Evaluating Order Throughput Time in 2-Block Warehouses with Time Window Batching

Inneke Van Nieuwenhuyse

Research Center for Operations Management, K.U. Leuven
Naamsestraat 69, 3000 Leuven, Belgium
Inneke.vannieuwenhuyse@econ.kuleuven.be

René B.M. de Koster
RSM Erasmus University
PO Box 1738, 3000 DR Rotterdam, The Netherlands
rkoster@rsm.nl

ABSTRACT
Given the increased pressure on short delivery lead times, minimizing customer order throughput times is an important objective in warehousing operations. Warehouse managers can influence the responsiveness of their system through a number of controls, such as the order batching policy, the capacity of the picking and sorting operations, and the picking policy used (pick-and-sort versus sort-while-pick). This paper studies the impact of these controls on the average customer order throughput time in a numerical way, for warehouses with time window batching and separate picking and sorting functions.

Keywords: Warehousing, order batching, order picking and sorting

1. INTRODUCTION
Warehouses play an important role in companies’ supply chains. Among the many activities carried out in a warehouse, order picking - the process of retrieving products from storage in response to a specific customer request - is the most critical one: it has to be carried out in a short available time, meeting truck departure due times. Order picking may consume as much as 60% of all labor activities in the warehouse, and, for a typical warehouse, the cost of order picking is estimated to be as much as 55% of the total warehouse operating expense (Tompkins et al., 2003). For these reasons, warehousing professionals consider order picking as the highest-priority activity for productivity improvements.
Four operational decision problems influencing the performance of (manual) order picking systems have received attention from researchers (see De Koster et al., 2007 for a comprehensive literature review):

- **Storage assignment.** Storage assignment methods assign stock keeping units (SKUs) to storage locations. The main storage policies mentioned in the literature are randomized, class-based and dedicated storage. The easiest storage method is to randomly allocate incoming products to available storage locations. It is often used in systems that exclusively consist of fast-moving items, or items with a short lifecycle (e.g. fashion items). In other settings, pick-frequency class-based storage strategies tend to be popular.

- **Layout problem.** This is the problem of finding a good aisle configuration (i.e. the optimal number and length of aisles) minimizing order picking throughput time. Little research has been done in this area (see de Koster et al, 2007 for references).

- **Routing order pickers.** This well-researched problem considers the determination of the optimal sequence of visits to pick up a number of requested items as quickly as possible, depending on the layout and working methods. The disadvantage of exact algorithms is that they are not available for all types of layouts, and that the resulting routes may be too complicated for pickers to follow. For large and more complicated layouts (more than two blocks) several heuristics are documented. While the Combined heuristic is probably the best routing heuristic known so far (Roodbergen and De Koster, 2001), the S-shape heuristic is most popular in practice. It is easy to implement, reduces aisle congestion by forcing single-directional traffic in aisles, and the generated pick sequences suggest a 'logical' way of working for pickers.

- **Batching and zoning.** Batching determines which orders are released together to the picking operation. With batch picking, multiple orders are picked together in one pick tour and need to be sorted by order. The main advantage of batch picking over single-order picking is that the average travel time per order is reduced by sharing a pick tour. In stochastic settings, time-window batching is prevalent. With variable time window batching (which will be referred to as VTWB), a predetermined number of orders needs to be grouped in a batch (hence, the time window for collection of the orders is variable). With fixed time window batching (referred to as FTWB), the batch contains all orders that have arrived during a given (fixed) time interval. Zoning is closely related to batching; it divides the pick area into sub-divisions (or zones), each with one or few pickers dedicated to it. The major advantages of zoning are: reduction of the travel time (because of the smaller traversed area and also the familiarity of the picker with the zone) and of the traffic congestion.

As stated in Gu et al. (2007), the literature on optimal order batching is rather scarce. The majority of current batching models (e.g. Gademann et al. 2001, Gademann and Van de Velde 2005, Chen and Wu 2005, Elsayed et al. 1993) assume that the set of customer orders is known at the time the batching decision is
taken, and then focus on which orders to assign to a single tour, in view of a given performance criterion (e.g. picker capacity, order due dates). Consequently, these models are ill-suited to guide batching decisions in a realistic (stochastic, online) setting, where there is no advance information about the actual composition or size of a customer order, nor about the timing of customer order arrival. As mentioned above, time-window batching is the most popular batching method in this type of warehouses. Rather than assigning orders of a given set to tours, a central question in this setting is how to choose the fixed number of orders (in case of VTWB) or the fixed time window (in case of FTWB) in view of minimizing the average throughput time of the orders.

Only a few papers have studied order batching in a stochastic (online) context. Chew and Tang (1999) and Le-Duc and De Koster (2007) have studied VTWB, assuming exponential interarrival and/or processing times, and single server stations. These papers do not model the sorting function explicitly; this limits the insights that may be gained from the models, as the trade-offs between the picking and sorting functions cannot be accounted for. Choe et al. (1993) consider a single-block layout for the picking area and model the sorting function as a single (in their case, automated) server. They consider both FTWB and VTWB; however, their FTWB model assumes that pick batch interarrival times are deterministic, which is not necessarily the case (see the discussion in Section 4 below). To the best of our knowledge, this paper is the first to study the responsiveness of 2-block warehouses with random customer order sizes, consecutive sorting and packing processes, multi-server picking and sorting operations, and general setup and service time distributions, in case of VTWB as well as FTWB. To this end, we model the warehouse operation as a 2-stage general queueing system.

In the next section, we describe the warehouse operation and introduce notations and assumptions. To allow for analytical tractability (more specifically, of the travel time component), we need to make specific assumptions concerning storage, routing and layout. In section 3, we present the model for approximating the average and variance of the customer order throughput time as a function of batch size for the VTWB case. In Section 4, we outline the changes that would be required to reflect FTWB. In this way, we can compare both batching methods in the subsequent analysis. Section 5 shows the results of the validation experiment, and discusses the trade-offs that imply the existence of an optimal batching policy yielding a minimum average throughput time. In section 6, the impact of order picking policies (pick-and-sort versus sort-while-pick) and different worker allocations on this minimum average throughput time is evaluated numerically. Section 7 summarizes the conclusions.

2. NOTATION AND ASSUMPTIONS

We consider a setting in which orders arrive online and are then batched for picking. Batched orders are picked in one pick tour in a rectangular 2-block warehouse, as sketched in Figure 1. This layout is rather
common in practice and used by Caron et al. (1998), among others. The warehouse is assumed to have random storage.

Customer orders are grouped either according to variable time window batching (implying that a fixed number of customer orders $k$ is grouped into an order picking batch), or according to fixed time window batching (all customer orders that arrived in a given time interval $\Delta t$ are grouped in a pick batch). The number of order lines in a customer order follows a known discrete distribution, and the arrival process of customer orders is Poisson (which is acceptable, assuming that customer orders originate from a large number of customers ordering independently).

When an order batch is started for picking, setup time ($SU_p$) is required. This time includes the time for the picker to obtain a pick list and an empty pick cart at the input/output (I/O) point of the picking zone and, at the end of the pick process, drop-off the pick cart at the I/O point. It is assumed to be independent of the order batch size, and to be generally distributed. Pickers travel along the aisles at constant speed, according to a certain routing policy (we assume S-shape routing). They can pick in the rack at the left and right side of the aisle without additional travel (this is called ‘two-sided picking’). All rack levels can be reached by the picker without additional vertical travel (like in conventional shelf racks, or pallet racks where picking occurs from the bottom levels only). The picking times per order line are IID distributed (denoted by $X_p$) with a general probability distribution.

\[\text{(Insert Figure 1 here)}\]

As the picked items on the pick cart belong to multiple orders, they have to be sorted and, in most cases, packed by order. If the picker has already grouped the order lines on the pick cart per order, only consecutive packing is needed. Such a picking process is called ‘sort-while-pick’ (SWP) and is possible only if the number of orders in the batch is not too large, since sorted orders occupy more space, and to avoid sorting errors by the picker. In a pick-and-sort policy (PAS), the picker does not sort the orders during the picking process; both sorting and packing happen at a second process. Regardless of the policy used, pick batches that are complete are dropped off by the picker at the I/O point and moved to the (sorting and) packing area. Here, a setup time ($SU_s$) is required to retrieve the cart from the depot, and for retrieving packing instructions and materials. The sorting plus packing time per order line is denoted by $X_s$. Both $SU_s$ and $X_s$ are random with a general probability distribution. The picking and sorting processes are carried out by either single or multiple parallel operators.

In the next section, we propose the queueing model that allows to estimate the average customer order throughput time for the VTWB case, in terms of the order batch size. Section 4 then describes the changes required to incorporate FTWB. The following notation will be used:

- $d$: length (in travel time units) of a pick aisle

\[d\]
The notation $E(Z)$ denotes the expected value of random variable $Z$; $\text{Var}(Z)$ denotes its variance, $E(Z^2)$ its second moment, and $c_Z^2$ its squared coefficient of variation (SCV).

3. ESTIMATION OF AVERAGE CUSTOMER ORDER THROUGHPUT TIME: VTWB CASE

The throughput time $W$ of a customer order can be written as a sum of different components:

- the time that the order spends waiting for the pick batch to be formed (the collection time), denoted by $W_{coll}$;
- the time that the pick batch spends in queue at the picking area and the sorting area, denoted by $W_{q,p}$ and $W_{q,s}$ respectively;
- the total service time needed for a pick batch in the picking area, $S_p$, which consists of the setup time $SU_p$, the travel time $TR$ and the picking time $W_p$;
- the total service time needed for a pick batch in the sorting area, $S_s$, which consists of the setup time $SU_s$ and the sorting time $W_s$.

Each of these components (except the setup times) depends on the number of customer orders $k$ that are collected into the batch. Hence, we may write the following expressions:
As the customer order interarrival times $Y$ are IID distributed, the average of the collection time of an arbitrary order in a batch containing $k$ orders is given by:

$$E(W_{coll|k}) = \frac{k-1}{2} E(Y)$$

(2)

In what follows, we first discuss the first and second moments of the service time for a pick batch in the picking and sorting area; next, we present the queueing expressions which are used to approximate the average queueing time at the picking and sorting operations.

### 3.1 Average and Variance of Service Time in the Sorting Area

The service time for a pick batch in the sorting area consists of the setup time and the sorting time:

$$S_s = SU_s + W_s$$

(3)

As $SU_s$ and $W_s$ are independent, the average and variance of $S_s$ can be written conditional upon the number of customer orders $k$ contained in the pick batch:

$$E(S_s|k) = E(SU_s) + E(W_s|k)$$

$$Var(S_s|k) = Var(SU_s) + Var(W_s|k)$$

(4)

The average and variance of the setup time $SU_s$ are given, and are independent of the pick batch size $Q$. As the sorting times for individual order lines in the order batch ($X_s$) are IID distributed, the sorting time $W_s$ for the order batch can be interpreted as the sum of a random number ($Q$) of IID random variables ($X_s$), with $Q$ and $X_s$ independent. For a specific number of customer orders $k$, the first moment of $W_s$ is then given by:

$$E(W_s|k) = E(Q|k)E(X_s)$$

(5)

where $E(Q|k)$ refers to the expected number of order lines in the pick batch, given that it consists of $k$ customer orders:

$$E(Q|k) = \sum_{n=k}^{\infty} nP^k(n) = kE(O)$$

(6)

Here, $P^k(n)$ denotes the probability that the order batch size $Q$, for a given number of customer orders $k$, consists of $n$ order lines. The expression for $P^k(n)$ is given by the k-fold convolution of the probability mass function of the customer order size $O$. Note that the number of order lines in the pick batch will be larger than or equal to $k$, as a single customer order consists of at least one order line.

The variance of $W_s$, for a given number of customer orders $k$ in the pick batch, can be written as

$$Var(W_s|k) = E(Q|k)Var(X_s) + Var(Q|k)[E(X_s)]^2$$

(7)

where $Var(Q|k)$ is given by
By means of expressions (4) to (8), \( E(S_jk) \) and \( \text{Var}(S_jk) \) in expression (1) can be determined for any arbitrary number of customer orders \( k \) contained in the pick batch, and any arbitrary discrete distribution of the customer order size.

### 3.2 Average and Variance of Service Time in the Picking Area

The service time for a pick batch in the picking area, \( S_p \), is given by the sum of the setup time, the actual picking time, and the travel time of the order picker:

\[
S_p = SU_p + W_p + TR
\]

As the three random variables are mutually independent, the average and variance of \( S_p \) are given by:

\[
\begin{align*}
E(S_p|k) &= E(SU_p) + E(W_p|k) + E(TR|k) \\
\text{Var}(S_p|k) &= \text{Var}(SU_p) + \text{Var}(W_p|k) + \text{Var}(TR|k)
\end{align*}
\]

for any given number of customer orders \( k \) contained in the pick batch. The average and variance of \( SU_p \) are given, and are independent of the pick batch size \( Q \). The random variable \( W_p \) depends on \( Q \) in a way that is similar to \( W_s \):  

\[
\begin{align*}
E(W_p|k) &= E(Q|k)E(X_p) \\
\text{Var}(W_p|k) &= E(Q|k)\text{Var}(X_p) + \text{Var}(Q|k)\left[E(X_p)\right]^2
\end{align*}
\]

with \( E(Q|k) \) given by expression (6), and \( \text{Var}(Q|k) \) by expression (8). The average and variance of the travel time \( TR \) depend not only on the pick batch size, but also on three additional factors: warehouse layout, storage strategy, and routing policy of the order pickers through the warehouse. For a 2-block rectangular warehouse with S-shape routing policy and random storage, as we are considering here, an approximation for the first and second moment of travel time has been described in Le-Duc and De Koster (2007). In what follows, we present simplified versions of these expressions and adapt them to incorporate the impact of variable pick batch sizes.

The first moment of travel time for a pick batch containing \( n \) order lines consists of four components:

\[
E[TR|n] = E[TR_{wa}|n] + E[TR_{ca}|n] + E[AT_1|n] + E[AT_2|n].
\]

Both \( E[AT_1|n] \) and \( E[AT_2|n] \) are adjustment terms (see Le-Duc and de Koster, 2007). For a random storage strategy, the expressions for the average within-aisle travel time \( E[TR_{wa}|n] \) and the average cross-aisle travel time \( E[TR_{ca}|n] \) derived in Le-Duc and De Koster (2007) in fact reduce to:
The first moment of travel time for a pick batch consisting of k customer orders is then given by:
\[
E[TR_{WA}|k] = dm \left[1 - \left(1 - \frac{1}{m}\right)^k\right]
\]
(12)

\[
E[TR_{CA}|n] = 2w_c \left[\frac{m}{2} - \sum_{i=1}^{m/2-1} \left(\frac{(2m-1)i}{m^2}\right)^n\right] = 2w_c \left[\frac{m}{2} - \frac{(2m-1)^n}{m^{2n}} \sum_{i=1}^{m/2-1} l^n\right]
\]

The second moment of travel time for a given pick batch size of n order lines is given by (Le-Duc and De Koster, 2007):
\[
E[TR^2|n] = d^2 E[J^2|n] + (2w_c)^2 E[L^2|n] + 4w_c dE[JL|n]
\]
(14)

where J refers to the number of aisles visited, and L to the pick line (see Figure 1) of the farthest visited aisle. In case of a random storage policy, these components reduce to:
\[
E[J^2|n] = m \left( (m-1) \left(\frac{m-2}{m}\right)^n + (1-2m) \left(\frac{m-1}{m}\right)^n + m \right)
\]
\[
E[L^2|n] = (m/2)^2 - \sum_{i=1}^{m/2-1} (2i+1) \left(\frac{(2m-1)i}{m^2}\right)^n = (m/2)^2 - \frac{(2m-1)^n}{m^{2n}} \sum_{i=1}^{m/2-1} (2i+1)i^n
\]
(15)
\[
E[JL|n] = 2^{n+1} \sum_{i=1}^{m/2} i \left[1 - \left(1 - \frac{1}{2l}\right)^n\right] \frac{l^{n+1}}{m^n} + \left(1 - \left(1 - \frac{1}{2(l-1)}\right)^n\right) \frac{(l-1)^{n+1}}{m^n}
\]

The variance of travel time, for a given number of order lines n, can be approximated using expressions (12) and (14):
\[
Var[TR|n] = E[TR^2|n] - (E[TR_{CA}|n] + E[TR_{WA}|n])^2
\]
(16)

The variance of travel time for a pick batch consisting of k customer orders is then given by:
\[
Var[TR|k] = \sum_{n=k}^{\infty} P^k(n) E[TR^2|n] - \left(\sum_{n=k}^{\infty} P^k(n)(E[TR_{CA}|n] + E[TR_{WA}|n])\right)^2
\]
(17)

Using expressions (11), (13) and (17) in expression (10), we obtain an approximation for \(E[S_p|k]\) and \(Var[S_p|k]\).
### 3.3 Approximation for the average of $W_{q,p}$ and $W_{q,s}$

In the system under study, pick batches may have to wait in queue both at the picking area and at the sorting area. Single-class $GI/G/1$ and $GI/G/m$ queueing expressions may be used to approximate the average and variance of these waiting times.

In case of a single-server station, we rely upon the approximations of Whitt (1983). Assuming that $k$ customer orders are contained in a pick batch, we can write the following for $E(W_{q,p})$:

$$E(W_{q,p}|k)_{GI/G/1} = \begin{cases} \frac{\rho_p^2 \left( c_{a_p}^2 + c_{s_p}^2 \right)}{2\lambda_p \left[ 1 - \rho_p \right]} & \text{if } c_{a_p}^2 \geq 1 \\ \rho_p \left( c_{a_p}^2 + c_{s_p}^2 \right) \exp \left\{ -2 \left[ 1 - \rho_p \right] \left[ 1 - c_{a_p}^2 \right]^2 \right\} & \text{otherwise} \end{cases}$$

(18)

where $\lambda_p$ denotes the arrival rate of pick batches at the picking area ($\lambda_p = \frac{1}{E(Y_p)}$), $c_{a_p}^2$ denotes the SCV of interarrival times of pick batches at the picking area, $c_{s_p}^2$ refers to the SCV of pick batch service time and $\rho_p$ to the utilization of the picking server. These parameters depend on the number of customer orders $k$ being grouped in the pick batch:

$$\lambda_p = \frac{1}{k E(Y)}, c_{a_p}^2 = \frac{k \text{Var}(Y)}{(k E(Y))^2}, c_{s_p}^2 = \frac{\text{Var}(S_p|k)}{(E(S_p|k))^2}, \rho_p = \lambda_p E(S_p|k)$$

(19)

with $E(S_p|k)$ and $\text{Var}(S_p|k)$ given by expression (10). In fact, the expression for $c_{a_p}^2$ is valid as long as the interarrival times are IID distributed, regardless of the actual probability distribution. In the case of a Poisson customer order arrival process, $c_{a_p}^2 = \frac{1}{k}$ (as $Y_p$ follows an Erlang distribution). The average waiting time of a pick batch at the sorting server, $E(W_{q,s})_{GI/G/1}$ can be derived in an analogous way, using the following input parameters:

$$\lambda_s = \frac{1}{k E(Y)}, c_{s_s}^2 = \frac{\text{Var}(S_s|k)}{(E(S_s|k))^2}, \rho_s = \lambda_s E(S_s|k)$$

(20)

Here, $E(S_s|k)$ and $\text{Var}(S_s|k)$ are given by expression (4) above. The expression for $\lambda_s$ follows from the conservation of flow in the system: this implies $\lambda_s$ is equal to $\lambda_p$. As the sorting operation takes place after the picking operation, the SCV of interarrival times of pick batches at the sorting area ($c_{a_s}^2$) equals the SCV of interdeparture times of pick batches from the picking area (which we will denote by $c_{d_p}^2$). When the picking station is a single server station, $c_{a_s}^2$ can be approximated by the following linking equation (Marshall, 1968):
In case of multi-server stations, we approximate \( E(W_{q,p})_{G|G_{in}} \) and \( E(W_{q,s})_{G|G_{in}} \) by the expressions developed by Whitt (1993), which are given in Appendix A. When the picking station is a multiserver station, \( ca_s^2 \) can be approximated by the following linking equation (Whitt, 1983; Hopp and Spearman, 2000):

\[
ca_s^2 = cd_p^2 = ca_p^2 + 2\rho_p^2 cs_p - 2\rho_p(1 - \rho_p) \frac{E(W_{q,p}|k)}{E(S_p|k)}
\]

(21)

Note that, both in the single-server and the multi-server case, \( \rho_p \) and \( \rho_s \) need to be strictly smaller than 1 in order to maintain system stability; otherwise, the respective waiting times grow to infinity.

### 4. ESTIMATION OF AVERAGE CUSTOMER ORDER THROUGHPUT TIME WITH FTWB

With fixed time window batching, the system checks at regular time intervals (denoted by \( \Delta t \)) if any new customer orders have arrived. All customer orders present at the time of review are grouped in a pick batch, so the number of orders grouped into a batch will be variable. Assuming a Poisson arrival process of customer orders with rate \( \lambda \), the probability that \( k \) customer orders arrive in an interval \( \Delta t \) is given by:

\[
P(k|\Delta t) = \frac{e^{-\lambda\Delta t}(\lambda\Delta t)^k}{k!} \quad \forall k \geq 0
\]

(23)

Evidently, when \( k=0 \) (no customer orders have arrived during the time interval), no pick batch is formed and the system will wait for another interval \( \Delta t \) before checking the status of the order list again. Denoting the probability \( P(0|\Delta t) = e^{-\lambda\Delta t} \) by \( P_0 \), the average and variance of the interarrival time of pick batches at the picking station is given by:

\[
E(Y_p|\Delta t) = \sum_{i=0}^{\infty} (i+1)(\Delta t)(1-P_0)(P_0)^i = (\Delta t) \frac{1}{1-P_0}
\]

\[
Var(Y_p|\Delta t) = \left[ \sum_{i=0}^{\infty} (i+1)(\Delta t)^2(1-P_0)(P_0)^i \right] - [E(Y_p)]^2 = (\Delta t)^2 - \frac{1}{(P_0-1)^2} - [E(Y_p)]^2
\]

(24)

The choice of \( \Delta t \) impacts the average and variance of the number of order lines contained in the batch:

\[
E(Q|\Delta t) = \sum_{k=1}^{\infty} P(k|\Delta t) \left[ \sum_{n=0}^{\infty} np^k (n) \right]
\]

\[
Var(Q|\Delta t) = \sum_{k=1}^{\infty} P(k|\Delta t) \left[ \sum_{n=0}^{\infty} n^2 p^k (n) \right] - [E(Q|\Delta t)]^2
\]

(25)

Note that the probabilities \( P(k|\Delta t) \) (\( k \geq 1 \)) need to be truncated, as with probability \( P_0 \) no batch is formed.

The expressions in (25) influence the average and variance of the service time in the picking and sorting area. For the sorting area, we have:
For the picking area, we obtain:

\[
E(S_p | \Delta t) = E(SU_p) + E(W_p | \Delta t) + E(TR | \Delta t)
\]
\[
Var(S_p | \Delta t) = Var(SU_p) + Var(W_p | \Delta t) + Var(TR | \Delta t) + Var(Q | \Delta t)[E(X_p)]^2
\]

(27)

with

\[
E(W_p | \Delta t) = E(Q | \Delta t)E(X_p)
\]
\[
Var(W_p | \Delta t) = Var(Q | \Delta t)Var(X_p) + Var(Q | \Delta t)[E(X_p)]^2
\]

(28)

As the status of the order list is checked at regular intervals \( \Delta t \), the average collection time of a customer order is now given by:

\[
E(W_{coll} | \Delta t) = \frac{\Delta t}{2}
\]

(29)

5. IMPACT OF \( k \) AND \( \Delta t \) ON AVERAGE ORDER THROUGHPUT TIME

The models presented in Sections 2 and 3 were validated by means of a discrete-event simulation experiment. The parameters used in this experiment are shown in Table 1.

Combining all parameters at all levels yields 72 different scenarios. For the VTWB case, we varied the number of customer orders per pick batch (\( k \)) from 1 to 15. Equivalently, for the FTWB case, the time interval (\( \Delta t \)) was chosen as a multiple of the average customer order interarrival time \( E(Y) \): \( \Delta t = zE(Y) \), with \( z \)
varying between 1 and 15. Evidently, for each scenario, only those values of \( k \) (resp. \( z \)) that yield a stable system (\( \rho_p < 1 \) and \( \rho_s < 1 \)) were validated. For every setting, the average throughput time of a customer order as observed in the simulation (\( E(W)_{\text{simul}} \)) was compared to the corresponding average throughput time obtained in the queuing model (\( E(W)_{\text{analytic}} \)). The resulting error is defined as \( \varepsilon = \frac{E(W)_{\text{analytic}} - E(W)_{\text{simul}}}{E(W)_{\text{simul}}} \).

The results show that both the VTWB and the FTWB models provide adequate approximations of the average customer order throughput time\(^1\). The VTWB model in general tends to slightly overestimate \( E(W) \): an average absolute error of about 1% was observed. The FTWB model in general underestimates \( E(W) \): here, an average absolute error of 5.3% was observed. Moreover, the results reveal that there is a convex relationship between \( k \) and \( E(W) \) in the VTWB setting, and \( z \) and \( E(W) \) in the FTWB setting. Figure 2a shows the impact of \( k \) on \( E(W) \) and its components for the VTWB case with \( E(Y) = 60 \text{ sec}, d = 60 \text{ sec}, m = 20, E(O) = 1.5, n_p = 4 \) and \( n_s = 2 \). Figure 2b shows the relationship between \( z \) and \( E(W) \) for the same setting, in case of FTWB. Other settings yielded similar results, showing that \( E(W) \) is, in general, convex in \( k \) (resp. \( z \)).

The convex relationship is the result of a trade-off between two underlying effects, known as the saturation effect and the batching effect (Karmarkar et al., 1985; Vandaele et al. 2003). When \( k \) (resp. \( z \)) is small, the collection time and service time are small, but the waiting time will be large. This is due to the fact that, with small values of \( k \) (resp. \( z \)), setups need to be performed frequently, which causes the utilization of the picking and sorting stations to increase. An increase in utilization inevitably aggravates congestion: the pick batches will need to wait longer in queue in front of the picking and sorting process. As \( \rho_p \) and/or \( \rho_s \) approach 1, the system becomes saturated (hence the term saturation effect). In the picking area, this saturation effect is further emphasized by the travel times: even for small values of \( k \) (resp. \( z \)), an expected travel time is incurred, which may be large especially in large warehouses (with a large number of aisles). This further exacerbates the saturation effect. Due to the saturation effect, the smallest value of \( k \) (resp. \( z \)) that yields a stable system (referred to as \( k_{\text{crit}} \) resp. \( z_{\text{crit}} \)) may be larger than one (as in the VTWB case shown in Figure 2a). Note that the congestion effect in our model refers to the time that pick batches wait in queue in front of both processes; it does not refer to traffic congestion that might occur due to an excessive number of pickers being present in a given aisle. Traffic congestion is ignored in the current model, as we assume that pickers travel at constant speed (see Section 2).

\(^1\) The VTWB model was also validated on settings where the customer interarrival time \( Y \) is not exponentially distributed. Details on the results can be found in Van Nieuwenhuyse et al. (2007).
In contrast, when \( k \) (resp. \( z \)) is large, the waiting time will be small as the utilization of the system goes down (fewer setups need to be performed). The collection time, service time and travel time though will be large, and will hence be the dominant components of \( E(W) \). This is referred to as the batching effect.

The combination of the saturation effect and the batching effect yields a convex relationship, implying the existence of an optimal \( k \) (for VTWB) or optimal \( z \) (for FTWB), that minimizes \( E(W) \). The actual location of this optimum depends on the relative magnitude of the saturation effect vs. the batching effect (and, as such, on the system’s characteristics): when the saturation effect is weak relative to the batching effect, the smallest feasible value of \( k \) (resp. \( z \)) may be optimal. Otherwise, the saturation effect tends to push the optimal value of \( k \) (resp. \( z \)) upwards. Further validation results have shown that the queueing models are adequate in approximating \( k_{opt} \) (resp. \( z_{opt} \)) and the corresponding optimal value of the customer order throughput time \( E(W_{opt}) \). In the VTWB case, the optimal \( k \) determined by the queueing model coincides with the optimal \( k \) determined by simulation, for all scenarios in the validation experiment. In the FTWB case, the optimal \( z \) determined by the queueing model coincides with the optimal \( z \) determined by simulation, for 69 of the 72 scenarios in the validation experiment: in the other three scenarios, the difference in \( z_{opt} \) was \( \leq 1 \) (in these settings, the curve for \( E(W) \) turned out to be very flat near the optimum).

The trade-off between the batching and saturation effect leads to a number of interesting insights that are particular for the warehousing setting. At small values of \( k \) (resp. \( z \)), the picking station will typically be the bottleneck of the warehousing system, due to the possibly large travel times involved in order picking. As discussed previously, these travel times may rule out small batch sizes particularly in large warehouses, and exert an upward pressure on the optimal value of \( k \) and the corresponding optimal \( E(W) \). Hence, the responsiveness gains obtained by optimal batching are essentially the result of a more efficient use of the pickers’ capacities.

It follows that for a given warehouse setting, \( k_{opt} \) (resp. \( z_{opt} \)) as well as \( E(W_{opt}) \) can be influenced by changing the allocation of workers to picking and sorting operations. Indeed, allocating more personnel to the picking operation at the expense of the sorting operation will increase the picking capacity, and reduce the bottleneck effect (note that the allocation of personnel only impacts the queueing times in the system, and does not influence any other throughput time component). By contrast, increasing the sorting capacity tends to emphasize the bottleneck effect. This raises the question of how to allocate personnel optimally to the picking and sorting operations, in view of minimizing overall waiting time in the system.

Moreover, the question arises whether overall system performance can be improved by shifting tasks between picking and sorting stations, e.g. by moving from a pick-and-sort strategy to a sort-while-pick strategy. Task shifting obviously impacts the relative capacities of the picking and sorting stations, and hence influences optimal system behavior.

Both questions are studied in the next section.
6. INSIGHTS ON THE OPTIMAL $k$ AND OPTIMAL $E(W)$

6.1. Impact of Workforce Allocation to Picking and Sorting Operations

In this section, we use the queuing models developed in Sections 3 and 4 to investigate the allocation of a given workforce to the picking and sorting operations in order to minimize the average customer order throughput time. For this purpose, we applied the model to all scenarios used in the validation experiment (see Section 5, Table 1 for the relevant input data), assuming a total workforce $n_{tot}$ equal to 4 or 8 people. For all possible combinations of $n_s$ and $n_p$, the optimal pick batch size $k_{opt}$ (and corresponding optimal customer order throughput time $E(W)_{opt}$) was determined, along with the utilization of the pickers ($\rho_p$) and of the sorters ($\rho_s$) in the optimum.

Figure 3 shows the resulting optimal customer order throughput times $E(W)_{opt}$ obtained with VTWB and FTWB batching, for different allocations of $n_s$ and $n_p$, for $n_{tot}=4$ (top panes) and $n_{tot}=8$ (bottom panes). For reasons of clarity, the figure only shows a subset of the scenarios ($m=10$ aisles or 20 aisles, $E(O)=1.5$ or 3), but all scenarios revealed the following main result: the overall minimal value of $E(W)_{opt}$ (from now on referred to as $E(W)_{opt}^*$) is obtained by allocating $n_s$ and $n_p$ in such a way that the picking and sorting operations become more closely balanced in the corresponding optimum ($\rho_p/\rho_s$ close to one). Although this result is not straightforward, it is intuitively appealing: when both stations are closely balanced, there is no longer an outspoken bottleneck left in the system. The choice of personnel allocation, along with the corresponding optimal batch size, makes the bottleneck disappear. As the personnel allocation only impacts the waiting time component of $E(W)$, this observation implies that balancing lowers the total waiting time in the system.

For each figure, only one personnel allocation scheme provides the best overall performance (e.g., $n_p=3$ and $n_s=1$ in the top panes, and $n_p=6$ and $n_s=2$ in the bottom panes), for all values of $m$ and $E(O)$. Apparently, the actual choice regarding personnel allocation is a main determinant for the system’s throughput time performance, both in the VTWB and the FTWB case: while optimal batch sizing can be used to fine-tune performance for a given allocation, it is unable to compensate for the inherent performance loss caused by an ill-considered allocation decision. More specifically, an understaffing of the picking (resp. sorting) operation will cause this operation to behave like a bottleneck, a behavior that can only be mitigated (not eliminated) by fine-tuning the batch size. The optimal personnel allocation appears to be quite robust with regard to changes in the warehouse size ($m$) and expected customer order volumes $E(O)$.

(Insert Figure 3 here)
The experiment revealed that, in about 90% of the settings studied, \( E(W)_{\text{opt}} \) obtained with FTWB slightly outperforms \( E(W)_{\text{opt}} \) with VTWB. These results should be interpreted with caution though, due to the approximation errors present in both models.

### 6.2. Comparison of PAS and SWP Strategy

In this section, we study the impact of shifting workload between picking and sorting operations, by moving from a PAS policy to a SWP policy. As mentioned in section 2, the SWP policy implies that the picker sorts the different line items into customer orders before handing them over to the sorting operation; the sorting personnel primarily takes care of packaging the customer orders and preparing them for shipment. Note that, all other things being equal, such a shift increases the inherent imbalance in workload between the two stations, and causes the picking station to exhibit even worse bottleneck behavior. The question then is whether personnel allocation and/or optimal batch sizing can ever counteract this shift in workload, to such an extent that SWP outperforms PAS.

The impact of a move from PAS to SWP on the operating characteristics of the system (average and variability of \( SU_p, X_p, SU_s \) and \( X_s \)) is hard to evaluate in general, as it will depend on the actual system being studied. To compare the performance of PAS with SWP, we assumed that \( E(X_s) \) decreases with \( \delta=30\% \) (i.e., 3 seconds) or \( \delta=50\% \) (i.e., 5 seconds) while \( E(X_p) \) increases accordingly with 3 seconds or 5 seconds. Setup time characteristics and SCVs of processing times were left unchanged. Table 2 compares the results for \( E(W)_{\text{opt}} \) obtained with the PAS policy (as shown in Figure 3) and the two SWP policies, for the VTWB and FTWB case. The first row in each cell refers to the VTWB result (including \( k_{\text{opt}} \)); the second row shows the FTWB result (including \( z_{\text{opt}} \)). The results for \( E(W)^*_{\text{opt}} \) (i.e., with optimal personnel allocation) for each scenario are indicated in bold. From this table, we observe the following:

(i) For small values of \( n_p \), the same optimal value of \( k \) (resp. \( z \)) and the same value of \( E(W)_{\text{opt}} \) is obtained regardless of \( n_{\text{tot}} \). This is true for both PAS and SWP. This is an intuitive result: the understaffing of the picking operation implies that the picking operation is the bottleneck, and hence constrains the system. Adding personnel without changing \( n_p \) is futile in such a setting: it merely leads to an unnecessarily high number of sorters, without impacting system performance.

(ii) Moving from a PAS policy to a SWP policy tends to leave the optimal value of \( k \) (resp. \( z \)) for any given workforce allocation unchanged, except in settings where \( n_p \) is small. In those cases, a move towards SWP increases the optimal value of \( k \) (resp. \( z \)). The higher the value of \( \delta \) (i.e., the higher the workload that is shifted from the sorting station to the picking station), the more pronounced this effect is. This is again intuitive: when \( n_p \) is small, the picking operation is the bottleneck operation, so the increase in \( X_p \) due to the SWP policy only further aggravates the bottleneck situation, and hence further constrains the system.

(iii) Only in those cases where the sorting area is understaffed (\( n_s=1 \)), the SWP policies may yield a lower optimal \( E(W) \) than the PAS policy. In these cases, sorting tends to be the bottleneck operation due to the
limited personnel. The reduction in $X$, caused by the SWP policy then allows to alleviate the bottleneck, and enables lower throughput times throughout the entire system. Note, however, that the shifting of workloads from sorting to picking should be done with caution: when too much workload is shifted ($\delta=0.5$), it may backfire, especially in cases where the total personnel is limited and the travel times for picking are large ($n_{tot}=4, m=20$).

(iv) When $n_p$ is small, PAS consistently outperforms SWP.

(v) As observed before, the optimal personnel allocations appear to be robust with regard to differences in warehouse size ($m$) and average customer order size ($E(O)$). Furthermore, they appear to be robust with regard to shifts in the relative workloads. The differences in $E(W)_{opt}$ between PAS and SWP are only slight. This indicates that, in spite of even rather significant shifts in workload, the achievement of workforce balance in our setting remains primarily determined by factors that do not depend on the policy used (e.g., picker travel times, setup times). As, in typical warehouse settings, picking and sorting times per order line are small compared to the travel times and setup times required, switching between a PAS and SWP policy will not significantly affect $E(W)_{opt}$.

7. CONCLUSIONS

In this paper we have provided an analytical approach for approximating the expected system throughput time, for online order systems with either variable time window batching (VTWB) or fixed time window batching (FTWB). The model includes arbitrary distributions for customer order size, picking time, sorting time, and setup times for picking or sorting a batch. The average throughput time of a customer appears to be a convex function of the number of customer orders grouped in a batch (in the case of VTWB), or the length of the time window (in the case of FTWB). This implies that, for both cases, an optimal batching policy (i.e., an optimal number of customer orders or an optimal time window) can be determined, which minimizes the expected customer order throughput time. Results from simulation experiments confirm that the model yields good approximations of the system behavior (in particular, optimal system behavior). As such, it can be used to determine the optimal batch sizes for a given workforce allocation, or to determine the optimal allocation of workforce to the picking and sorting processes. While it is likely that real-life picking operation will at times deviate from the optimal batching policy suggested by the models, these optimal policies can be considered as target policies for real-life operations.

The analysis has shown that, in both cases, the overall average throughput time is minimized if the workforce is allocated to the picking and sorting operations in such a way that the stations are approximately
balanced. Comparing SWP with PAS operations depends on the particular situation (i.e., ratio of set-up times to service times, and the increase in picking time versus the reduction in sorting time if we move from PAS to SWP). In our experiments we found that, in view of minimizing the average throughput time, SWP can only marginally outperform PAS for a given number of workers, when nearly all personnel is allocated to the picking operation.

The analytical model may be used as a decision support tool for management: it enables to numerically compare different system alternatives in a stochastic setting without having to resort to lengthy simulations. The model technically enables to include other storage strategies and/or routing policies; this would only impact the average and variance of the travel time component in the picking area. The main reason why such other combinations could not be investigated is the fact that the required expressions for the moments of travel time for other storage or routing schemes in 2-block warehouses do not exist yet in the literature. Developing such approximations remains a big challenge. Another extension of the model could be to include a zoned picking system with parallel pickers. In such a case, a customer order can be picked by multiple pickers simultaneously, and an order has to be assembled at the sorting and packing stations.

**APPENDIX A: APPROXIMATION FOR E(W_q) G/G/m (WHITT, 1993)**

Whitt (1993) provides the following approximation for the expected waiting time in general multiserver queues:

\[
E(W_q)_{G/G/m} = \phi(\rho, ca^2, cs^2, m) \left( \frac{ca^2 + cs^2}{2} \right) E(W_q)_{M/M/m} \]  

(A.1)

For a multiserver station consisting of \( m \) servers, the utilization is given by \( \rho = \frac{\lambda}{m \mu} \) with \( \lambda \) the arrival rate of entities at the station, and \( \mu \) the service rate of a single server at the station. The exact expression for \( E(W_q)_{M/M/m} \) is given by

\[
E(W_q)_{M/M/m} = \frac{P(N \geq m)}{\mu m (1 - \rho)} \quad P(N \geq m) = \left[ \frac{(m \rho)^m}{m! (1 - \rho)} \right] \zeta \quad \text{with} \quad \zeta = \left[ \frac{(m \rho)^m}{m! (1 - \rho)} + \sum_{k=0}^{m-1} \frac{(m \rho)^k}{k!} \right]^{-1}.
\]

\( P(N \geq m) \) denotes the probability that the number of customers in the system \( (N) \) exceeds the number of servers \( (m) \), and is equivalent to the probability that all servers are busy. The expression for \( \phi \) in (A.1.) is given by:

\[
\phi(\rho, ca^2, cs^2, m) = \begin{cases} 
\frac{4(ca^2 - cs^2)}{4ca^2 - 3cs^2} \phi_1(m, \rho) + \frac{c^2}{4ca^2 - 3cs^2} \psi(c^2, m, \rho) & \text{if } ca^2 \geq cs^2 \\
\frac{cs^2 - ca^2}{2(ca^2 + cs^2)} \phi_1(m, \rho) + \frac{cs^2 + 3ca^2}{2(ca^2 + cs^2)} \psi(c^2, m, \rho) & \text{if } ca^2 \leq cs^2
\end{cases}
\]
with $\gamma(m, \rho) = \min \left\{ 0.24, \frac{(1-\rho)(m-1)(4+5m)^{0.5}}{16m\rho} \right\}$, $\phi_1(m, \rho) = 1 + \gamma(m, \rho)$,

$\phi_2(m, \rho) = (1 - 4\gamma(m, \rho))e^{\frac{2(1-\rho)}{3\rho}}$, and

$\psi(c^2, m, \rho) = \begin{cases} 1 & c^2 > 1 \\ \phi_4(m, \rho)^{2(1-c^2)} & 0 \leq c^2 \leq 1 \end{cases}$

with

$c^2 = \frac{ca^2 + cs^2}{2}$, $\phi_4(m, \rho) = \min \left\{ 1, \frac{\phi_1(m, \rho) + \phi_3(m, \rho)}{2} \right\}$.

REFERENCES


Figure 1: Detailed flowchart of a 2-block warehouse with an S-shape picking route
Figure 2: Average throughput time $E(W)$ for a customer order (in minutes), along with its components, for the case with $E(Y)=60$ sec, $d=60$ sec, $m=20$, $E(0)=1.5$, $n_p=4$ and $n_s=2$: (a) in terms of $k$ for the VTWB case; (b) in terms of $z$ for the FTWB case.
Figure 3: Optimal throughput time $E(W)$ for the VTWB and FTWB case, obtained for different allocations of $n_s$ and $n_p$, for $n_{tot}=4$ (top panes) and $n_{tot}=8$ (bottom panes)
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>Exponential with average = 40, 60 or 90 seconds</td>
<td>$SU_p$</td>
<td>GAMM with average = 60 seconds, SCV = 2</td>
</tr>
<tr>
<td>$w_d$</td>
<td>6 seconds</td>
<td>$SU_s$</td>
<td>LOGN with average = 30 seconds, SCV = 1</td>
</tr>
<tr>
<td>$w_c$</td>
<td>10 seconds</td>
<td>$X_p$</td>
<td>LOGN with average = 8 seconds, SCV = 1</td>
</tr>
<tr>
<td>$d$</td>
<td>15, 30 or 60 seconds</td>
<td>$X_s$</td>
<td>GAMM with average = 10 seconds, SCV = 0.5</td>
</tr>
<tr>
<td>$(n_p,n_s)$</td>
<td>(6,3) or (4,2)</td>
<td>$m$</td>
<td>10 aisles or 20 aisles</td>
</tr>
<tr>
<td>$O$</td>
<td>$1+ Poiss(b)$, with $b=0.5$ or 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1**: Parameters for the validation experiment
Table 2: Comparison of $E(W)_{\text{opt}}$ (in minutes) and corresponding $k_{\text{opt}}$ (resp. $z_{\text{opt}}$) for different allocations of $n_p$ versus $n_s$, in the PAS and SWP scenario. The results for $E(W)^*_{\text{opt}}$ (with optimal workforce allocation) are indicated in bold.