EURANDOM PREPRINT SERIES

2015-008

April, 2015

Stochastic bounds for order flow times in warehouses with remotely located order-picking workstations

D. Claeys, I. Adan, O. Boxma
ISSN 1389-2355

Stochastic bounds for order flow times in warehouses with remotely located order-picking workstations

Dieter Claeys[*], Ivo Adan[†], Onno Boxma[‡]

Abstract

This paper focuses on the mathematical analysis of order flow times in parts-to-picker warehouses with remotely located order-picking workstations. To this end, a polling system with a new type of arrival process and service discipline is introduced as a model for an order-picking workstation. Stochastic bounds are deduced for the cycle time, which corresponds to the order flow time. These bounds are shown to be adequate and aid in setting targets for the throughput of the storage area. The paper thus complements existing literature, which mainly focuses on optimizing the operations in the storage area.

Keywords: warehousing, order-picking workstation, order flow time, polling system, cycle time, stochastic bounds

1 Introduction

Warehouses currently endure various challenges [1]. In addition to tighter delivery schedules being requested, the number of stock keeping units (SKUs) has also shot up, a challenge coined SKU proliferation [2]. Moreover, the order profiles have changed. E-commerce, for instance, has led to significantly more order requests, albeit of smaller size, i.e., requiring fewer products.

Order picking is one of the most important aspects in warehousing. It is the process of retrieving products from the warehouse to satisfy customer orders. Other warehouse operations can be categorized into receiving, putting away, storing and shipping. As compared to these, order picking is much more expensive [3], highlighting the importance of understanding this process thoroughly when designing warehouses.

In this paper, we focus on an automated end-of-aisle parts-to-picker order-picking system with miniloads and remotely located manual workstations [4, 5] (see Figure 1). Such systems consist of a storage area, a closed-loop conveyor and order-picking workstations. The storage

[*]SMACS Research Group, Department TELIN, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium, Dieter.Claeys@telin.ugent.be

[†]Department of Mechanical Engineering, Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands, i.j.b.f.adan@tue.nl

[‡]EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, o.j.boxma@tue.nl

area is a collection of miniloads, each being composed of one crane and two parallel racks at either side of the aisle. The closed-loop conveyor deals with transport between the storage area and the order-picking workstations, which are each operated by an order picker. An automated end-of-aisle parts-to-picker order-picking system with miniloads and remotely located workstations has many advantages, such as being able to retrieve products from several orders simultaneously, being robust and providing a solution to SKU proliferation [1].

A customer order consists of one or more order lines, each describing the required quantity of a specific product. For instance, an order can comprise two DDR4 SDRAM modules, one particular CPU and one specific motherboard. The number of order lines is called the order size and equals three in the example. When an order is released into the system, the cranes in the aisle where the products are stored retrieve the corresponding totes (also called bins), each containing a single SKU, and transfer them to the end of the aisle where they are put onto the closed-loop conveyor for transport to one of the workstations. At the workstation, the order picker collects the required number of products from the totes and puts them in a so-called order tote. The totes with the remaining products are placed back onto the closed-loop conveyor and are subsequently returned to the storage area.

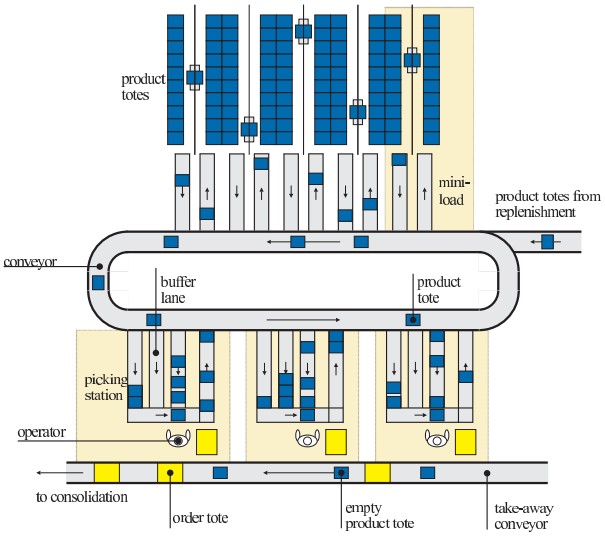


Figure 1: An automated end-of-aisle parts-to-picker order-picking system with miniloads and remotely located manual workstations [6]

Literature on end-of-aisle parts-to-picker order-picking systems mainly focuses on operations in the storage area (also called automatic storage/retrieval system (AS/RS)), especially on reducing travel times of cranes. Finding an optimal combination between number of racks, their length and height [3] is an example of this. Storage assignment strategies have also been investigated for that purpose, as well as batching strategies, where cranes retrieve several totes simultaneously. The optimal position of an idle crane has also been examined, a topic coined dwell-point strategies. A comprehensive overview of research on storage area operations is provided by Roodbergen and Vis [5].

While the storage area has been studied extensively, little attention has been paid to the

design and performance analysis of order-picking workstations. Andriansyah et al. [7] developed a new conceptual design for an automated order-picking workstation including a carrousel to overcome deadlock problems. They also introduced several picking policies for the robot and evaluated these policies through simulation experiments, revealing that 'smart' policies can increase the throughput significantly. In Andriansyah et al. [6], the performance of a manual order-picking workstation is assessed through simulation, using the concept Effective Process Time (EPT), which aggregates all individual components making up the process (picking) time; EPT simplifies the modeling aspect while yielding accurate results. To the best of our knowledge, no mathematical analysis has been carried out of a model for a remotely located workstation, a gap which this paper aims to fill. An analytical approach aids in providing quick fundamental insights into the performance, which is particularly useful during the design phase of warehouses.

The major contribution of this paper is the analytic study of an order-picking workstation. A queueing (polling) model is developed (Section 2) that strikes a balance between capturing the essential features of an order-picking workstation while allowing to deduce stochastic bounds for the order flow time (Section 3), which are shown to be accurate for a reasonable range of mean arrival (retrieval) rates (Section 4). Furthermore, guidelines and insights are provided that are useful to practitioners. For instance, the bounds are valuable in setting targets for the desired retrieval rate of totes from the storage area.

2 Model description

This section first discusses the operation of an order-picking workstation (Section 2.1) and then explains how an order-picking workstation is modelled as a polling system (Section 2.2). The model is mainly based on that of Andriansyah et al. [6], except for some small adaptations that allow mathematical analysis without compromising the applicability of the model.

2.1 Order-picking workstation operation

At an order-picking workstation (see Figure 2), totes arrive randomly; totes do not necessarily arrive in the sequence of release times of orders, i.e., some totes destined for order $i+j$ might arrive before some totes of order i. Arriving totes are put at the tail of one of the buffer lanes. The exact location of a tote - buffer and position within the buffer - is immaterial since the routing of totes to the buffer lanes is such that whenever a tote is needed, it is within reach of the picker: this is why several (short) buffers are provided instead of one (long) buffer.

In order to restrict the number of picking errors, the order picker handles one order at a time. This means that when the order picker is processing an order with order size N (stochastic), he consecutively picks from each of the N associated totes the required number of products and puts them into the active order tote. When the order picker has dealt with all order lines of an order, he carries out some completion work, which accounts for packing, putting the order tote on another conveyor, scanning the barcodes of the new order, et cetera [6]. Afterwards, the order picker starts picking another order and a new order is released into the system, meaning

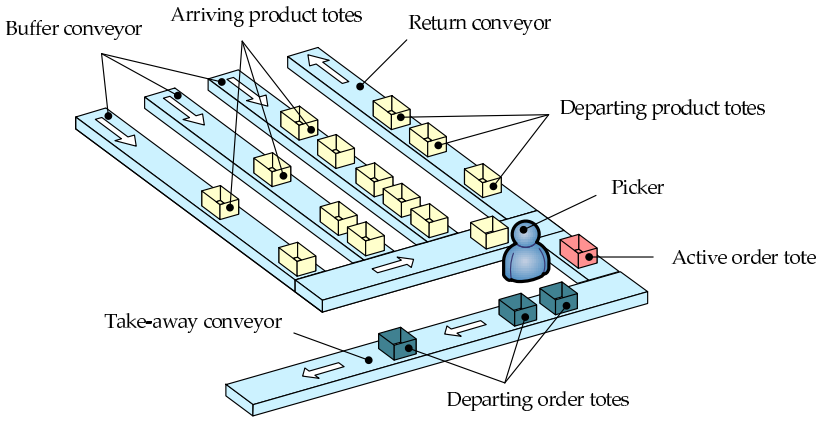


Figure 2: Operation of an order-picking workstation [6]

that totes destined for this new order start arriving. At most K orders can simultaneously be active, i.e., released into the system and not yet completed; K is called the line capacity. This is to avoid deadlocks of the closed-loop conveyor due to full buffer lanes. It is possible that the order picker is handling an order while none of the remaining totes for that order are available yet at the workstation. In this case, the picker waits until one of these totes has arrived, even if totes for other orders are already present. Although this policy might be somewhat inefficient, its simplicity leads to fewer picking errors.

2.2 Polling model of workstation

An order-picking workstation is modelled as a polling system (see Figure 3), a queueing system with a single server that cyclically serves K queues, serving customers at those queues according to a particular service discipline (see e.g. [8, 9]):

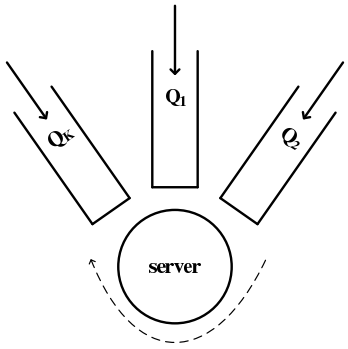


Figure 3: Polling model for an order-picking workstation

- The single server corresponds to the order picker, and customers represent totes containing a single SKU.

- K queues $(Q_1, \ldots, Q_K)$ correspond to K consecutive orders. Hence, a queue in the polling system does not necessarily model a buffer lane in the order-picking workstation. In fact, in the polling system, arriving totes are sorted based on the released customer orders and placed accordingly into virtual queues $Q_1, \ldots, Q_K$. A tote is placed in a specific queue, say queue i, simply because its SKU is requested by the $jK + i$-th released order into the system for some $j \geq 0$. This fictitious ordering, however, does not affect the behaviour of the workstation, due to the picking policy and because the order picker can reach and thus pick each tote whenever needed.
- In each queue, totes arrive randomly. As soon as all N totes of an order have arrived, the arrival process at the associated queue is turned off. When the server leaves that queue, this arrival process is immediately switched on again, reflecting that a new order is released into the order-picking system. The latter means that we assume that always new orders are available to be released into the system. In other words, we focus on the maximum throughput capability of the order-picking workstation. Note that the on-off feature of the arrival processes does not occur in the model of Andriansyah et al. [6]. However, the combination of K queues and the particular on-off arrival process captures that at most K orders can simultaneously be active, avoiding deadlocks of the closed-loop conveyor.
- When the server arrives at a queue, he will serve exactly as many customers as the order size N before proceeding to the next queue: when the queue becomes empty before N customers have been served, the server becomes idle and waits for more customers to arrive, even if other queues are not empty. Serving a customer thus corresponds to handling an order line, i.e., picking the required number of products from a tote, and visiting a queue represents carrying out a complete order.
- After having finished the last order line of an order, the server moves on to the next queue where it will process another order. This requires a switching time, constituting the completion time of the order.

The results in this paper hold under the following assumptions:

- As long as not all N totes of an order have arrived, its associated totes arrive according to a Poisson process with parameter λ, referred to as arrival rate, or retrieval rate. This assumption is reasonable since totes come from random locations in a large warehouse processing many orders simultaneously.
- The service times of consecutive customers are independent and identically distributed (IID), with generic random variable B, density $b(\cdot)$, and Laplace-Stieltjes transform (LST) $\tilde{B}(\cdot)$. This corresponds to the concept EPT [6], incorporating all individual processes such as raw pick time, picker availability and picking errors.
- The consecutive switching times are IID with common random variable C, density $c(\cdot)$, and LST $\tilde{C}(\cdot)$.
- The consecutive order sizes are IID with common random variable N; N has a finite support of size $\mathrm{M} < \infty$, i.e., its probability generating function (PGF) is a polynomial:

$N(z) = \sum_{w=1}^{\mathrm{M}} \mathbb{P}[N = w]\, z^w$. This is a natural assumption in the context of e-commerce, where order sizes are generally small. In addition, large orders are split into smaller suborders in practice.

- As all processes and distributions are independent of the queue number, the queues are symmetric. This is realistic, since totes from a new order are placed in a specific (virtual) queue, say queue i, simply because it is the $jK + i$-th order released into the system (for some $j \geq 0$).

In contrast with the mainstream papers on polling systems, the objective of this paper is to examine the steady-state delay D of an order, coined order flow time, and not the delay of a random tote (customer). The order flow time D is a natural performance measure for the order-picking process under investigation. As a new order is released into the system when another order has been completed, the order flow time is defined as the time between two consecutive server departures from the same queue. Hence, it corresponds to the *cycle time* in polling systems.

Remark 1. *The order flow time is different from that studied by Andriansyah et al. who defined it as the time between the arrival of the first tote in the workstation and the departure of the last product tote of the order [6]. Note that it is premature to state that the order flow time is the sum of the order flow time defined by Andriansyah et al. [6] plus an independent exponentially distributed time, exactly because the arrival process can be switched off in the present paper to avoid blocking of the closed-loop conveyor.*

Remark 2. *The order flow time is strongly related to the maximum throughput of an order-picking workstation, which is defined as the line capacity K divided by the mean cycle time $\mathbb{E}[D]$. Hence, a lower (upper) bound for the mean order flow time immediately yields an upper (lower) bound for the maximum throughput.*

Remark 3. *The policy of serving exactly N customers during a visit has recently also been studied by Boxma et al. [10]. The model in this paper differs in two aspects. First, Boxma et al. [10] assume that the length of the intervisit period is independent of the length of the preceding visit period. Secondly, the on-off feature of the arrival process is not incorporated by Boxma et al. The papers also differ in the obtained performance measures; whereas Boxma et al. deduce the PGF of the system occupancy at various time epochs and the LST of the delay of a customer, this paper is devoted to the delay of a random order, i.e., the order flow time.*

3 Analysis of the polling system

The polling system under investigation differs from traditional polling systems, not only because the arrival process can be switched off, but also because the service discipline is peculiar: during a visit of a queue, the server processes exactly N customers, even if that entails that the server is idle for some time while other queues are not empty. Unfortunately, this service discipline does not satisfy the so-called branching property, making it very unlikely that the joint queue length distribution can be determined exactly [11, 12]. Note that this service discipline is similar to

the k-limited service policy, where the server continues serving a particular queue until either a predefined number of k customers is served or until the queue becomes empty, whichever occurs first. Our service discipline can be considered as an N-limited service policy with the extra feature that the server serves *exactly* N customers and N is random. Polling models with k-limited services also do not satisfy the branching property, making them notoriously hard to analyse; only in a few special two-queue cases an exact analysis has turned out to be feasible, see e.g. [13, 14, 15, 16, 17].

Although the polling model under investigation does not satisfy the branching property, expressions for the cycle time will be established, representing the flow time of an order. Due to the symmetry of the queues, the distribution of the cycle time does not depend on its starting point, allowing us, without loss of generality, to study cycles that start when the server ends visiting Q_1.

We first relate cycle and visit time to intervisit time (Section 3.1) and after that, we compute the distribution of the intervisit time, both for the two-queue case (Section 3.2) and for the case of an arbitrary number of queues (Section 3.3). As will become clear, the approach for the former case cannot be applied to an arbitrary number of queues, exactly because the polling system is of non-branching type. Therefore, stochastic bounds will be established in Section 3.3.

3.1 Relation between cycle, visit and intervisit time

Focussing on Q_1, we define:

- V_n: n-th visit time, the length of the period between the server arriving at Q_1 and the server leaving Q_1, switching (completion) time included.
- I_n: n-th intervisit time, i.e., length of period between end of n-th visit and start of $n+1$-th visit of Q_1.
- D_n: n-th cycle time, i.e., $D_n \triangleq I_{n-1} + V_n$, representing the flow time of the n-th order processed at Q_1.

These definitions immediately lead to

$$\begin{aligned}\tilde{D}(s) &\triangleq \lim_{n\to\infty} \mathbb{E}\left[e^{-sD_n}\right] = \lim_{n\to\infty} \mathbb{E}\left[e^{-s(I_{n-1}+V_n)}\right] \\ &= \lim_{n\to\infty} \int_0^\infty e^{-sx}\mathbb{E}\left[e^{-sV_n}|I_{n-1}=x\right] \mathrm{d}\mathbb{P}[I_{n-1}\le x] \quad . \end{aligned} \tag{1}$$

We now deduce an expression for $\mathbb{E}\left[e^{-sV_n}|I_{n-1}=x\right]$. Define X_n as the number of customers present at Q_1 when the server arrives at Q_1 for the n-th time. Due to the particular on-off arrival process, this is also equal to the number of arrivals at Q_1 during the $n-1$-th intervisit period. As I_{n-1} influences V_n only through X_n, we obtain, after defining N_n as the size of the n-th order at Q_1,

$$\mathbb{E}\left[e^{-sV_n}|I_{n-1}=x\right] = \sum_{w=1}^{\mathrm{M}} \mathbb{P}[N_n=w] \sum_{k=0}^{w-1} e^{-\lambda x}\frac{(\lambda x)^k}{k!}\mathbb{E}\left[e^{-sV_n}|X_n=k, N_n=w\right]$$

$$+\sum_{w=1}^{\mathrm{M}}\mathbb{P}[N_n=w]\left[1-\sum_{k=0}^{w-1}e^{-\lambda x}\frac{(\lambda x)^k}{k!}\right]\mathbb{E}\left[e^{-sV_n}|X_n=N_n=w\right]\ . \quad (2)$$

Expression (2) takes into account that $\mathbb{P}[X_n = N_n|I_{n-1} = x]$ equals the probability of a Poisson process with rate λ generating at least N_n arrivals in an interval of length x. Obviously,

$$\mathbb{E}\left[e^{-sV_n}|X_n=N_n=w\right]=\tilde{B}(s)^w\tilde{C}(s)\ . \quad (3)$$

On the other hand, when $X_n < N_n$, the server might become idle during the visit. In order to deal with this complicating behaviour, the visit time is divided into four parts:

$$V_n=T_n+W_n+B_n+C_n\ , \quad (4)$$

whereby T_n is the length of the period starting at the instant the server arrives for the n-th time at Q_1 until all N_n customers have arrived, W_n is the waiting time of that N_n-th customer, B_n is the service time of that customer and C_n is the switching time at the end of the n-th visit of Q_1. Note that B_n and C_n are mutually independent and also independent of T_n and W_n. Hence, translating (4) into LSTs and conditioning on X_n and N_n yields

$$\mathbb{E}\left[e^{-sV_n}|X_n=k,N_n=w\right]=\tilde{B}(s)\tilde{C}(s)\mathbb{E}\left[e^{-s(T_n+W_n)}|X_n=k,N_n=w\right]\ . \quad (5)$$

To find the last factor at the RHS of (5), we examine the closed cyclic queueing network with 2 queues ($\hat{Q}_1$ and $\hat{Q}_2$) presented in Figure 4. Service times in $\hat{Q}_1$ are generally distributed with

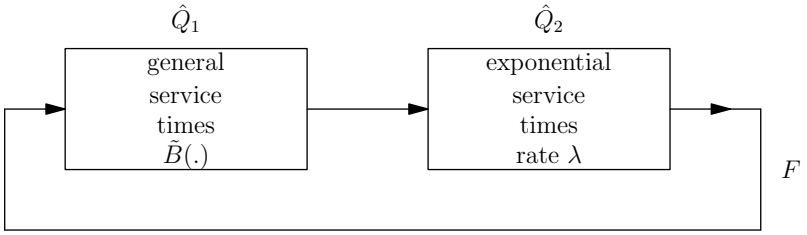


Figure 4: The cyclic queueing network system with F circulating customers

LST $\tilde{B}(\cdot)$ and service times in $\hat{Q}_2$ are exponentially distributed with rate λ. The closed network contains F customers. Formula (2.22) of [18] provides an expression for

$$\mathbb{E}\left[e^{-sA_m}e^{-\theta\hat{W}_m}|Z_0=k\right]\ ,$$

with A_m the time until the m-th arrival at $\hat{Q}_1$ since a randomly chosen service completion epoch in $\hat{Q}_1$, denoted by τ, $\hat{W}_m$ the waiting time (i.e., time spent in $\hat{Q}_1$, service time excluded) of the m-th customer arriving at $\hat{Q}_1$ after epoch τ and Z_0 the number of customers present at $\hat{Q}_1$ immediately after epoch τ. Although the cyclic queueing network is different from the polling system, the following observations are essential to determine $\mathbb{E}\left[e^{-s(T_n+W_n)}|X_n=k,N_n=w\right]$:

- Taking epoch τ as time origin, as long as the $(F - Z_0)$-th customer has not yet arrived at $\hat{Q}_1$, $\hat{Q}_1$ is fed by a Poisson arrival process with rate λ. In the polling system, as long as the N_n-th customer has not yet arrived, Q_1 is fed by a Poisson arrival process with rate λ.
- In the polling system, the arrival process is turned off upon the arrival of the N_n-th customer, whereas this is not necessarily the case in the cyclic model. However, due to the FCFS scheduling discipline, the arrival process after the arrival of the N_n-th customer does not affect the waiting time of that customer.
- A service completion at $\hat{Q}_1$ means that the service of an other customer in $\hat{Q}_1$ can commence. In the polling system, when the server arrives at Q_1, it can initiate service of a customer in that queue. Hence, in both systems, service initiation opportunities arise.

As a consequence, the polling and the cyclic system can be related as follows:

$$\mathbb{E}\left[e^{-s(T_n+W_n)}|X_n = k, N_n = w\right] = \mathbb{E}\left[e^{-s(A_{w-k}+\hat{W}_{w-k})}|Z_0 = k\right] \quad , \quad 0 \leq k \leq w-1 \ . \tag{6}$$

The combination of (5) and (6) with Formula (2.22) from [18] produces

$$\mathbb{E}\left[e^{-sV_n}|X_n = k, N_n = w\right] = \tilde{B}(s)^w \tilde{C}(s) \Bigg[1 - \frac{s}{\lambda+s} \sum_{j=0}^{w-k-1} \tilde{B}(s)^{-k-j} \sum_{i=0}^{j} \left(\frac{\lambda}{\lambda+s}\right)^i \frac{k+i}{k+j} \int_0^\infty e^{-(s+\lambda)t} \frac{(\lambda t)^{j-i}}{(j-i)!} \mathrm{d}B^{(k+j)*}(t)\Bigg] \quad , \quad k < w \ , \tag{7}$$

where $B^{(k+j)*}$ denotes the $(k+j)$-fold convolution of B. For $k = j = 0$, the sum over i is equal to 1 by convention. Throughout this section, we tacitly define each sum over i equal to 1 if $k = j = 0$. Substituting (3) and (7) into (2) yields:

$$\begin{aligned}\mathbb{E}\left[e^{-sV_n}|I_{n-1} = x\right] =& N(\tilde{B}(s))\tilde{C}(s) \\ &- \frac{s}{\lambda+s}\tilde{C}(s) \sum_{k=0}^{\mathrm{M}-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!} \sum_{j=0}^{\mathrm{M}-k-1} \tilde{B}(s)^{-k-j} \left\{ \sum_{w=k+1+j}^{\mathrm{M}} \mathbb{P}[N = w]\, \tilde{B}(s)^w \right\} \\ &\sum_{i=0}^{j} \left(\frac{\lambda}{\lambda+s}\right)^i \frac{k+i}{k+j} \int_0^\infty e^{-(s+\lambda)t} \frac{(\lambda t)^{j-i}}{(j-i)!} \mathrm{d}B^{(k+j)*}(t) \ . \end{aligned} \tag{8}$$

Remark 4. *In case of exponentially distributed service times, i.e., $\tilde{B}(s) = \mu/(\mu+s)$, the integral in (8) can be evaluated explicitly:*

$$\int_0^\infty e^{-(s+\lambda)t} \frac{(\lambda t)^{j-i}}{(j-i)!} dB^{(k+j)*}(t) = \binom{2j+k-i-1}{j-i} \frac{\mu^{j+k}\lambda^{j-i}}{(s+\lambda+\mu)^{2j+k-i}} \ .$$

The combination of (1) and (8) and letting $n \to \infty$ leads to the LST of the steady-state cycle time D:

$$\tilde{D}(s) = N(\tilde{B}(s))\tilde{C}(s)\tilde{I}(s)$$

$$- \frac{s}{\lambda+s}\tilde{C}(s)\sum_{k=0}^{\mathrm{M}-1}\int_0^\infty e^{-(s+\lambda)x}\frac{(\lambda x)^k}{k!}\mathrm{d}\mathbb{P}[I\le x]$$
$$\sum_{j=0}^{\mathrm{M}-k-1}\tilde{B}(s)^{-k-j}\left\{\sum_{w=k+1+j}^{\mathrm{M}}\mathbb{P}[N=w]\,\tilde{B}(s)^w\right\}$$
$$\sum_{i=0}^{j}\left(\frac{\lambda}{\lambda+s}\right)^i\frac{k+i}{k+j}\int_0^\infty e^{-(s+\lambda)t}\frac{(\lambda t)^{j-i}}{(j-i)!}\mathrm{d}B^{(k+j)*}(t)\ , \tag{9}$$

where I denotes the length of a random intervisit period of Q_1, with corresponding LST $\tilde{I}(\cdot)$. The mean cycle time is obtained from (9) by applying the moment generating property of LSTs:

$$\begin{aligned}\mathbb{E}[D] =&\mathbb{E}[N]\,\mathbb{E}[B]+\mathbb{E}[C]+\mathbb{E}[I]\\ &+\frac{1}{\lambda}\sum_{k=0}^{\mathrm{M}-1}\int_0^\infty e^{-\lambda x}\frac{(\lambda x)^k}{k!}\mathrm{d}\mathbb{P}[I\le x]\sum_{j=0}^{\mathrm{M}-k-1}\mathbb{P}[N\ge k+1+j]\\ &\sum_{i=0}^{j}\frac{k+i}{k+j}\int_0^\infty e^{-\lambda t}\frac{(\lambda t)^{j-i}}{(j-i)!}\mathrm{d}B^{(k+j)*}(t)\ .\end{aligned} \tag{10}$$

The calculation of both integrals in (10) can be avoided by observing that the first integral represents the probability that k customers arrive during a random intervisit period and the second corresponds to the probability that $j-i$ customers arrive during $k+j$ consecutive service periods. As a result, (10) can be rewritten in terms of finite sums of derivatives:

$$\begin{aligned}\mathbb{E}[D] =&\mathbb{E}[N]\,\mathbb{E}[B]+\mathbb{E}[C]+\mathbb{E}[I]\\ &+\frac{1}{\lambda}\sum_{k=0}^{\mathrm{M}-1}\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}z^k}\tilde{I}(\lambda(1-z))\Big|_{z=0}\sum_{j=k}^{\mathrm{M}-1}\mathbb{P}[N\ge j+1]\sum_{i=0}^{j-k}\frac{j-i}{j}\frac{1}{i!}\frac{\mathrm{d}^i}{\mathrm{d}z^i}\tilde{B}^j(\lambda(1-z))\Big|_{z=0}\ .\end{aligned} \tag{11}$$

The first three terms in (11) correspond to the parts of the cycle time that always occur: N service periods, a switching and an intervisit period. The last term represents the mean idle time during a visit: the average time $1/\lambda$ that an idle server has to wait for a customer to arrive multiplied by the average number of times the server becomes idle during a visit.
Note that from (11) (or from (10)), it can be observed that

$$\mathbb{E}[D]\sim\mathbb{E}[N]\,\mathbb{E}[B]+\mathbb{E}[C]+\mathbb{E}[I]\ ,$$

as $\lambda\to\infty$, expressing that the server is never idle when customers arrive sufficiently fast.
For use later on, note that, once the distribution of I_{n-1} is known, the LST of the visit time can be deduced from (8):

$$\tilde{V}(s)\triangleq\lim_{n\to\infty}\mathbb{E}\left[e^{-sV_n}\right]=\int_0^\infty\lim_{n\to\infty}\mathbb{E}\left[e^{-sV_n}|I_{n-1}=x\right]\mathrm{d}\mathbb{P}[I_{n-1}\le x]\ . \tag{12}$$

Substitution of (8) into (12), interpreting integrals as probabilities, and manipulating summation indices gives:

$$\begin{aligned}\tilde{V}(s) =& N(\tilde{B}(s))\tilde{C}(s) \\ & - \frac{s}{\lambda+s}\tilde{C}(s)\sum_{k=0}^{\mathrm{M}-1}\frac{1}{k!}\left.\frac{\mathrm{d}^k}{\mathrm{d}z^k}\tilde{I}(\lambda(1-z))\right|_{z=0} \\ & \sum_{j=k}^{\mathrm{M}-1}\tilde{B}(s)^{-j}\left\{\sum_{w=j+1}^{\mathrm{M}}\mathbb{P}[N=w]\,\tilde{B}(s)^w\right\}\left(\frac{\lambda}{\lambda+s}\right)^{j-k} \\ & \sum_{i=0}^{j-k}\frac{j-i}{j}\frac{1}{i!}\left.\frac{\mathrm{d}^i}{\mathrm{d}z^i}\tilde{B}^j((s+\lambda)(1-z))\right|_{z=0} \quad . \end{aligned} \tag{13}$$

3.2 Distribution of intervisit time in case of two queues

Now that the cycle time (and also the visit time) has been expressed in terms of the intervisit time, the remaining challenge is to determine the distribution of the intervisit time. To set our mind, we first consider the case of two queues (i.e., $K = 2$) and then continue with an arbitrary number of queues in Section 3.3. Define:

- $V_n^{(k)}$: length of n-th visit period of Q_k.
- $I_n^{(k)}$: length of n-th intervisit period of Q_k.
- $N_n^{(k)}$: size of n-th order at Q_k.

Crucial is the observation that the visit of one queue coincides with the intervisit of the other queue:

$$I_n^{(1)} = V_n^{(2)} \quad , \tag{14}$$

and

$$I_n^{(2)} = V_{n+1}^{(1)} \quad . \tag{15}$$

This leads to the following iterative algorithm: start with an initial assessment of the distribution of $I_0^{(1)}$ (e.g., assume that $I_0^{(1)}$ equals the sum of N IID service times and 1 switching time); then use (13) to assess the distribution of $V_1^{(1)}$. Due to (15), this distribution equals the distribution of $I_0^{(2)}$. Using (13), the distribution of $V_1^{(2)}$ is obtained and thus, owing to (14), the distribution of $I_1^{(1)}$. Keep on iterating until the distributions converge.

We do not investigate whether this algorithm converges, because in this two queue case, there is a better alternative to find the cycle time: $\lim_{n\to\infty}\mathbb{P}\left[X_n^{(k)}=i, N_n^{(k)}=w\right]$ can be determined, with $X_n^{(k)}$ the number of customers present in Q_k at the beginning of the n-th visit of Q_k. Given $X_n^{(k)}$ and $N_n^{(k)}$, $V_n^{(k)}$ is deduced via (3) if $X_n^{(k)} = N_n^{(k)}$ and via (7) if $X_n^{(k)} < N_n^{(k)}$. In addition,

$$\mathbb{P}\big[X_n^{(2)}=i_n^{(2)}, N_n^{(2)}=w_n^{(2)} | V_n^{(1)}=t\big] = \begin{cases} \mathbb{P}\left[N=w_n^{(2)}\right] e^{-\lambda t}\frac{(\lambda t)^{i_n^{(2)}}}{i_n^{(2)}!} & \text{if } i_n^{(2)} < w_n^{(2)} \\ \mathbb{P}\left[N=w_n^{(2)}\right]\left[1-\sum_{k=0}^{w_n^{(2)}-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\right] & \text{if } i_n^{(2)} = w_n^{(2)} \end{cases}$$

and $\mathbb{P}\Big[X_{n+1}^{(1)} = i_{n+1}^{(1)}, N_{n+1}^{(1)} = w_{n+1}^{(1)} | V_n^{(2)} = t\Big]$ can be established analogously. Hence, by applying the law of total probability, the one-step transition probabilities of the Markov chains $\{(X_n^{(k)}, N_n^{(k)}); n \geq 1\}$ can be computed. As these chains are irreducible, aperiodic and positive recurrent due to the finite state space, the one-step transition probabilities can be used, in combination with the normalization condition, to compute $\lim_{n\to\infty} \mathbb{P}\Big[(X_n^{(k)}, N_n^{(k)}) = (i, w)\Big]$.

Finally, the cycle time is obtained from these joint distributions:

$$\begin{aligned}
\tilde{D}(s) =& \sum_{w=1}^{\mathrm{M}} \sum_{i=0}^{w} \lim_{n\to\infty} \mathbb{P}\Big[X_{n-1}^{(2)} = i, N_{n-1}^{(2)} = w\Big] \, \mathbb{E}\left[e^{-s(I_{n-1}^{(1)} + V_n^{(1)})} | X_{n-1}^{(2)} = i, N_{n-1}^{(2)} = w\right] \\
=& \sum_{w=1}^{\mathrm{M}} \sum_{i=0}^{w} \lim_{n\to\infty} \mathbb{P}\Big[X_{n-1}^{(2)} = i, N_{n-1}^{(2)} = w\Big] \\
& \int_0^\infty e^{-sx} \lim_{n\to\infty} \mathbb{E}\left[e^{-sV_n^{(1)}} | I_{n-1}^{(1)} = x\right] \mathrm{d}\mathbb{P}\Big[V_{n-1}^{(2)} \leq x | X_{n-1}^{(2)} = i, N_{n-1}^{(2)} = w\Big] \quad ,
\end{aligned}$$

where the expectation is given by (8) and the last probability is found by inverting expression (3) numerically if $i = w$ or expression (7) in the other case. We do not provide further details, as we have now set our mind to move on to an arbitrary number of queues.

3.3 Stochastic bounds for intervisit time and cycle time

In order to find the distribution of the intervisit time of Q_1 in case of an arbitrary number K of queues, we should, as in the two-queue case, for each Q_j, $j = 2, \ldots, K$, compute the distribution of the number of customers present when the server starts visiting that queue. However, this is infeasible for polling systems in which the service policy does not satisfy the branching property. Therefore, in the remainder, we establish stochastic bounds for the intervisit time and consequently, for the cycle time.

The following two relations are our starting point:

$$I_n^{(1)} = V_n^{(2)} + V_n^{(3)} + \ldots + V_n^{(K)} \quad , \tag{16}$$

$$V_n^{(i)} \geq \sum_{k=1}^{N_n^{(i)}} B_n^{(k,i)} + C_n^{(i)} \quad , \qquad i = 1, \ldots, K \quad , \tag{17}$$

with $B_n^{(k,i)}$ the service time of the k-th customer during the n-th visit of Q_i and $C_n^{(i)}$ the switching time at the end of the n-th visit of Q_i. Relation (16) expresses that an intervisit of Q_1 consists of consecutive visits of $Q_2, \ldots, Q_K$ and inequality (17) formulates that any visit of a queue comprises at least N services and a switching period. Equality occurs in (17) if and only if the server never becomes idle during the n-th visit of Q_i. Combination of (16) and (17) yields

$$I_n^{(1)} \geq \underline{I}_n^{(1)} \triangleq \sum_{i=2}^{K} \left[\sum_{k=1}^{N_n^{(i)}} B_n^{(k,i)} + C_n^{(i)} \right] \quad , \tag{18}$$

(henceforth, lower and upper bounds of a random variable X will be denoted by $\underline{X}$ and $\overline{X}$ respectively).

The next step is to translate (18) into a stochastic bound for the visit time of Q_1 and the cycle time. Therefore, we introduce a new system which is identical to the system under investigation, with the only exception that each intervisit time $I_n^{(1)}$ of Q_1 is replaced by its smaller counterpart $\underline{I}_n^{(1)}$. For this new system, the length of the n-th visit period of Q_1 is denoted by $\overline{V}_n^{(1)}$ and the n-th cycle time by $\underline{D}_n$. Theorem 1 gives the following stochastic bounds:

Theorem 1.

$$\mathbb{P}\big[V_n^{(1)} \geq x\big] \leq \mathbb{P}\Big[\overline{V}_n^{(1)} \geq x\Big] \quad , \quad n \geq 1 \ . \tag{19}$$

$$\mathbb{P}[D_n \geq x] \geq \mathbb{P}[\underline{D}_n \geq x] \quad , \quad n \geq 1 \ . \tag{20}$$

Proof. Let $P_n^{(j,i)}$ be the j-th customer served during the n-th visit of Q_i and define $A_n^{(j,i)}$ as the interarrival time of $P_n^{(j,i)}$. The interarrival time should be defined carefully due to the on-off feature of the arrival process; for $j \geq 2$, $A_n^{(j,i)}$ is the time between the arrival epochs of $P_n^{(j-1,i)}$ and $P_n^{(j,i)}$, whereas $A_n^{(1,i)}$ is the time between the start of the $n-1$-th intervisit of Q_i and the arrival epoch of $P_n^{(1,i)}$.
Our proof is based on classical coupling arguments, see e.g. Chapter 9 of [19]. We show that $V_n^{(1)} \leq \overline{V}_n^{(1)}$ and $D_n \geq \underline{D}_n \ \forall n \geq 1$, for all realisations of input sequences $\{N_n^{(1)}\}$, $\{A_n^{(j,1)}\}$, $\{B_n^{(j,1)}\}$, $\{C_n^{(1)}\}$, $\{I_n^{(1)}\}$ and $\{\underline{I}_n^{(1)}\}$, yielding (19) and (20). Note that the actual input sequences are in fact $\{N_n^{(i)}\}$, $\{A_n^{(j,i)}\}$, $\{B_n^{(j,i)}\}$, $\{C_n^{(i)}\}$ and that the above input sequences can be expressed in terms of the actual input sequences, while this is not necessarily possible in the other direction. However, for the cycle times and visit times of Q_1, the input sequences $\{N_n^{(1)}\}$, $\{A_n^{(j,1)}\}$, $\{B_n^{(j,1)}\}$, $\{C_n^{(1)}\}$, $\{I_n^{(1)}\}$ and $\{\underline{I}_n^{(1)}\}$ provide enough information. As the new system only differs from the original through the intervisit times $I_n^{(1)}$ being replaced by their smaller counterpart $\underline{I}_n^{(1)}$, we have that $\{N_n^{(1)}\}$, $\{A_n^{(j,1)}\}$, $\{B_n^{(j,1)}\}$ and $\{C_n^{(1)}\}$ are identical in both systems.

The n-th visit period of Q_1 can be divided into $N_n^{(1)}$ services, 1 switching period and possibly some idle periods, producing

$$V_n^{(1)} = \sum_{j=1}^{N_n^{(1)}} B_n^{(j,1)} + C_n^{(1)} + G_n^{(1)} \quad ,$$

with $G_n^{(1)}$ the total idle time during the n-th visit of Q_1. For fixed $N_n^{(1)}, A_n^{(1,1)}, \ldots, A_n^{(N_n^{(1)},1)}$ and $B_n^{(1,1)}, \ldots, B_n^{(N_n^{(1)}-1,1)}$, $G_n^{(1)}$ is a non-increasing function of $I_{n-1}^{(1)}$. This holds for all input sequences, proving (19).

To prove (20), we relate the n-th cycle time with the input sequences and with $W_n^{(N_n^{(1)})}$, whereby $W_n^{(j)}$ represents the waiting time of $P_n^{(j,1)}$ (see Figure 5 for an illustration):

$$D_n = \sum_{j=1}^{N_n^{(1)}} A_n^{(j,1)} + W_n^{(N_n^{(1)})} + B_n^{(N_n^{(1)},1)} + C_n^{(1)} \quad . \tag{21}$$

$W_n^{(N_n^{(1)})}$ can be calculated recursively via the following Lindley equations:

$$W_n^{(j)} = \left(W_n^{(j-1)} + B_n^{(j-1,1)} - A_n^{(j,1)}\right)^+ \quad , \quad 2 \le j \le N_n^{(1)} \quad , \tag{22}$$

$$W_n^{(1)} = \left(I_{n-1}^{(1)} - A_n^{(1,1)}\right)^+ \quad , \tag{23}$$

with $(\cdot)^+ \triangleq \max(\cdot, 0)$. From (23), we observe that $W_n^{(1)} \ge \underline{W}_n^{(1)}$, with $\underline{W}_n^{(j)}$ the waiting time of $P_n^{(j,1)}$ in the new system. Consequently, (22) yields $W_n^{(N_n^{(1)})} \ge \underline{W}_n^{(N_n^{(1)})}$, which, in combination with (21), proves (20). □

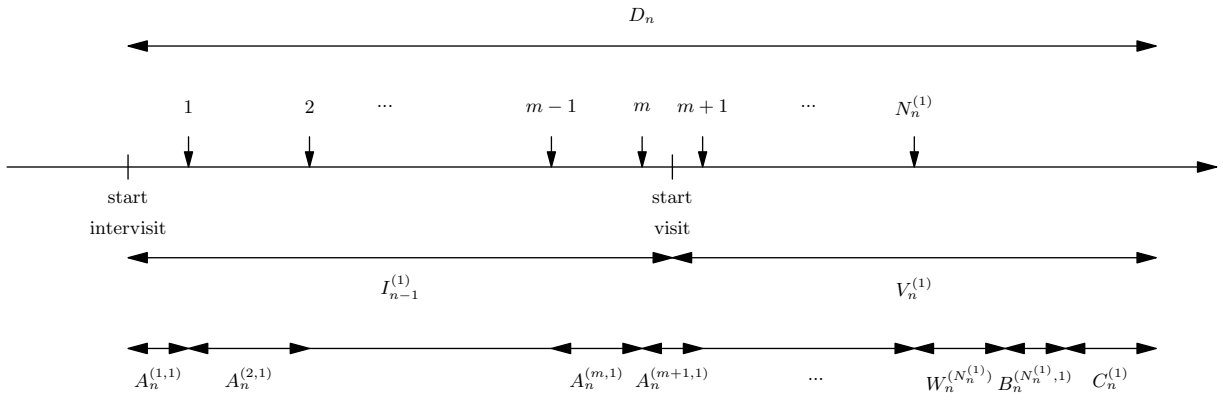


Figure 5: Illustration of relation (21)

As (20) holds for all n, we have obtained a stochastic *lower bound* for a random cycle time. This lower bound is computed by replacing $\tilde{I}(s)$ by $\left[N(\tilde{B}(s))\tilde{C}(s)\right]^{K-1}$ in expression (9) for $\tilde{D}(s)$, which boils down to assuming that the server never becomes idle during an intervisit.

The stochastic lower bound can immediately be translated into a lower bound for the mean value:

$$\mathbb{E}\left[D\right] = \int_0^\infty \mathbb{P}[D \ge x]\,\mathrm{d}x \ge \int_0^\infty \mathbb{P}[\underline{D} \ge x]\,\mathrm{d}x = \mathbb{E}\left[\underline{D}\right] \quad ,$$

where $\underline{D} \triangleq \lim_{n\to\infty} \underline{D}_n$. Hence, substitution of $\tilde{I}(s)$ by $[N(\tilde{B}(s))\tilde{C}(s)]^{K-1}$ in (11) yields

$$\begin{aligned}\mathbb{E}\left[D\right] \ge & \mathbb{E}\left[N\right] K \mathbb{E}\left[B\right] + K\mathbb{E}\left[C\right] \\ & + \frac{1}{\lambda}\sum_{k=0}^{\mathrm{M}-1} \frac{1}{k!}\,\frac{\mathrm{d}^k}{\mathrm{d}z^k}\left[N(\tilde{B}(\lambda(1-z)))\tilde{C}(\lambda(1-z))\right]^{K-1}\Bigg|_{z=0} \\ & \sum_{j=k}^{\mathrm{M}-1} \mathbb{P}[j+1 \le N \le \mathrm{M}] \sum_{i=0}^{j-k} \frac{j-i}{j}\frac{1}{i!}\,\frac{\mathrm{d}^i}{\mathrm{d}z^i}\tilde{B}^j(\lambda(1-z))\Bigg|_{z=0} \quad .\end{aligned} \tag{24}$$

In the remainder, we deduce an *upper bound*. The key idea is to apply the previous reasoning to the visit times of $Q_2, \ldots, Q_K$, i.e., assuming for each Q_i, $i = 2, \ldots, K$, that when the server arrives, it did not have to wait during the preceding intervisit. Application of Theorem 1 then produces

$$\mathbb{P}\big[V_n^{(i)} \geq x\big] \leq \mathbb{P}\Big[\overline{V}_n^{(i)} \geq x\Big] \quad .$$

Repeating this for every queue i, it can be proved by using the same coupling argument as in Theorem 1 that

$$\mathbb{P}\big[I_n^{(1)} \geq x\big] = \mathbb{P}\big[V_n^{(2)} + \ldots + V_n^{(K)} \geq x\big] \leq \mathbb{P}\Big[\overline{V}_n^{(2)} + \ldots + \overline{V}_n^{(K)} \geq x\Big] \quad . \tag{25}$$

and that replacing $I_n^{(1)}$ by $\overline{V}_n^{(2)} + \ldots + \overline{V}_n^{(K)}$ yields an upper bound for D_n.

However, upper bound (25) is difficult to compute, because the $\overline{V}_n^{(i)}$'s are dependent: $\overline{V}_n^{(i)}$ and $\overline{V}_n^{(i+1)}$ both depend on the order sizes, service and completion times during the $(n-1)$-th visit of $Q_{i+2}, \ldots, Q_K$ and the n-th visit of $Q_1, \ldots Q_i$. The upper bound is only easily computable in case of two queues and in case of deterministic order sizes in combination with deterministic service and completion times. We thus seek for another upper bound, that does not suffer from this issue.

We start from the following lower bound for $X_n^{(i)}$:

$$X_n^{(i)} \geq \underline{X}_n^{(i)} \quad ,$$

where $\underline{X}_n^{(i)}$ represents the number of customers that arrived in Q_i during the service periods and switching period of the $(n-1)$-th visit of Q_{i+1} (thus, arrivals during possibly idle periods during the visit period of Q_{i+1} and during visits of queues other than Q_{i+1} are excluded) for $i = 2, \ldots, K-1$; $\underline{X}_n^{(K)}$ represents the number of customers that have accumulated in Q_K during the service periods and switching period during the n-th visit of Q_1. Crucial here is that $\underline{X}_n^{(2)}, \ldots, \underline{X}_n^{(K)}$ are independent.

By using the same coupling argument as in Theorem 1, it can be proved that

$$\mathbb{P}[D_n \geq x] \leq \mathbb{P}\big[\overline{D}_n \geq x\big] \quad ,$$

with $\overline{D}_n$ the cycle time whose LST is obtained by replacing $\tilde{I}(s)$ by $\overline{V}(s)^{K-1}$ in (9), whereby $\overline{V}(s)$ is found by replacing $\tilde{I}(s)$ in Formula (13) for $\tilde{V}(s)$ by $N(\tilde{B}(s))\tilde{C}(s)$.

This stochastic upper bound yields an upper bound for the mean cycle time:

$$\begin{aligned}
\mathbb{E}\left[D\right] \leq & \mathbb{E}\left[N\right]\mathbb{E}\left[B\right] + \mathbb{E}\left[C\right] + (K-1)\mathbb{E}\left[\overline{V}\right] \\
& + \frac{1}{\lambda}\sum_{k=0}^{\mathrm{M}-1} \frac{1}{k!}\,\frac{\mathrm{d}^k}{\mathrm{d}z^k}\overline{V}^{K-1}(\lambda(1-z))\Big|_{z=0} \sum_{j=k}^{\mathrm{M}-1} \mathbb{P}[j+1 \leq N \leq \mathrm{M}] \\
& \sum_{i=0}^{j-k} \frac{j-i}{j}\frac{1}{i!}\,\frac{\mathrm{d}^i}{\mathrm{d}z^i}\tilde{B}^j(\lambda(1-z))\Big|_{z=0} \quad ,
\end{aligned} \tag{26}$$

with (cf. (13))

$$\begin{aligned}\overline{V}(s) =& N(\tilde{B}(s))\tilde{C}(s) \\ &- \frac{s}{\lambda+s}\tilde{C}(s)\sum_{k=0}^{\mathrm{M}-1}\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}z^k}N(\tilde{B}(\lambda(1-z)))\tilde{C}(\lambda(1-z))\Big|_{z=0} \\ &\sum_{j=k}^{\mathrm{M}-1}\tilde{B}(s)^{-j}\left\{\sum_{w=j+1}^{\mathrm{M}}\mathbb{P}[N=w]\,\tilde{B}(s)^w\right\}\left(\frac{\lambda}{\lambda+s}\right)^{j-k} \\ &\sum_{i=0}^{j-k}\frac{j-i}{j}\frac{1}{i!}\frac{\mathrm{d}^i}{\mathrm{d}z^i}\tilde{B}^j((s+\lambda)(1-z))\Big|_{z=0}\Bigg] \ . \end{aligned} \tag{27}$$

Expressions (26)-(27) are easy to implement as they contain only finite sums, and neither integrals nor convolutions. Note that, in the two-queue case, upper bound (26) is the same as the upper bound obtained by applying upper bound (25) for $I_n^{(1)}$.

Remark 5. *Since the upper bound follows from excluding all possible arrivals (in Q_i) during $K-2$ visits (of $Q_{i+2},\ldots,Q_K$ and $Q_1,\ldots,Q_{i-1}$) to avoid complicating dependencies, it is likely to become inaccurate for large K. This will be confirmed and discussed in Section 4. When the picking and completion times have strictly positive lower bounds, i.e.,*

$$b_{\min} \triangleq \inf\{x\in\mathbb{R} : b(x)>0\} > 0 \ ,$$

$$c_{\min} \triangleq \inf\{x\in\mathbb{R} : c(x)>0\} > 0 \ ,$$

a modified, more accurate, upper bound can be developed for $K>2$, by including during each of the $K-2$ visits possible arrivals during the first $b_{\min}$ time units of the picking time of the first order line and the first $c_{\min}$ time units of the completion time. Hence, the more accurate approximation is found by replacing $N(\tilde{B}(\lambda(1-z)))\tilde{C}(\lambda(1-z))$ by $N(\tilde{B}(\lambda(1-z)))\tilde{C}(\lambda(1-z))e^{\lambda(b_{\min}+c_{\min})(z-1)(K-2)}$ in (27).

4 Evaluation of bounds and discussion

In this section, we examine the accuracy of the bounds for the order flow time and provide some insights and guidelines that are useful to practitioners. The bounds as well as simulated values of the actual flow times are depicted versus the retrieval rate λ for three special cases: exponential picking (service) and completion (switching) times (Figure 6), i.e.,

$$\tilde{B}(s) = \frac{\mu}{\mu+s} \ , \quad \mathbb{E}[B] = \frac{1}{\mu} \ ,$$

$$\tilde{C}(s) = \frac{\alpha}{\alpha+s} \ , \quad \mathbb{E}[C] = \frac{1}{\alpha} \ ,$$

deterministic picking and completion times (Figure 7), i.e.,

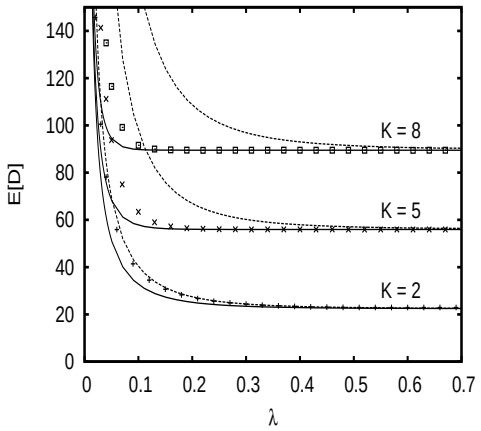
$$\tilde{B}(s) = e^{-s\beta} \ , \quad \mathbb{E}[B] = \beta \ ,$$

$$\tilde{C}(s) = e^{-s\zeta} \ , \quad \mathbb{E}[C] = \zeta \ ,$$

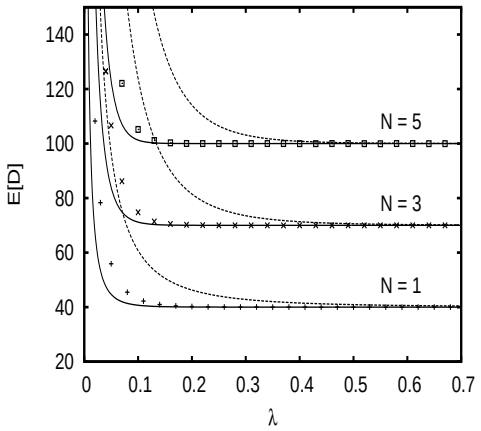
and shifted exponential picking and completion times (Figure 8):

$$\tilde{B}(s) = e^{-s\beta}\frac{\mu}{\mu+s} \ , \quad \mathbb{E}[B] = \beta + \frac{1}{\mu} \ ,$$

$$\tilde{C}(s) = e^{-s\zeta}\frac{\alpha}{\alpha+s} \ , \quad \mathbb{E}[C] = \zeta + \frac{1}{\alpha} \ .$$



(a) Various values of line capacity K; N stochastic with mass function (28)



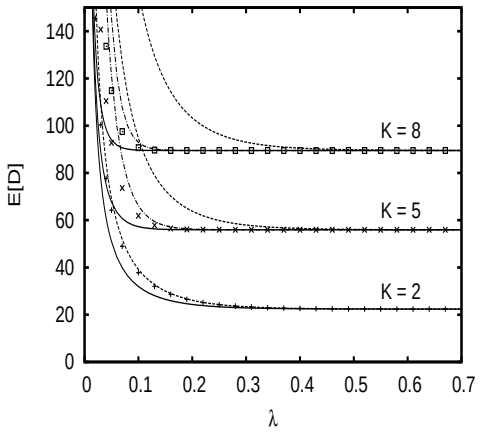
(b) Various values of (deterministic) order size N; $K = 5$

Figure 6: Lower (solid lines) and upper (dashed lines) bounds as well as simulated values (dots) of the average flow time $\mathbb{E}[D]$ versus retrieval rate λ; exponential picking times ($\mu = 1/3$), exponential completion times ($\alpha = 1/5$)

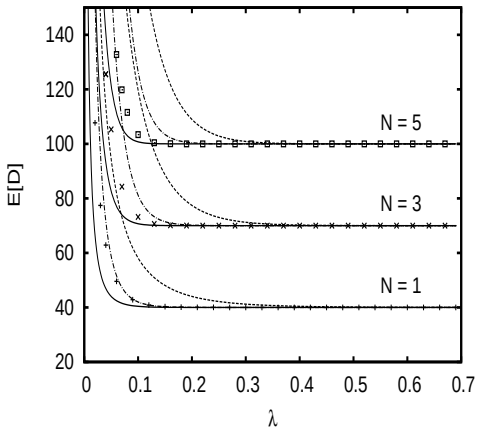
In the left part of Figures 6-8, curves are depicted for several values of the line capacity K, while the distribution of the order size N is

$$\mathbb{P}[N = w] = \begin{cases} 1/2 & \text{if } w = 1 \\ 3/16 & \text{if } w = 2 \\ 2/16 & \text{if } w = 3 \\ 2/16 & \text{if } w = 4 \\ 1/16 & \text{if } w = 5 \end{cases} \tag{28}$$

On the other hand, K is kept constant equal to 5 in the right parts, whereas different deterministic order sizes N are considered. In case of deterministic service and completion times

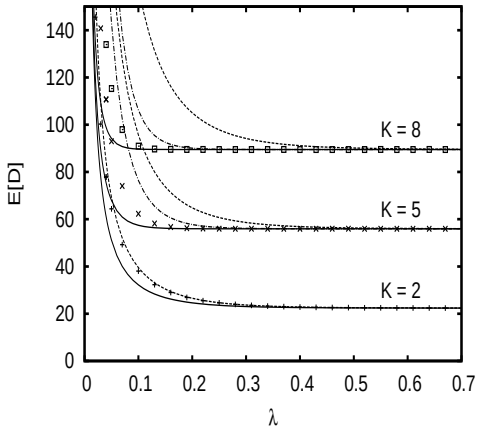


(a) Various values of line capacity K; N stochastic with mass function (28)

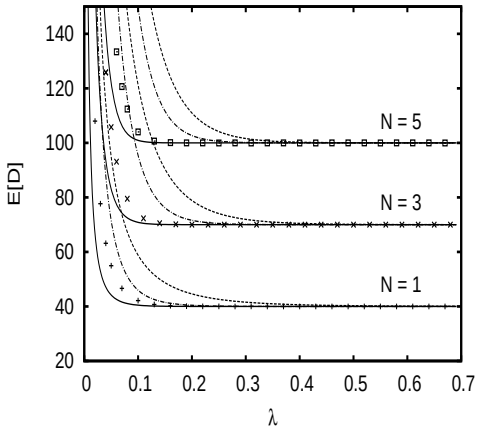


(b) Various values of (deterministic) order size N; $K = 5$

Figure 7: Lower (solid lines), upper (dashed lines), and modified upper (dot-dashed lines) bounds as well as simulated values (dots) of the average flow time $\mathbb{E}[D]$ versus retrieval rate λ; deterministic picking times ($\beta = 3$), deterministic completion times ($\zeta = 5$)



(a) Various values of line capacity K; N stochastic with mass function (28)

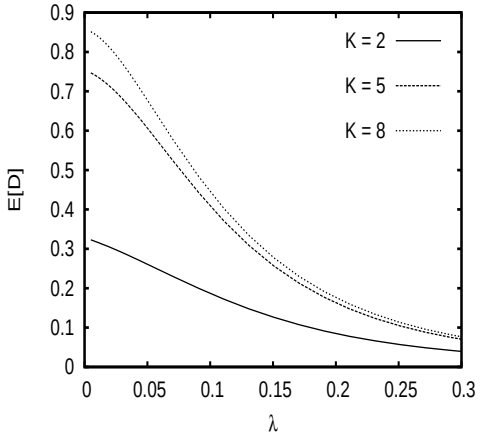


(b) Various values of (deterministic) order size N; $K = 5$

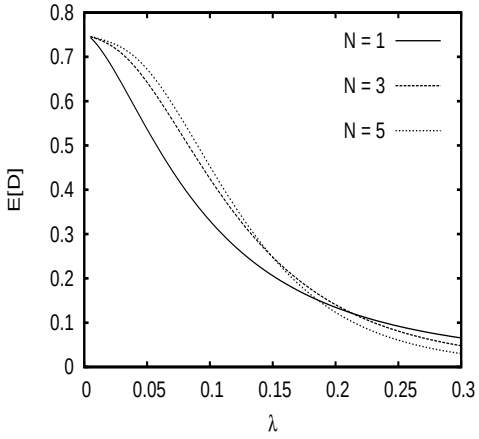
Figure 8: Lower (solid lines), upper (dashed lines), and modified upper (dot-dashed lines) bounds as well as simulated values (dots) of the average flow time $\mathbb{E}[D]$ versus retrieval rate λ; shifted exponential picking times ($\beta = 1$, $\mu = 1/2$), shifted exponential completion times ($\zeta = 3$, $\alpha = 1/2$)

($b_{\min} = 3$, $c_{\min} = 5$) and in case of shifted exponential service and completion times ($b_{\min} = 1$, $c_{\min} = 3$), the modified upper bound is applicable and therefore also depicted (Figures 7-8).

Figures 6-8 show that the bounds differ significantly for very small retrieval rate λ and that the difference rapidly decreases for increasing λ. Although the bounds are not accurate for very small λ, Figures 6-8 reveal that the flow times are unacceptably long in that case,



(a) Various values of line capacity K; N stochastic with mass function (28)



(b) Various values of (deterministic) order size N; $K = 5$

Figure 9: Relative differences between bounds (i.e., (upper bound-lower bound)/upper bound) versus retrieval rate λ; exponential picking times ($\mu = 1/3$), exponential completion times ($\alpha = 1/5$)

implying that λ should not be too small in practice. Figures 6-8 also illustrate that the lower and upper bounds virtually coincide when λ is large enough and that the values of the bounds become insensitive to the exact distribution of the picking and completion times. This can be understood by examining Formulas (24) and (26)-(27) for the bounds, uncovering that both the upper and lower bounds converge to $K(\mathbb{E}[N]\,\mathbb{E}[B] + \mathbb{E}[C])$ as $\lambda \to \infty$. Hence, for large λ, the distributions of the picking and completion times only play a role through their first moments and the flow time only consists of picking and completion times, i.e., no idle times occur.

Remark 6. *Note that the analysis of the order flow time is partially based on a relation with the cyclic queueing network system depicted in Figure 4. In [20], it is shown that the average cycle time in the cyclic system, defined as the time between consecutive departures of the same customer from the same queue, tends to $F\mathbb{E}[B]$ for F large enough and $\mathbb{E}[B] > \lambda^{-1}$, which is called "influence of the slowest server". In our model, $\lambda \to \infty$ implies $\mathbb{E}[B] > \lambda^{-1}$ and $\mathbb{E}[C] > \lambda^{-1}$. The cycle times in the slightly related systems (set $F = K\mathbb{E}[N]$ and assume that a fraction of $1/\mathbb{E}[N]$ customers receives an additional service of length C in $\hat{Q}_1$) thus exhibit similar asymptotic behaviour. This can be understood intuitively by letting the end of an order completion correspond to "firing a token" (of the last order line of the next order processed in Q_1) from $\hat{Q}_1$ to $\hat{Q}_2$ (customers in the cyclic network system thus represent tokens in this regard). The newly released order is satisfied the next time the same token is fired from $\hat{Q}_1$ to $\hat{Q}_2$.*

Figures 6-8 reveal that the order flow time is increasing in K and N, as expected. In addition, N only slightly affects the accuracy of the bounds, whereas increasing K has a detrimental effect on the accuracy of the upper bound; this is further illustrated in Figure 9, which shows the relative difference of the bounds in case of exponential service and completion times.

Figures 6-8 illustrate that, for $K = 2$, the bounds are tight even for quite small λ and that the upper bound is slightly better than the lower bound. The latter can be understood by examining the key features of the bounds. The lower bound underestimates the intervisit time of Q_1 by assuming that it contains no idle periods and compensates this partially by a longer visit time of Q_1. The upper bound overestimates the intervisit time of Q_1 by overestimating the total idle time during the intervisit and compensates this partially by an underestimation of the length of the following visit period of Q_1. The overestimation of the idle times during the intervisit, though, is minor as it takes into account the number of tote arrivals in Q_2 during the services and completion of the preceding visit of Q_1.

For K larger than 2, the bounds differ more for small λ, mainly because the upper bound becomes less accurate. The reason is that to assess the total idle time during a visit of some Q_i ($i > 2$) (note that this is part of the intervisit time of Q_1), the number of totes present at Q_i upon the arrival of the picker at that queue is assessed. This is underestimated by taking into account the number of arrivals in Q_i during the services and completion of the visit of Q_{i+1} during the preceding cycle, thus not only ignoring the idle periods during that visit, but also ignoring the picking and completion times during the visits of the other $K - 2$ queues. This explains why the upper bound is excellent for $K = 2$ and becomes worse for larger K. Note, however, that K should not be too large in practice in order to avoid deadlocks of the closed-loop conveyor and that the modified upper bound leads to a considerably better accuracy than the original upper bound. Nevertheless, we advise practitioners to use the lower bound to assess the flow time, as this bound is much more accurate (except for $K = 2$) than the upper bound and, in general[1], more accurate than the modified upper bound. In addition, the lower bound is easier to compute because it only requires calculation of derivatives of $\tilde{B}^j(\lambda(1-z))$ and $N(\tilde{B}(\lambda(1-z)))^{K-1}\tilde{C}(\lambda(1-z))^{K-1}$, which, in many occasions, can be expressed explicitly. For instance, in case of exponential service and completion times, it holds:

$$\frac{\mathrm{d}^i}{\mathrm{d}z^i}\tilde{B}^j(\lambda(1-z))\bigg|_{z=0} = \frac{(j+i-1)!}{(j-1)!}\left(\frac{\mu}{\mu+\lambda}\right)^j\left(\frac{\lambda}{\mu+\lambda}\right)^i \quad ,$$

and

$$\begin{aligned}\frac{\mathrm{d}^k}{\mathrm{d}z^k}&\left[N(\tilde{B}(\lambda(1-z)))\tilde{C}(\lambda(1-z))\right]^{K-1}\bigg|_{z=0}\\ &= \sum_{w=1}^{M}\mathbb{P}[N=w]\sum_{i=0}^{k}\binom{k}{i}\frac{(w(K-1)+i-1)!}{(w(K-1)-1)!}\left(\frac{\mu}{\mu+\lambda}\right)^{w(K-1)}\left(\frac{\lambda}{\mu+\lambda}\right)^i\\ &\qquad\qquad \frac{(K+k-i-2)!}{(K-2)!}\left(\frac{\alpha}{\alpha+\lambda}\right)^{K-1}\left(\frac{\lambda}{\alpha+\lambda}\right)^{k-i} \quad .\end{aligned}$$

Figures 6-8 clearly indicate that the performance of warehouse systems can be improved considerably by speeding up the retrieval process of totes from the storage area. It should, however, also be noticed that the performance improvement becomes gradually less pronounced,

[1]Note that in Figure 7, the modified upper bound is exceptionally accurate, because of the deterministic service and completion times.

and eventually becomes negligible. As expediting the retrieval process corresponds to significant investments in e.g. faster cranes and conveyor belts, a cost-benefit analysis is required. This is where the bounds play a role: they clearly show the effect of a particular retrieval rate. For instance, Figures 6-8 illustrate that from $\lambda = 0.2$ onwards, the performance cannot be improved considerably by investing in the retrieval process. In conclusion, we advise practitioners to plot the lower bound versus the retrieval rate as this will prove to be valuable when designing the storage area cost effectively: it aids in setting targets for the throughput of the storage area.

5 Conclusions

In this paper, stochastic bounds have been established for the order flow time in parts-to-picker order-picking systems with remotely located manual workstations. The lower bound is more accurate than the upper bound, except when the line capacity equals two, and it is easy to compute. Therefore, we advise practitioners to plot the lower bound of the order flow time versus the retrieval rate of totes from the storage area when designing a parts-to-picker order-picking system with remotely located manual workstations. This procedure shows that the order flow time decreases for increasing retrieval rate, and, more importantly, it indicates from which retrieval rate onwards the order flow time cannot be improved considerably, providing a useful tool to set targets for the retrieval rate. As such, we believe that the results from this paper are complementary to those of existing literature, which mainly focuses on operations in the storage area.

Besides its contribution to warehousing management science, this paper also contributes to the field of queueing theory: a polling system with a new type of arrival process and service discipline has been analysed. Although this polling system does not satisfy the branching property, useful stochastic bounds for the cycle time have been deduced.

There are several opportunities for future research. For instance, the distribution of the total queue length (i.e., the aggregation of the number of totes in all virtual queues) is of interest as it can help to determine the number of actual buffer lanes and their sizes required to achieve a certain target deadlock rate of the closed-loop conveyor. Another possibility is to deduce stochastic bounds for the variance of the order flow time, which could be carried out along the same lines as in this paper.

Acknowledgment Dieter Claeys is a Postdoctoral Fellow with the Research Foundation Flanders (FWO-Vlaanderen), Belgium. His research was conducted during a visit of the author at the EURANDOM Research institute, and was supported by a travel grant of the FWO-Vlaanderen. Part of his research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. The research of Onno Boxma was supported by the NETWORKS program of the Dutch government and by the IAP BESTCOM program of the Belgian government. The authors gratefully acknowledge stimulating discussions with Pascal Etman.

References

[1] R. Andriansyah, Order-picking workstations for automated warehouses, Ph.D. thesis, Eindhoven University of Technology (2011).

[2] D. Twist, The impact of radio frequency identification on supply chain facilities, Journal of Facilities Management 3(3) (2005) 226–239.

[3] J. Tompkins, J. White, Y. Bozer, J. Tanchoco, Facilities Planning, 4th Edition, John Wiley and Sons, 2010.

[4] F. Dallari, G. Marchet, M. Melacini, Design of order picking system, The International Journal of Advanced Manufacturing Technology 42 (1-2) (2009) 1–12.

[5] K. Roodbergen, I. Vis, A survey of literature on automated storage and retrieval systems, European Journal of Operational Research 194 (2009) 343–362.

[6] R. Andriansyah, L. Etman, J. Rooda, Flow time prediction for a single-server order picking workstation using aggregate process times, International Journal on Advances in Systems and Measurements 3 (2010) 35–47.

[7] R. Andriansyah, L. Etman, I. Adan, J. Rooda, Design and analysis of an automated order-picking workstation, Journal of Simulation 1 (2013) 1–13.

[8] R. Bekker, P. Vis, J. L. Dorsman, R. D. van der Mei, E. M. M. Winands, The impact of scheduling policies on the waiting-time distributions in polling systems, Queueing Systems 79 (2015) 145–172.

[9] O. Boxma, O. Kella, K. Kosinski, Queue lengths and workloads in polling systems, Operations Research Letters 39 (6) (2011) 401–405.

[10] O. Boxma, D. Claeys, L. Gulikers, O. Kella, A queueing system with vacations, submitted to Naval Research Logistics.

[11] S. Fuhrmann, Performance analysis of a class of cyclic schedules, Technical Memorandum 81-59531-1 Bell Laboratories, 1981.

[12] J. Resing, Polling systems and multitype branching processes, Queueing Systems 13 (1993) 409–426.

[13] O. Boxma, Models of two queues: a few new views, in: Teletraffic Analysis and Computer Performance Evaluation, eds. O.J. Boxma. J.W. Cohen and H.C. Tijms (North-Holland Publ. Cy., Amsterdam), 1986, pp. 75–98.

[14] D.-S. Lee, A two-queue model with exhaustive and limited service disciplines, Stochastic Models 12 (2) (1996) 285–305.

[15] T. Ozawa, Alternating service queues with mixed exhaustive and k-limited services, Performance Evaluation 11 (1990) 165–175.

[16] T. Ozawa, Waiting time distribution in a two-queue model with mixed exhaustive and gated-type k-limit service, in: Proceedings of International Conference on the Performance and Management of Complex Communication Networks, Tsukuba, 1997, pp. 231–250.

[17] E. Winands, I. Adan, G. van Houtum, A two-queue model with alternating limited service and state-dependent setups, in: Proceedings of Analysis of Manufacturing Systems, Production Management, Thessaloniki, Greece, 2005, pp. 200–208.

[18] O. Boxma, The cyclic queue with one general and one exponential server, Advances in Applied Probability 15 (1983) 857–873.

[19] S. Ross, Stochastic Processes, second edition, John Wiley and Sons, New York, 1996.

[20] O. Boxma, Sojourn times in cyclic queues - the influence of the slowest server, in: Computer Performance and Reliability, Proceedings of the Second International MCPR Workshop, Rome, eds. G. Iazeolla, P.J. Courtois and O.J. Boxma (Elsevier Science Publishers, North-Holland), 1988, pp. 13–24.