,1}^{\mathrm{C}} + a_i \geq \min\left\{\min_{j\in\{1,\ldots,N-1\}}\{h_{i-1,j}^{\mathrm{D}}\}, d_{i-1}^{\mathrm{D}} + b_i\right\} + a_i = d_i^{\mathrm{D}}, \tag{21}$$

where the inequality holds since $h_{i-1,1}^{\mathrm{C}} \geq \min_{j\in\{1,\ldots,N-1\}}\{h_{i-1,j}^{\mathrm{D}}\}$ as per the induction assumption.

For the latter statement $h_{i,1}^{\mathrm{C}} \geq \min_{j\in\{1,\ldots,N-1\}}\{h_{i,j}^{\mathrm{D}}\}$, we derive from (15) that for the cyclic case

$$\begin{aligned} h_{i,1}^{\mathrm{C}} &= \max\{d_i^{\mathrm{C}}, h_{i-1,2}^{\mathrm{C}}\} = \max\{d_i^{\mathrm{C}}, h_{i-2,3}^{\mathrm{C}}\} \\ &= \cdots = \max\{d_i^{\mathrm{C}}, h_{i-N+2,N-1}^{\mathrm{C}}\} = \max\{d_i^{\mathrm{C}}, d_{i-N+1}^{\mathrm{C}} + b_{i-N+1}\}. \end{aligned} \tag{22}$$

Similarly, it can be derived from (19) that there exist $k, l \in \{1, \ldots, N-1\}$ so that $h_{i,k}^{\mathrm{D}} = \max\{d_i^{\mathrm{D}}, h_{i-N+2,l}^{\mathrm{D}}\}$. This leads to the inequality

$$\min_{j\in\{1,\ldots,N-1\}} h_{i,j}^{\mathrm{D}} \leq \max\left\{d_i^{\mathrm{D}}, \max_{j\in\{1,\ldots,N-1\}}\{h_{i-N+2,j}^{\mathrm{D}}\}\right\}. \tag{23}$$

We now proceed to show that $h_{i,1}^{\mathrm{C}} \geq \min_{j\in\{1,\ldots,N-1\}}\{h_{i,j}^{\mathrm{D}}\}$ by arguing that $h_{i,1}^{\mathrm{C}}$ is not smaller than each of the arguments in the outer maximum operator in the right-hand side of (23). For the first argument, we have by using (22) and (21), respectively, that

$$h_{i,1}^{\mathrm{C}} = \max\{d_i^{\mathrm{C}}, d_{i-N+1}^{\mathrm{C}} + b_{i-N+1}\} \geq d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}}.$$

To deal with the second argument of the maximum operator, we observe that by (19) $\max_{j\in\{1,\ldots,N-1\}}\{h_{i-N+2,j}^{\mathrm{D}}\}$ can evaluate either to (a) d_{i-N+2}^{D}, to (b) one of the values from the set $\{d_j^{\mathrm{D}} + b_j : j \in \{1, \ldots, i-N+1\}\}$ or to (c) $\zeta_{(N)}$. We treat each of these cases separately below.

(a) By (22) and (21), respectively, we have that

$$h_{i,1}^{\mathrm{C}} = \max\{d_i^{\mathrm{C}}, d_{i-N+1}^{\mathrm{C}} + b_{i-N+1}\} \geq d_i^{\mathrm{C}} \geq d_i^{\mathrm{D}} \geq d_{i-N+2}^{\mathrm{D}}.$$

(b) We show that $h_{i,1}^{\mathrm{C}}$ is not smaller than any value in the set $\{d_j^{\mathrm{D}} + b_j : j \in \{1, \ldots, i-N+1\}\}$. To this end, observe that $h_{k,1}^{\mathrm{C}} \geq h_{l,1}^{\mathrm{C}}$ for any $k \geq l$, since

$$h_{l,1}^{\mathrm{C}} \leq d_{l+1}^{\mathrm{C}} \leq d_k^{\mathrm{C}} \leq h_{k,1}^{\mathrm{C}}$$

for all $k > l$. For any $j \in \{1, \ldots, i-N+1\}$, it follows from (22), (21) and this observation that

$$h_{i,1}^{\mathrm{C}} \geq h_{j+N-1,1}^{\mathrm{C}} = \max\{d_{j+N-1}^{\mathrm{C}}, d_j^{\mathrm{C}} + b_j\} \geq d_j^{\mathrm{C}} + b_j \geq d_j^{\mathrm{D}} + b_j.$$

(c) By (22) and again the observation that in the cyclic case $h^{\mathrm{C}}_{k,1} \geq h^{\mathrm{C}}_{l,1}$ if $k \geq l$, we have that

$$h^{\mathrm{C}}_{i,1} \geq h^{\mathrm{C}}_{N,1} \geq d^{\mathrm{C}}_{N,1} \geq \zeta_{(N)},$$

where the first inequality again follows from the observation that $h^{\mathrm{C}}_{k,1} \geq h^{\mathrm{C}}_{l,1}$ if $k \geq l$. The second inequality follows from the fact that at time $d^{\mathrm{C}}_{N,1}$, the server has served exactly one customer at each of the service stations, and therefore $d^{\mathrm{C}}_{N,1}$ cannot be smaller than each of the initial residual preparation times $\zeta_1, \ldots, \zeta_N$.

By these observations, we have that $h^{\mathrm{C}}_{i,1} \geq \min_{j\in\{1,\ldots,N-1\}}\{h^{\mathrm{D}}_{i,j}\}$, which concludes the induction step. The lemma now follows by induction on i. ■

A combination of Lemmas 5.1 and 5.2 now leads to the following theorem.

THEOREM 5.3: *Given any two non-negative distributions for the service time A and the preparation time B, we have that $\mathbb{E}[W^{\mathrm{C}}] \geq \mathbb{E}[W^{\mathrm{D}}]$.*

PROOF: Given any two sets of i.i.d. sequences $\{a_i, i \geq 1\}$ and $\{b_i, i \geq 1\}$ from the random variables A and B, and any initial set of preparation times $(\zeta_1, \ldots, \zeta_N)$, Lemma 5.2 states that $d^{\mathrm{C}}_i \geq d^{\mathrm{D}}_i$ for all $i \geq 1$.

Observe that $d^{\mathrm{C}}_i = \sum_{j=1}^{i}(w^{\mathrm{C}}_j + a_j)$, where w^{C}_j is the time the server has to wait directly before the start of the jth service in the cyclic scenario. Likewise, we have that $d^{\mathrm{D}}_i = \sum_{j=1}^{i}(w^{\mathrm{D}}_j + a_j)$, where w^{D}_i is defined similarly to w^{C}_i for the dynamic scenario. Therefore, the lemma implies that for all $i > 0$,

$$\sum_{j=1}^{i}(w^{\mathrm{C}}_j + a_j) \geq \sum_{j=1}^{i}(w^{\mathrm{D}}_j + a_j), \tag{24}$$

which, after subtracting $\sum_{j=1}^{i} a_j$, dividing by i and taking limits on both sides, leads to

$$\lim_{i\to\infty} \frac{\sum_{j=1}^{i} w^{\mathrm{C}}_j}{i} \geq \lim_{i\to\infty} \frac{\sum_{j=1}^{i} w^{\mathrm{D}}_j}{i}.$$

The left-hand side (right-hand side) represents the asymptotic mean waiting time of the server in the cyclic (dynamic) scenario given the realizations $\{b_i, i \geq 1\}$, $\{a_i, i \geq 1\}$, and $(\zeta_1, \ldots, \zeta_N)$. Therefore, the theorem follows by conditioning on these realizations. ■

Remark 5.1: It is suggested by (24) that $\sum_{j=1}^{i} W^{\mathrm{C}}_j$ is stochastically larger than or equal to $\sum_{j=1}^{i} W^{\mathrm{D}}_j$ for all $i > 0$, where W^{C}_j (W^{D}_j) is the random variable representing the jth waiting time of the server in the cyclic (dynamic) case. Although there is not necessarily a stochastic ordering in the limiting distributions of the waiting times W^{C} and W^{D} (cf. Section 5.1), it thus appears that there exists a stochastic ordering in partial sums of transient waiting times starting at $j = 1$.

5.3. Numerical Comparison

In Section 4, we obtained several insights into the effect of the system parameters on its performance in the cyclic model. More specifically, we commented on the effect of variability of the preparation and service times, we observed the correlation structure of the waiting

times and we studied the number of stations to be assigned to a server. In this section, we compare the insights obtained for the cyclic model with equivalent observations for the dynamic model based on additional simulation results, and we explicitly comment on similarities and differences between the two models.

Variability of preparation and service times. We observed in Section 4 that the variability of the preparation time in the cyclic model seems to have a bigger impact on the server's waiting-time process than the variability of the service times. This observation does not extend to the dynamic case. Although the impact of the variability of the service times is similar, the variability of the preparation times hardly seems to matter for the waiting-time process. In Figure 6, we have plotted the counterpart of Figure 1 where the server now visits the service stations dynamically rather than cyclically. Thus, for the same variability settings considered before, we now plot the throughput θ^{D} versus the number of queues N.

It turns out that the solid curve and the dotted curve corresponding to moderately variable preparation times are similar to the ones corresponding to the cyclic model, other than the fact that these curves converge faster to the maximum throughput as expected. However, whereas the dashed curve corresponding to highly variable preparation times was the farthest away from the solid curve in Figure 1, the solid and dashed curves now almost coincide. This indicates that the variability of the preparation times hardly matters for the server's waiting time in the dynamic model.

This phenomenon can be explained by the fact that the dynamic model has many similarities with the Erlang loss model. In fact, if the service time A were exponentially distributed, the dynamic model would reduce to an $M/G/N/N$ queueing system. The service completions in the dynamic model are then equivalent to Poisson arrivals in the $M/G/N/N$ queue, of which the number of customers present represents the number of preparations in progress. A distinctive feature of the $M/G/N/N$ queue is that its performance measures are insensitive to the distribution of B apart from its first moment (see, e.g., [10]). Thus, if we would have chosen exponential service times in Figure 6, the solid and the dashed curves would have coincided. As this is not the case in our current example, the curves do not completely coincide, but the majority of the insensitivity remains.

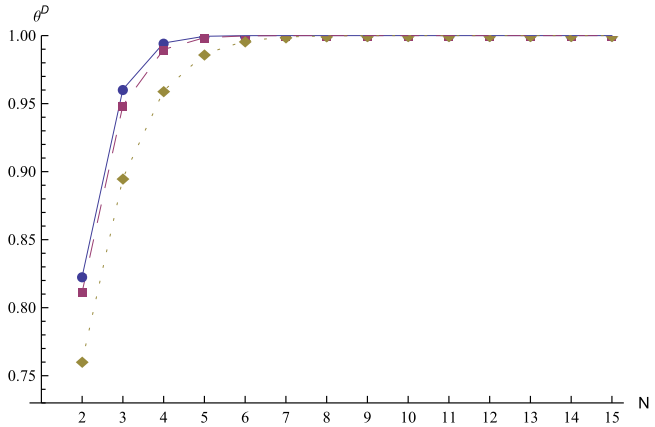

FIGURE **6.** (Color online) The throughput versus the number of stations for moderately variable preparation and service times (solid curve), highly variable service times (dotted curve) and highly variable preparation times (dashed curve) in the dynamic model.

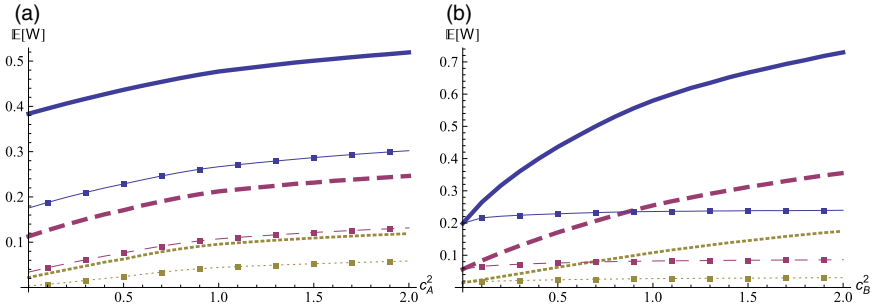

FIGURE **7.** (Color online) Waiting time versus c_A^2 (a) and c_B^2 (b) for the cyclic (thick line) and dynamic (thin and marked line) model with the values $r = 0.5$ (solid curve), $r = 0.8$ (dashed curve), and $= 1.2$ (dotted curve).

To further study the effects of the variability of the two time components, define the squared coefficient of variation $c_A^2 = \mathrm{Var}[A]/\mathbb{E}[A]^2$. Let c_B^2 be defined similarly and let $r = \mathbb{E}[B]/\mathbb{E}[A]$ represent the ratio of the two time components. Consider the systems with $N = 3$, $\mathbb{E}[A] = 1$, and the values $r = 0.5$, $r = 0.8$, and $r = 1.2$. Figures 7(a) and 7(b) plot the waiting time $\mathbb{E}[W]$ versus c_A^2 (keeping c_B^2 fixed at 1.5) and c_B^2 (keeping c_A^2 fixed at 1.5), respectively. In these two graphs, thick lines correspond to the cyclic case, whereas the thin, marked lines indicate results where the server visits the stations dynamically. From Figure 7(a), we conclude that as c_A^2 increases, the waiting time also increases for both cases, but that the rate of change is bigger in the cyclic case. The difference between a curve corresponding to the dynamic case and its equivalent for the cyclic case is, however, eventually almost constant and this difference increases as the value of r decreases. In Figure 7(b), we see that the mean waiting time in the cyclic model is more sensitive to c_B^2 than c_A^2 as observed before. However, for the dynamic system it is indeed almost insensitive to c_B^2. From these graphs, we conclude that the mean waiting time generally becomes larger for smaller values of r. This may strike as odd, since r is a measure of the workload offered to the server, while in most queueing models waiting times decrease as r decreases. However, recall that we study the waiting time of the server rather than that of the customers. Finally, we observe that in case $c_B^2 = 0$ (i.e., deterministic preparation times), the mean waiting times for the cyclic and the dynamic model coincide. Since deterministic preparation phases will always complete in the order they were initiated, the server will also serve the service points in a fixed cyclic order in the dynamic case, which leads to this behavior.

Correlations. In Section 4, we observed that the correlation structure of the waiting times behaves rather surprisingly for the cyclic model. The correlation structure in the dynamic model not only turns out to behave as unexpectedly, but it also behaves very differently from the cyclic case. In Figure 8, we plot the correlation structure of the dynamic model based on the same system settings as those used to construct Figure 3, namely exponentially (1) distributed preparation times, exponentially (10) distributed service times and $N = 5$. However, apart from the exponential case $c_B^2 = 1$, we now also observe the correlation structure for the values $c_B^2 = 0.5$ and $c_B^2 = 10$. In the cyclic case, increasing the value of c_B^2 does not alter the shape of the curve depicted in Figure 3, although the correlation generally becomes less significant. Figures 8(a)–8(c) show not only that in the dynamic case the correlation becomes more significant and converges to zero slower as c_B^2 increases, but also that the shape of the curve is sensitive to c_B^2. Figures 8(a) and 8(b) clearly show that also in the dynamic model periodicity effects are resent, as alternatingly convex and alternating

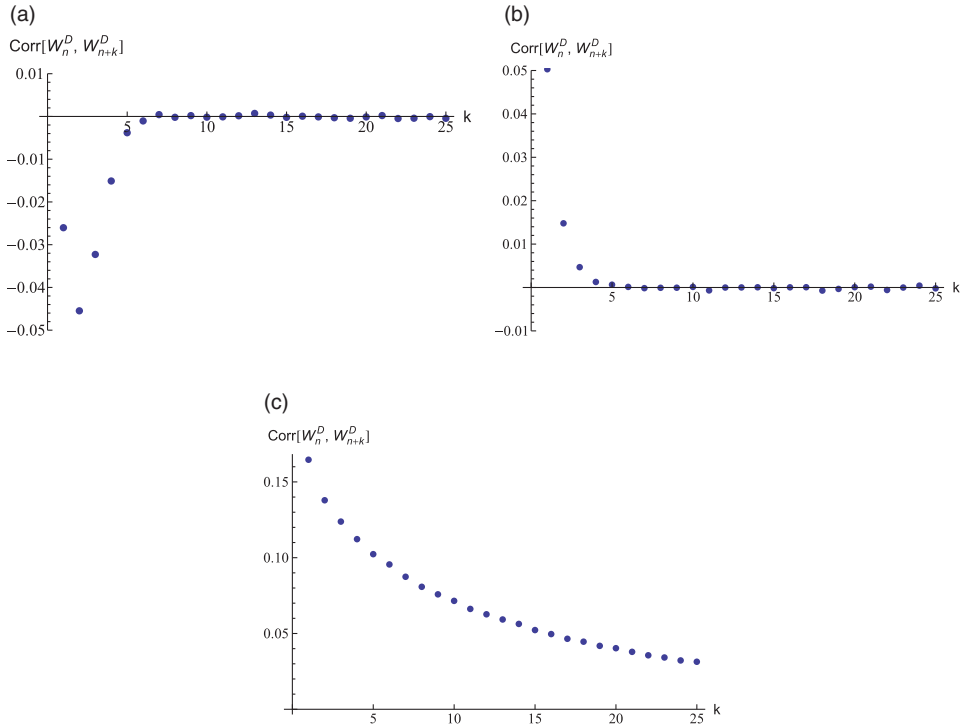

FIGURE **8.** (Color online) Correlation structure for $c_B^2 = 0.5$ (a) $c_B^2 = 1$ (b), and $c_B^2 = 10$ (c).

loops can be observed. However, an increasing c_B^2 also seems to have a significant effect on the correlation itself. For $c_B^2 = 0.5$, the correlation is negative for small k, whereas this is not the case for $c_B^2 = 1.0$. For $c_B^2 = 10$, even Figure 8(c) shows a monotonously decreasing curve.

It is not clear why these effects are present nor why the behavior for the dynamic model is so much different from the cyclic case. The increased sensitivity to the variability of the preparation times to the correlation of the waiting times in the dynamic model is highly surprising, as we observed that the waiting-time distribution itself in the dynamic case is hardly sensitive to c_B^2. Such peculiar behavior is also present for the variability of the service time, but in an opposite fashion. Whereas the waiting-time distribution is sensitive to c_A^2 in the dynamic case (cf. Figure 7(a)), numerical results show that this number has little effect on the correlation curves as depicted in Figure 8.

Number of stations to be assigned to a server. We now study how the number of stations to be assigned to a server changes when one switches from a cyclic to a dynamic regime. In Figure 9, we plot the same curves as those depicted in Figure 4, and we add the curves one would obtain when the server visits the service stations dynamically. This figure shows intuitive results. Obviously, the throughput θ^{D} for the dynamic model is larger than its equivalent θ^{C} for the cyclic model. This is not surprising, since we found in Section 5.2 that $\mathbb{E}[W^{\mathrm{C}}] \geq \mathbb{E}[W^{\mathrm{D}}]$. As such, the number of stations to be assigned to a service in order to be close to maximum throughput decreases. Whereas we concluded before that about 5 or 6 generally are needed for the cyclic case, it seems that for the dynamic case about 3 to 4 servers is already enough.

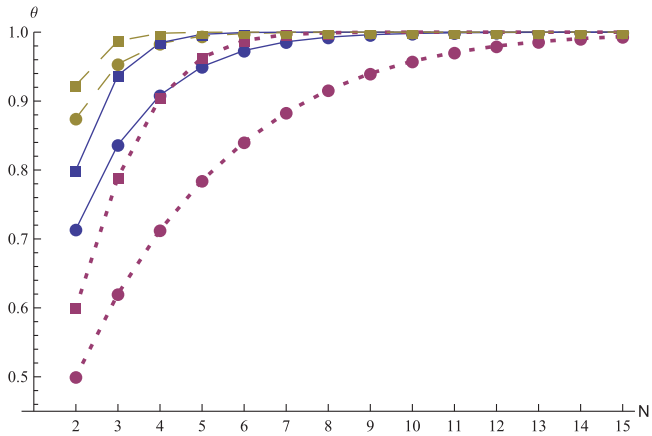

FIGURE **9.** (Color online) Throughput versus the number of stations for small (dashed curve), moderate (solid curve), and large preparation times (dotted curve) for the cyclic (circles) and the dynamic model (squares).

Acknowledgements

The authors thank Onno Boxma for helpful comments on earlier drafts of the present paper. This research was funded in the framework of the STAR-project "Multilayered queueing systems" by the Netherlands Organization for Scientific Research (NWO), while the first author was a Ph.D. student at the Eindhoven University of Technology and the Centrum Wiskunde & Informatica (CWI). The research of the third author was also partly supported by an NWO individual grant through project 632.003.002.

References

1. Asmussen, S. (2003). *Applied Probability and Queues*. New York: Springer.
2. Bertsekas, D. & Gallager, R. (1992). *Data Networks*. Englewood Cliffs: Prentice-Hall.
3. Bingham, N.H., Goldie, C.M. & Teugels, J.L. (1989). *Regular Variation*. Cambridge: Cambridge University Press.
4. Boon, M.A.A., van der Mei, R.D. & Winands, E.M.M. (2011). Applications of polling systems. *Surveys in Operations Research and Management Science* 16: 67–82.
5. Boxma, O.J. & Groenendijk, W.P. (1988). Two queues with alternating service and switching times. In *Queueing Theory and its Applications (Liber Amicorum for J. W. Cohen)* O.J. Boxma & R. Syski, (eds.), Amsterdam: North-Holland, pp. 261–282.
6. Cline, D.B.H. & Samorodnitsky, G. (1994). Subexponentiality of the product of independent random variables. *Stochastic Processes and their Applications* 49: 75–98.
7. Eisenberg, M. (1979). Two queues with alternating service. *SIAM Journal on Applied Mathematics* 36: 287–303.
8. Franks, R.G., Al-Omari, T., Woodside, C.M., Das, O. & Derisavi, S. (2009). Enhanced modeling and solution of layered queueing networks. *IEEE Transactions on Software Engineering* 35: 148–161.
9. Gamarnik, D., Nowicki, T. & 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: 76–106.
10. Kelly, F.P. (1979). *Reversibility and Stochastic Networks*. Chichester: Wiley.
11. Kleinrock, L. (1976). *Queueing Systems, Volume II: Computer Applications*. New York: Wiley.
12. Litvak, N. & Vlasiou, M. (2010). A survey on performance analysis of warehouse carousel systems. *Statistica Neerlandica* 64: 401–447.
13. McGinnis, L.F., Han, M.H. & White, J.A. (1986). Analysis of rotary rack operations. In *Proceedings of the 7th International Conference on Automation in Warehousing*, pp. 165–171.
14. Park, B.C., Park, J.Y. & Foley, R.D. (2003). Carousel system performance. *Journal of Applied Probability* 40: 602–612.

15. Perel, N., Dorsman, J.L. & Vlasiou, M. (2013). Cyclic-type polling models with preparation times. In *Proceedings of the 2nd International Conference on Operations Research and Enterprise Systems*, pp. 14–23.
16. Resing, J.A.C. (1993). Polling systems and multitype branching processes. *Queueing Systems* 13: 409–426.
17. Szekli, R. (1995). *Stochastic Ordering and Dependence in Applied Probability*. New York: Springer.
18. Takács, L. (1962). *Introduction to the Theory of Queues*. New York: Oxford University Press.
19. Takagi, H. (1986). *Analysis of Polling Systems*. Cambridge: MIT Press.
20. Tijms, H.C. (1994). *Stochastic Models: An Algorithmic Approach*. Chichester: John Wiley & Sons.
21. van Vuuren, M. & Winands, E.M.M. (2007). Iterative approximation of k-limited polling systems. *Queueing Systems* 55: 161–178.
22. Vlasiou, M. (2006). *Lindley-Type Recursions*. Ph.D. thesis, Eindhoven University of Technology, Eindhoven, The Netherlands.
23. Vlasiou, M. (2007). A non-increasing Lindley-type equation. *Queueing Systems* 56: 41–52.
24. Vlasiou, M. & Adan, I.J.B.F. (2005). An alternating service problem. *Probability in the Engineering and Informational Sciences* 19: 409–426.
25. Vlasiou, M. & Zwart, B. (2007). Time-dependent behaviour of an alternating service queue. *Stochastic Models* 23: 235–263.
26. Yechiali, U. (1993). Analysis and control of polling systems. In L. Donatiello & R. Nelson, editors, *Performance Evaluation of Computer and Communication Systems*, vol. 729, Berlin/Heidelberg: Springer, pp. 630–650.