Throughput Time Distribution Analysis for a One-Block Warehouse

Marc Schleyer*, Kevin Gue

Department of Industrial & Systems Engineering
Auburn University, AL 36830
USA

Abstract

We develop discrete time models for the throughput time distribution of orders arriving to a one-block warehouse. The models accommodate single- or multi-line orders, and we show how to use them to determine the optimal batch size, given a desired probability of on-time order fulfillment. Experiments suggest that the optimal batch size is slightly higher than one would choose if minimizing average throughput time.

Keywords: Warehousing, order picking, queueing system, discrete time techniques, service level

1. Thinking in Distributions

In the context of an order-fulfillment warehouse, service performance is determined mostly by whether or not orders are ready "on time," which often means "as soon as possible." In a system with random arrivals and random processing times, the throughput time takes on random times, and so understanding service performance requires one to compute the distribution of throughput times. Having the full distribution and not just the mean makes it possible to know, for example, what percentage of orders are processed in fewer than 3 hours, or what promised turnaround time will be met 95% of the time.

*Corresponding author. Tel.: +49 152 5338 5714

Email addresses: marc.schleyer@ifl.uni-karlsruhe.de (Marc Schleyer), kevin.gue@auburn.edu (Kevin Gue)

To understand the motivation for our study, consider a distributor of industrial supplies whom we have visited. The central warehouse serves many retail outlets, but it also operates a will-call service, with which it serves local contractors. Contractors call throughout the day with orders, which are picked by one or two workers specifically dedicated to this service. Customers receive faster service than if they waited on a delivery truck, which must batch orders to provide transportation economies. The warehouse would like to know, as a matter of policy, how long it should tell customers to wait before arriving to pick up their orders from the will-call service. Answering this question requires knowledge of the distribution of throughput times.

We develop discrete time queueing models to calculate the throughput time distribution of a customer order in a one-block warehouse system. We adapt approaches originally developed for the evaluation of Asynchronous Transfer Mode (ATM) networks (see Ackroyd, 1980; Tran-Gia, 1996; Hübner and Tran-Gia, 1995; Hasslinger and Klein, 1999) in order to describe material and information flows in a material handling system. The advantage of discrete time queueing analysis is that, unlike in most continuous time models, performance measures such as the waiting time distribution of an order in queue can be calculated efficiently under general distribution assumptions (see Grassmann and Jain, 1989). Schleyer (2007), Schleyer (2010), and Schleyer and Furmans (2007) develop a framework for stochastic modeling of material and information flows in discrete time. Our work differs from these in that we describe models specifically for picking processes in a warehouse.

The existing warehousing literature has proposed queueing approaches in continuous time for analytical modeling. If input parameters such as the service time (e.g., picking time in order picking systems) are generally distributed, only mean values for system performance have been produced (Chew and Tang, 1999; Jewkes et al., 2004; Le-Duc and de Koster, 2007; Van Nieuwenhuyse and de Koster, 2009). Previous research on the estimation of expected travel time in one-block warehouses has been carried out by Kunder and Gudehus (1975), Hall (1993), Jarvis and McDowell (1991), Chew and Tang (1999), and Roodbergen and Vis (2006). Unfortunately, mean values for performance do not answer the service-oriented questions we are considering. Petersen (2000) and Petersen et al. (2004) determine the probability of on-time order fulfillment using simulation. Their goal is to choose warehouse operating parameters (batch sizes, wave length, etc.) with respect to a defined level of service. By contrast, the models we present below are analytical and fast, which permits extensive "what-if" analysis. Our models also produce throughput time distributions, rather than point estimates of performance for particular configurations.

In this paper, we analyze the batching of order lines before picking is performed, a problem known in the literature as the order batching problem (OBP). Many authors have modeled the OBP with a known set of orders (e.g., Rosenwein, 1996; de Koster et al., 1999; Gademann and van de Velde, 2005). Three publications have considered the stochastic nature of the OBP in warehouses. Chew and Tang (1999) determine the average throughput time for the OBP in a 1-block warehouse using an $E_n/G/c$ queueing system, where n denotes the batch size. They make the conservative assumption that the average throughput time of an arbitrary order is equal to the throughput time of the first order in a batch, which has the longest batching time. Below, we model order batching time explicitly. Le-Duc and de Koster (2007) model the OBP as an $M/G^n/1$ queueing system and derive the average throughput time in a 2-block warehouse. Le-Duc and de Koster (2007) assume the arrival of single-line orders. Van Nieuwenhuyse and de Koster (2009) extend the model of Le-Duc and de Koster (2007) to study the average customer order throughput time with time window batching. The arrival of multi-line orders is allowed. They assume that the arrival process of customer orders is Poisson. We present in this paper models for the calculation of the throughput time distribution for both multi-line and single-line orders. Our research is not restricted to the assumption that the arrival process is Poisson. We require only a stationary arrival stream of orders, which might represent arrivals during the heavy traffic period of a system.

Below, we explain the objective of our work and describe the model, including assumptions and variables. In Section 3 we build a model for single-line orders, and show that the optimal batch size for a high service level is slightly larger than the optimal batch size for mean throughput time. In Section 4 we extend our models to accommodate orders with multiple lines and demonstrate the same result for this case. We offer conclusions in Section 5.

2. Objective, Model Description and Assumptions

The objective of this paper is to present a new methodology to analyze order fulfillment processes in a warehouse. Our focus is on the service performance of a warehouse, by which we mean the probability of on-time

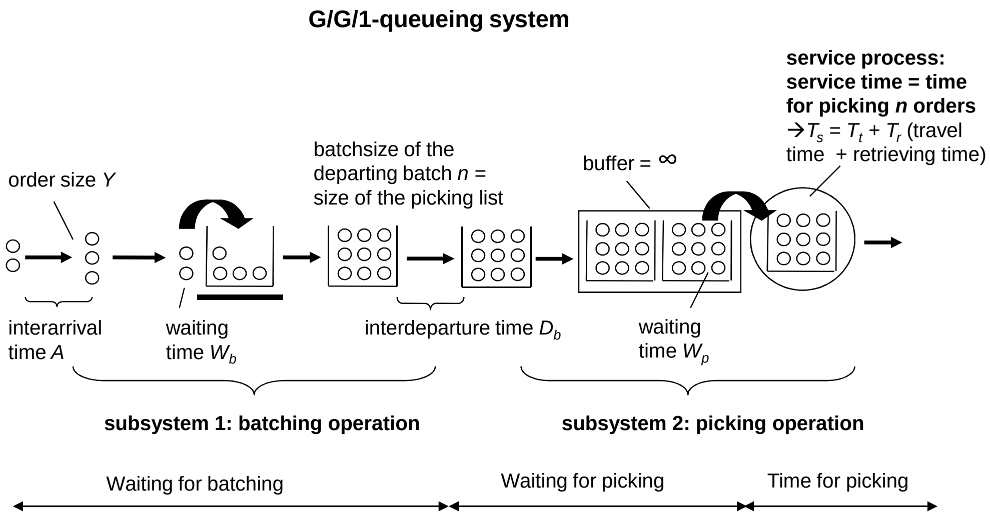

Figure 1: Queueing model of the picking system.

fulfillment of a costumer order, rather than on its throughput capacity. We believe ours is the first research that uses discrete time queues to model picking processes in a warehouse. We illustrate our techniques with a basic warehouse model in which a single order picker traveling through a one-block warehouse retrieves items using the well known S-shape routing strategy.

Consider a single-worker picking operation for a one-block, rectangular warehouse with M aisles in which items are randomly stored. The overall process is illustrated in Figure 1: Arriving orders are gathered into batches of size n before entry into a single-server queue, from which an order picker processes the batches (in our case, in a picking tour). The stochastic model for this picking system is comprised of two subsystems. The first describes the batching process, in which a batch of n order lines is formed. The second describes the picking operation itself, where a batch of n order lines is processed in one picking tour. The overall throughput time of a customer order consists of three components: (1) the waiting time due to batching W_b, (2) the waiting time of the batch when the picker is busy W_p, and (3) the service time to pick a batch of orders by the picker T_s.

The service time to pick a batch, T_s, consists of the travel time through the warehouse, T_t, and the time for retrieving a batch of n order lines from the shelves, T_r. The complete picking operation can be modeled using a $G/G/1$ queue in the discrete time domain. Both the arrival and the service process are described by independent and identically distributed (iid) random variables.

The batch size processed by the picker is n. We assume that the distribution of retrieval times, T_r, can be derived from empirical data, and that an order line can consist of one or more items. More specifically, the retrieval time distribution reflects time spent stopped at a *picking location*, which comprises multiple storage locations for products (both sides of the aisle and potentially multiple levels of storage).

More difficult is the derivation of the travel time, T_t, which depends on the pick locations in the warehouse and on the routing method, which in our case is the S-shape strategy. We assume that pick locations are randomly distributed in the warehouse, which means that each pick location in the warehouse has the same probability of being visited.

We assume a discrete time domain, which allows us to calculate the throughout time of customer orders when interarrival and service times are generally distributed. The length of a time unit should be short enough to yield a reasonable approximation of continuous time, but not so short that it causes computational difficulties. (We give examples below.)

We define the following variables:

$\vec{a}$ the (discrete) probability mass function for interarrival time,

a_i probability that an interarrival time takes on i time units, $i = 1, 2, 3, ..., max(a)$,

$\vec{y}$ the (discrete) probability mass function for the number of order lines in a customer order,

y_i probability that the number of order lines in a customer order is i, $i = 1, 2, 3, ..., max(y)$,

n batch size, the number of order lines which have to be picked in one picking tour,

M number of aisles,

N number of pick locations per aisle,

d time units to traverse an aisle (can be a continuous value),

w time units to travel the center-to-center distance between two adjacent aisles,

u time units to pick an order line,

$w_{b,i}$ probability that the waiting time of a customer order at the batching node is i time units, $i = 0, 1, 2, ...,$

$w_{p,i}$ probability that the waiting time of a customer order when the picker is busy is i time units, $i = 0, 1, 2, ...,$

$d_{b,i}$ probability that the interdeparture time of formed batches from the batching node is i time units, $i = 0, 1, 2, ...,$

T_s random variable describing the service time to process a batch order of size n by the picker,

T_t random variable describing the travel time through the warehouse for picking n order lines,

T_r random variable describing the time for retrieving a batch of n order lines from the shelves,

T_v random variable describing the picking sojourn time,

T_c random variable describing the total throughput time,

$\sigma_{0.95}(X)$ 95th percentile of a random variable X,

$\otimes$ convolution operator: The distribution of the sum of two independent nonnegative random variables X and Y, is called the convolution of their distributions and can be computed by $Z = X \otimes Y$ such that $z_i = \sum_{j=0}^{i} x_j y_{i-j}$. (Variable z_i is the probability that random variable Z takes on discrete value i, which is the sum of j and $i - j$. Variable x_j is the probability that random variable X takes on discrete value j, and y_{i-j} is the probability that random variable Y takes on discrete value $i - j$.)

$\{\vec{x}^{m\otimes}\}_i$ probability that the sum of m random variables, each described by pmf $\vec{x}$, is i time units.

Our analysis works as follows: First, we model the order picking process in a one-block warehouse. We derive the order picker travel time T_t, which leads to the service time to pick a batch T_s. Second, we analyze the batching

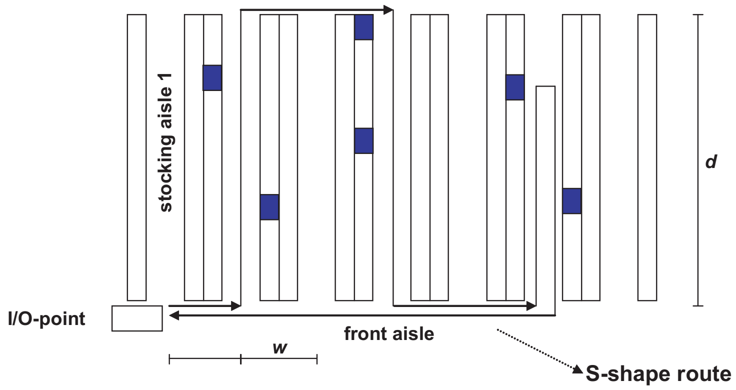

Figure 2: Warehouse with an S-shape pick route.

operation by computing the waiting time of an arbitrary order for batching, W_b, and the interdeparture time of formed batches, D_b. Given D_b and T_s, the waiting time for picking, W_p, can be determined using the discrete time $G/G/1$ queueing model by Grassmann and Jain (1989) (see Appendix 6). We then calculate the distribution of the total throughput time, T_c, by convolving the distributions of W_b, W_p, and T_s, assuming the independence of these random variables.

3. Analysis

3.1. Travel Time Analysis

Consider the warehouse layout depicted in Figure 2. The warehouse consists of a block of M aisles in which items are stored. The order picker traverses the warehouse using the S-shape strategy, which is quite efficient and, due to its simplicity, widely used in practice (Roodbergen et al., 2008). In an S-shape routing strategy, any aisle containing at least one item is traversed in its entirety. An exception is made if the number of aisles to be visited is odd, in which case the picker turns after he or she has picked the final item (see Figure 2).

We assume the order picker travels at constant speed, so travel time is directly proportional to the travel distance. The travel time consists of two components, the time spent within aisles and the time spent crossing between aisles. Both times depend on the number of aisles visited, which can

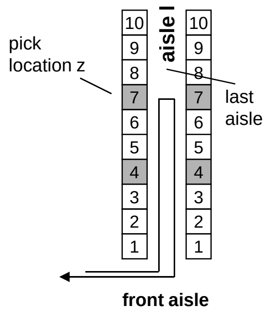

Figure 3: Farthest pick in the last aisle

be determined with a combinatorial approach given by Kunder and Gudehus (1975) or by solving the classical occupancy problem (Chew and Tang, 1999; Le-Duc and de Koster, 2007). In the latter problem, a fixed number of indistinguishable balls are randomly placed in a given number of distinguishable urns (see Johnson and Kotz, 1977). Treating the urns as aisles and the balls as order lines to be picked, the distribution of the number of aisles to be visited in a picking tour can be determined using Johnson and Kotz's formula,

$$P(x) = \frac{1}{M^n} \binom{M}{x} \sum_{i=0}^{x} (-1)^i \binom{x}{i} (x-i)^n \qquad x = 1, ..., \min(n, M). \quad (1)$$

If x is even, the within-aisle travel time is $d \cdot x$. If x is odd, the picker must traverse $x-1$ entire aisles. When the picker walks into the last aisle and picks the last order line, he or she must turn, in which case we must determine the distance from the center line of the front aisle to the location of the last item to be picked (see Figure 2). This distance depends on the number of picks, y, which must be performed in the last aisle. Under the condition that x aisles have to be visited, the probability that y picks are located in the last aisle is

$$P(y|x) = \binom{n-x}{y-1} \left(\frac{1}{x}\right)^{y-1} \left(1 - \frac{1}{x}\right)^{n-x-(y-1)}, \qquad y = 1, ..., n-x+1. \quad (2)$$

Equation 2 can be explained as follows: There must be at least one pick in each of the x aisles to be visited. At first, let us allocate exactly one pick to each of these x aisles. One of them is the last aisle to be visited, and there are $n-x$ picks remaining to be allocated to these x aisles. When one aisle is chosen out of x aisles, the last aisle is chosen with probability $1/x$ and one of the others with probability $1 - 1/x$. After the remaining $n - x$ picks have been allocated, there can be maximum $y = n - x + 1$ picks in the last aisle. Equation 2 follows from basic probability.

In each aisle there are N potential pick locations, each representing multiple storage locations on either side of the aisle. In the case the picker has to turn, the distance from the center line of the front aisle to the farthest pick in the aisle is a function of y, the number of picks in the final aisle. Let $P(z|y)$ be the probability that the farthest pick is at the zth location, given there are y picks in the aisle (see Figure 3 for an illustration). For example, if $y = 1$, the picker travels to only one location in the aisle and $P(z|y) = 1/N$, for all locations z. If $y > 1$, the $y - 1$ pick locations which are located closer to the center line must be chosen from the $z - 1$ closer located pick locations. To compute $P(z|y)$ in this case, we divide the number of ways to distribute $y - 1$ picks among these $z - 1$ potential locations by the total number of ways to distribute y picks in N locations,

$$P(z|y) = \frac{\binom{z-1}{y-1}}{\binom{N}{y}}, \qquad z = y, ..., N. \tag{3}$$

The probability that the number of aisles to be visited is even is easily calculated by

$$P_{even} = \sum_{x=0,\ even}^{\min(n,M)} P(x). \tag{4}$$

The probability that the number of aisles to be visited is odd and therefore that the picker has to turn is $P_{odd} = 1 - P_{even}$.

To derive the cross aisles travel time, we must determine the distance from the I/O-point to the farthest aisle to be visited (see Figure 2). We number the aisles from $\ell = 1$ for the closest aisle to $\ell = M$ for the farthest aisle from the I/O-point. $P(\ell|x)$ is the probability that ℓ is the farthest aisle given that x aisles have to be visited. $P(\ell|x)$ can be derived in a manner analogous to Equation 3. We get

$$P(\ell|x) = \frac{\binom{\ell-1}{x-1}}{\binom{M}{x}}, \qquad \ell = x, ..., M. \tag{5}$$

Because we model in the discrete time domain, we must define the time scale. One time unit could be a second, some seconds, one minute, etc., depending on the application. For most warehouse applications, a time unit of 5–20 seconds should provide sufficient resolution and still make the model computationally tractable. The results of the model are sensitive to selection of the time scale, of course: more precise results can be had with smaller time units, at the cost of longer computational times.

It takes d time units to traverse an entire aisle. If the picker has to turn at the zth location in an aisle, the travel time within this aisle is $2dz/N$ time units. Given x (x is odd), ℓ, and z, the travel time is $2w(\ell-1)+d(x-1)+2dz/N$ time units. Because we require a discrete value for time, we use notation $\lfloor 2w(\ell-1)+d(x-1)+2dz/N\rceil$ to indicate the nearest integer to $2w(\ell-1)+d(x-1)+2dz/N$. For the overall travel time it follows that

$$\begin{aligned} P(T_t =&\lfloor 2w(\ell-1)+dx\rceil | x \text{ even}) = \\ &\frac{1}{P_{even}} \sum_{x=2,\ even}^{\min(n,M)} \sum_{\ell=x}^{M} P(\ell|x)P(x) \\ P(T_t =&\lfloor 2w(\ell-1)+d(x-1)+2dz/N\rceil | x \text{ odd}) = \\ &\frac{1}{P_{odd}} \sum_{x=1,\ odd}^{\min(n,M)} \sum_{\ell=x}^{M} \sum_{y=1}^{n-x+1} \sum_{z=1}^{N} P(\ell|x)P(y|x)P(z|y)P(x). \end{aligned} \tag{6}$$

Using Bayes' formula we get the probability that the travel time is i time units by

$$P(T_t = i) = P(T_t = i|\text{ even})P_{even} + P(T_t = i|\text{ odd})P_{odd}. \tag{7}$$

This approach for the travel time distribution is exact in the discrete time domain, and the computation time is fast: a problem of size $M = 20$, $N = 50$, and $n = 20$ requires just 0.07 seconds on an AMD 2 GHz processor. In the cases $N = 100$ and $N = 150$, run times are 0.33 and 1.3 seconds.

The picking service time, T_s, is the sum of T_t (travel time) and T_r (time for retrieving items). T_r can be derived directly from order history data (Petersen, 2000). For example, if after a picker has reached a storage location, it takes x time units to retrieve a single item and place it in the picking cart, each additional item at the same storage location might require $y < x$ time units. The distribution of T_s is given by the convolution of the distributions of T_t and T_r.

3.2. *Analysis of the Batching Process for Single-Line Orders*

As illustrated in Figure 1, there is a batching process before the picking process begins. The batching process requires that we determine the number

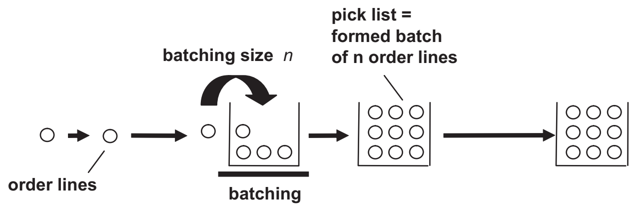

Figure 4: Batching of single-line orders.

of order lines in a picking tour. If too few order lines are picked in each picking tour, system capacity is reduced due to workers spending so much of their time traveling. Reduced capacity could lead to excessive queueing times for orders or, in the worst case, to an infeasible system (the so-called saturation effect, see Karmarkar et al., 1985; Van Nieuwenhuyse and de Koster, 2009). If order batches are too large, then orders could spend excessive time in queue waiting for other orders to complete the batch, also leading to long queues and high throughput times. This effect is called the batching effect by Karmarkar et al. (1985) and Van Nieuwenhuyse and de Koster (2009). An optimal solution to the order batching problem makes the best tradeoff: sufficient orders to make picking tours efficient, without causing arriving orders to wait too long to form batches.

3.3. *Distribution of Waiting Time for Batches*

To begin, we derive the waiting time distribution of an arbitrary customer order at the batching node in a picking system when single-line orders arrive. We represent interarrival times with a probability mass function $\vec{a}$, in which an element a_i corresponds to the probability that the interarrival time is i time units.

Suppose an arriving order line at the batching node encounters $0 \leq x < n$ waiting order lines. If $x < n - 1$, the order line must wait for the missing $n-x-1$ order lines. If $x = n-1$, the order line does not have to wait, because its arrival completes the batch, which may be transferred immediately to the picking station. The probability that an order line encounters x order lines

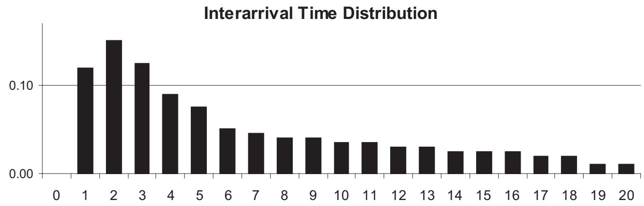

Figure 5: Interarrival time distribution.

upon arrival is $1/n$. Thus, the waiting time distribution at the batching node W_b is

$$w_{b,i} = \begin{cases} \frac{1}{n} \sum_{j=1}^{n-1} \{\vec{a}^{(n-j)\otimes}\}_i & \text{if } i > 1, \\ \frac{1}{n} & \text{if } i = 0. \end{cases} \quad (8)$$

Interdeparture times from the batching process are, by definition, interarrival times of formed batches to the picking station that follows it. We represent the interdeparture times of formed batches with a discrete random variable D_b, having probability mass function $\vec{d}$ in which an element $d_{b,i}$ represents the probability that an interdeparture time is i time units.

For a batch size of n, we get $d_{b,i}$ easily by the n-fold convolution of $\vec{a}$, because n arrivals are required to fill the departing batch. Therefore, we obtain

$$d_{b,i} = \{\vec{a}^{n\otimes}\}_i. \quad (9)$$

To summarize, we have built discrete time distributions for the waiting time due to batching, for the waiting time due to picker utilization, and for the picking service time. The throughput time distribution is their convolution: $T_c = W_b \otimes W_p \otimes T_s$.

3.4. Performance Analysis

Consider an interarrival time pmf $\vec{a} = \{0.0, 0.120, 0.150, 0.125, 0.090, 0.075, 0.050, 0.045, 0.040, 0.040, 0.035, 0.035, 0.030, 0.030, 0.025, 0.025, 0.025, 0.020, 0.020, 0.010, 0.010\}$, where the first element is a_0 (see Figure 5). The warehouse has $M = 20$ aisles, each of which has $N = 50$ pallet locations. The time to walk an entire aisle (one-way) is $d = 3$; the distance between aisle centers is $w = 1$; the time to make a single retrieval is a fixed $u = 0.25$; and the batch size $n = 12$. If the retrieval time for a batch is non-integer, we round to the nearest integer. In this case, for example, if n had been 11, we would round $11(0.25) = 3.75$ to 4. Figure 6 shows the throughput time distribution of a customer order.

This distribution gives us the probability that a customer order can be processed within a given time. In the example, an order request can be processed within 158 time units with probability 95 percent.

The spikes in the left part of the distribution correspond to orders for which travel time makes up the large majority of the throughput time. The irregularity of the travel time distribution is an artifact of the S-shape routing protocol and the nature of discrete time models, which in this case makes it more likely that throughput time is an even number. For higher values of throughput time, waiting time due to batch forming and due to high picker utilization tend to overwhelm the travel time, and the curve is more smooth.

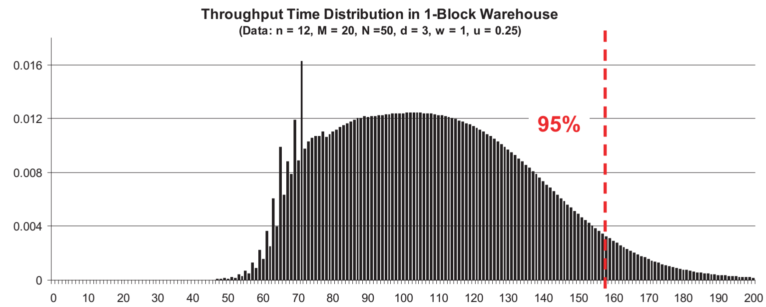

Figure 6: Throughput time distribution for a customer order in a 1-block warehouse.

3.4.1. Comparison with simulation results

To evaluate the accuracy of the discrete time approach we compared the analytical solutions with simulation experiments. The simulation to determine the throughput time distribution processed 2.5 million order batches. The interarrival time distribution was assumed to be negative exponential (a continuous distribution). For the analytical calculation we used a discrete time interarrival time distribution with the same mean as the simulation model (see an example in Figure 14, in Appendix 7). In real life applications the discrete time interarrival time distribution would be derived from empirical data. The following data were used both for the simulation and the discrete time calculation: The warehouse has 20 aisles, each having 50

			Utilization							
			80%			85%			90%	
			Throughput time			Throughput time			Throughput time	
		Calc	Sim	Deviation	Calc	Sim	Deviation	Calc	Sim	Deviation
Batch size	8	Mean	98.9	97.2	1.67%	101.5	101.3	0.20%	109.8	112.2	-2.13%
		85%-P	132	126	4.76%	135	131	3.05%	147	148	-0.68%
		90%-P	141	134	5.22%	144	140	2.86%	160	162	-1.23%
		92.5%-P	147	140	5.00%	151	147	2.72%	169	172	-1.74%
		95%-P	154	147	4.76%	160	155	3.23%	181	187	-3.21%
		97.5%-P	166	157	5.73%	174	170	2.35%	202	212	-4.72%
Batch size	12	Mean	111.9	110.8	1.00%	112.9	113.1	-0.14%	117.8	120.6	-2.33%
		85%-P	146	142	2.82%	146	144	1.39%	153	154	-0.65%
		90%-P	154	149	3.36%	155	152	1.97%	163	164	-0.61%
		92.5%-P	160	155	3.23%	159	157	1.27%	170	171	-0.58%
		95%-P	167	161	3.73%	168	164	2.44%	179	182	-1.65%
		97.5%-P	177	172	2.91%	180	176	2.27%	194	201	-3.48%
Batch size	20	Mean	133.0	131.8	0.92%	132.3	132.1	0.15%	134.6	136.1	-1.15%
		85%-P	171	168	1.79%	169	166	1.81%	171	170	0.59%
		90%-P	180	175	2.86%	177	174	1.72%	179	178	0.56%
		92.5%-P	185	181	2.21%	182	179	1.68%	185	184	0.54%
		95%-P	192	187	2.67%	189	185	2.16%	193	191	1.05%
		97.5%-P	202	198	2.02%	199	195	2.05%	205	204	0.49%

Table 1: Simulation results are compared to results obtained by discrete time queueing analysis

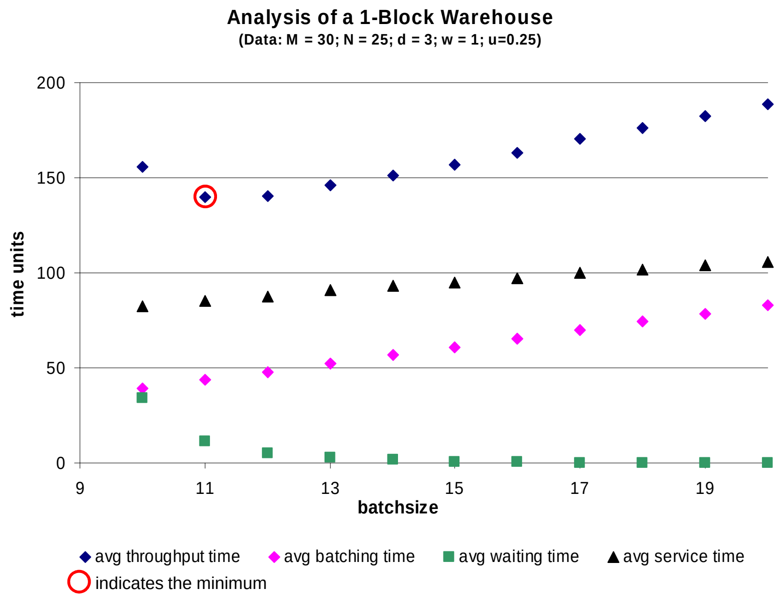

Figure 7: Mean throughput time, mean batching time, mean waiting time, and mean service time for an arbitrary customer order in a 1-block warehouse.

pallet locations. The speed of the picker is constant with 3 time units to walk an entire aisle (one-way), and 1 time unit to walk the distance between aisle centers. The time to make a single retrieval is a fixed $u = 0.25$. We did several experiments with different mean interarrival times resulting in a different system utilization (80%, 85% and 90%) and with different batch sizes (8, 12, and 20).

Table 1 shows a comparison of the simulation and discrete time queueing analysis. The discrete time calculation results match the simulation for most experiments. Deviations are likely due to the differences in the used interarrival time distributions (continuous vs. discrete distribution; see Figure 14 in the Appendix) and the assumed independence of W_b, W_p and T_s in our discrete time approach. In general, the model is more accurate when batch size is large and utilization is high. The computing times for the discrete time calculation take between 0.2 seconds and 3 seconds, depending on batch size and system utilization. The simulations took 15–25 minutes per run.

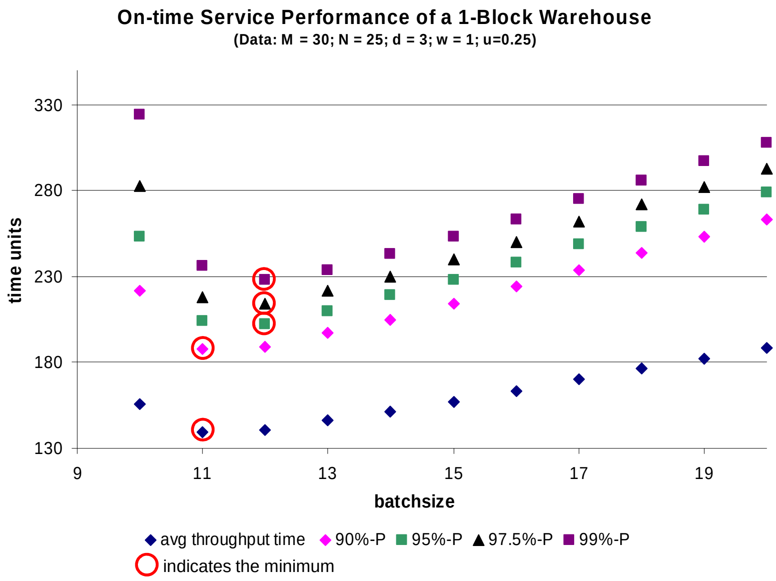

Figure 8: Optimal batch size in a 1-block warehouse.

3.4.2. Batch Size Optimization

In Section 3.2 we described the order batching problem which makes the best tradeoff between the saturation and batching effect.

Figure 7 illustrates the tradeoff in a second example, with $M = 30, N = 25, d = 3, w = 1$, and $u = 0.25$. The interarrival time distribution has pmf $\vec{a} = \{0.0, 0.05, 0.1, 0.125, 0.08, 0.07, 0.055, 0.05, 0.045, 0.04, 0.04, 0.04, 0.035, 0.035, 0.03, 0.03, 0.025, 0.025, 0.02, 0.02, 0.02, 0.015, 0.015, 0.015, 0.01, 0.01\}$.

The batch size that minimizes $E(T_c)$ is 11 (Figure 8); whereas the batch size that minimizes $\sigma_{0.95}(T_c)$ is 12. In several experiments we observed that the optimum batch size to ensure a high service level is slightly higher than the batch size that minimizes the mean throughput time. How can we explain this?

Figure 9 illustrates the impact of increasing n on the sum of waiting and service times (which we call the picking sojourn time T_v) for the same system we consider in Figure 8. For low values of n, the difference between $E(T_v)$ and $\sigma_{0.9}(T_v)$, $\sigma_{0.95}(T_v)$, $\sigma_{0.975}(T_v)$, or $\sigma_{0.99}(T_v)$ is high, and it decreases quickly with increasing n. There are two reasons for this behavior: First, as the batch size increases, the numbers of aisles visited per tour and therefore the picking travel times become more stable. In fact, for large batches it is likely that

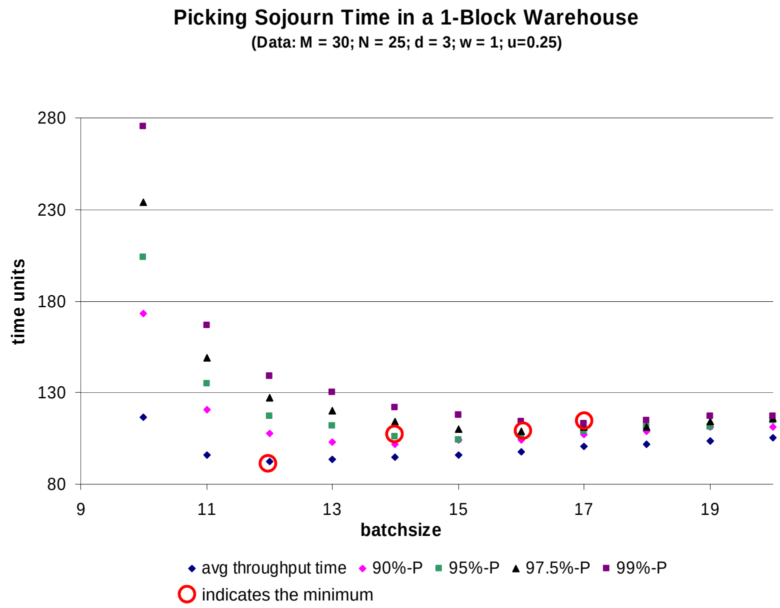

Figure 9: Optimal batch size to minimize the sojourn time of a picking system.

all aisles would visited, with the result that travel times are nearly identical (the only difference being due potentially to partial travel of the final aisle). Reduced variation in the picking travel times leads to lower variation of service times, and therefore to a lower variation of the picking sojourn times. Second, as the batch size increases, the picker utilization is lower. This leads to less waiting before picking, and especially to less variation of it. Again, an increasing batch size "tightens up" the distribution of the picking sojourn times, which draws the higher percentiles of the distribution closer to the mean (see again Figure 9). The effect of "tightening up" due to increasing batch size gives us different optimal batch sizes for the mean picking sojourn time and its high percentiles. For example, in Figure 9 the optimal batch size for $E(T_v) = 12$, and the optimal for $\sigma_{0.99}(T_v) = 17$. However, because T_b increases with n, the difference between the optimal n in a mean value analysis and the optimal value in a service level analysis is only small if the total throughout time of a customer order in the system is considered (see again Figure 8).

In summary, a service level analysis leads to a slightly greater value of n than a mean value analysis.

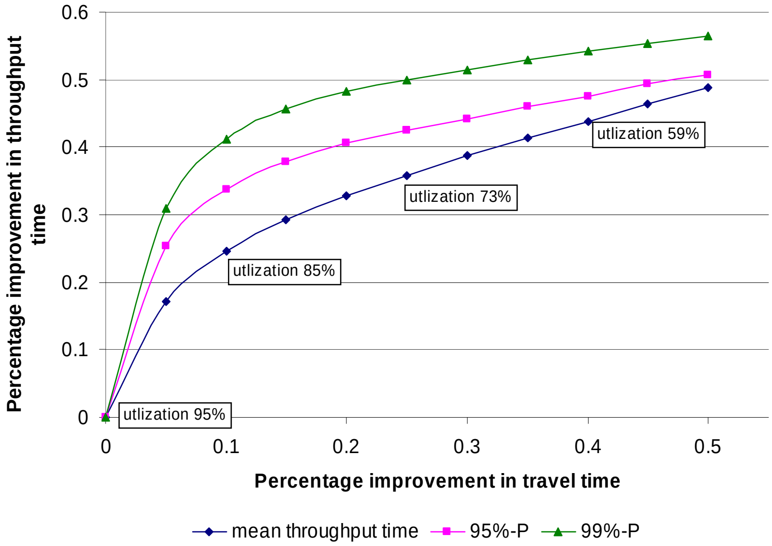

Figure 10: Percentage improvement of the throughput time depending on the percentage reduction in the travel time.

Two more observations: First, notice that at the optimal batch size, the difference between the mean throughput time and the 99th percentile throughput time is less than that difference for smaller *and* greater batch sizes (see Figure 8). In other words, the distribution of throughput times seems to be "tightest" at the optimal batch size, and therefore is most predictable there. Second, the graphs for the throughput time are flatter to the right of the optimal batch size. This means that a greater than optimal batch size has less negative impact on performance than a less than optimal batch size.

3.5. Applications

To illustrate the potential use of discrete time models of a warehouse system, suppose a firm is considering a new technology, such as a faster vehicle type, that would effectively increase the speed of an order picker. What happens to the throughput time distribution? Figure 10 illustrates the percentage improvement of the throughput time depending on the percentage reduction in the travel time. For this example, we assume $M = 20, N = 50, d = 3, w = 1, n = 12, u = 0.25$, and the interarrival time distribution has pmf $\vec{a} = \{0.0, 0.120, 0.150, 0.125, 0.090, 0.075, 0.050, 0.045, 0.040, 0.040, 0.035,$

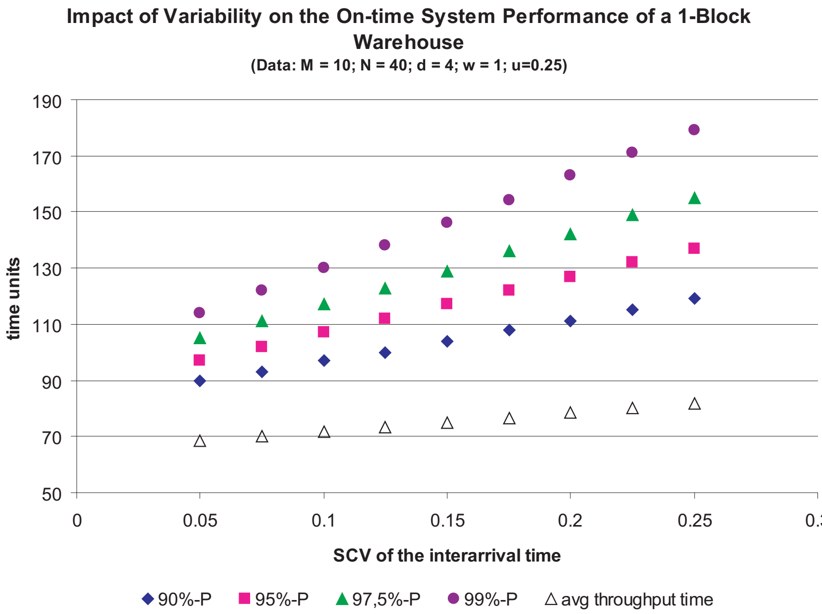

Figure 11: Impact of the variability of the interarrival process on throughput time.

$0.035, 0.030, 0.030, 0.025, 0.025, 0.025, 0.020, 0.020, 0.010, 0.010\}$. We increase the speed of the order picker stepwise by 5%, which leads to a decreasing utilization of the system. As we describe above, decreasing utilization reduces more the high percentiles of waiting time before picking than it does the mean value. Therefore, at a high utilization the positive effect of travel time reduction on the improvement of service level is considerably greater than on the improvement of mean throughput time.

For a second illustration, we performed a numerical analysis of the mean order throughput time $E(T_c)$ and several percentiles of throughput time with different levels of SCV of the interarrival time (for the data of the arrival process, see Table 2). Figure 11 shows that percentiles such as $\sigma_{0.95}(T_c)$, $\sigma_{0.975}(T_c)$, and $\sigma_{0.99}(T_c)$ vary much more strongly with increasing interarrival time variability than does $E(T_c)$, suggesting that the consequences of increasing variabilities on on-time order fulfillment will be underestimated if decisions are based on mean values.

4. A Model for Multi-Line Orders

In many applications, customers order several lines instead of just one. Analysis for this case is similar to the single-line order case, except for the batching process, which now must account for multiple lines arriving at the same time in an order.

4.1. Batching Process

We model the number of lines in a customer order with an iid discrete random variable Y, having probabilities y_i, $i = 1, ..., max(y)$ that an order contains i items. There are two possible policies: In the first, a batch of exactly n lines must be formed, and so it is possible that a customer order containing several order lines would be split between batches. For example, if $n = 20$ and 18 order lines have already arrived, an arriving order with 4 lines would have to be split between two picking tours. Splitting orders is undesireable because it requires a merging operation downstream, which could cost labor and throughput time. Therefore, we assume that batches must contain *at least* n items; our policy is to release a batch upon the first order arrival making the batch size greater than or equal to n.

Here we investigate the interdeparture time between batches, the waiting time of an arbitrary customer order, and the size of a formed batch for such a policy.

4.1.1. Interdeparture Time Distribution

The interdeparture time of formed batches depends on the number of arrivals required to fill a batch to at least n order lines. Let $P(N_a = k)$ be the probability that the number of customer orders in a batch of at least n is k, $k = 1, 2, ...$ Then,

$$P(N_a = k) = \begin{cases} \sum_{j=n}^{max(y)} y_j & \text{if } k = 1, \\ \sum_{i=1}^{n-k+1} \{\vec{y}^{(k-1)\otimes}\}_{n-i} \sum_{j=i}^{max(y)} y_j & \forall k > 1. \end{cases} \tag{10}$$

If $k > 1$, the first part of Equation (10) describes the probability that $k - 1$ arrivals fill the batch to $n - i$ order lines. In this case, the kth order arrival must contain at least $Y = i$ order lines.

The interdeparture time can be calculated by the k-fold convolution of $\vec{a}$, weighted with probability $P(N_a = k)$.

$$d_{b,i} = \sum_{k=1}^{k_{max}} P(N_a = k)\{\vec{a}^{k\otimes}\}_i \qquad k_{max} = \left\lceil \frac{n}{i_{min}} \right\rceil, \tag{11}$$

where i_{min} is the smallest order size having a non-zero probability of arriving.

4.1.2. Batch Size Distribution

The batch size must be at least of size n, so we must compute the probability that the number of collected order lines exceeds n by i, $i = 0, 1, ..., max(y) - 1$. If k arrivals are required to collect at least n order lines, the batch size distribution is

$$y_{b,i} = \begin{cases} y_i & \text{if } k = 1 \\ \sum_{k=2}^{k_{max}} \sum_{j=1}^{n-1} \{\vec{y}^{(k-1)\otimes}\}_{n-i} y_{i-(n-j)} & \forall k > 1, \end{cases} \quad (12)$$

where $i = n, n+1, ...$ Equation 12 consists of two components. First, we combine $k - 1$ arriving customer orders such that the number of order lines is $n - j$, $j = 1, ..., n - 1$. Then, the size of the kth customer order must be exactly $i - (n - j)$ in order to form a batch of size i.

4.1.3. Waiting Time Distribution

We derive the waiting time of arriving orders assuming that k arrivals are required to form a batch. If $k > 1$, the order lines of the first arriving customer order have to wait for $k - 1$ arrivals. In contrast, the order lines of an arriving customer order which fill the batch to at least n order lines do not have to wait at all. In general, order lines in the ith arrival, $i < k$, must wait for $k - i$ more arrivals for the batch to be completed. For these order lines, the waiting time distribution is the $(k-i)$-fold convolution of the interarrival time distribution. Thus, we must consider the probability that an arbitrary order line belongs to the ith arrival under the condition that $k > i$ arrivals are required to form a batch.

Define $E(Y|N_a = k)$ as the mean size of customer order $1, ..., k-1$ (mean number of order lines) given that $k > 1$ arrivals are required to fill a batch. If exactly two customer orders arrive to form a batch ($k = 2$), we get

$$E(Y|N_a = 2) = \frac{1}{P(N_a = 2)} \left(\sum_{j=1}^{n-1} y_{n-j}(n-j) \sum_{h=j}^{max(y)} y_h \right). \quad (13)$$

The size of the first customer order is $n - j$, and the size of the second is at least j. Equation 13 gives us the expected size of the first customer order. If $k > 2$, we get

$$E(Y|N_a = k) = \frac{1}{P(N_a = k)} \left(\sum_{x=1}^{n-1} y_x x \sum_{j=1}^{n-x-1} \{\vec{y}^{(k-2)\otimes}\}_{n-j-x} \sum_{h=j}^{max(y)} y_h \right). \quad (14)$$

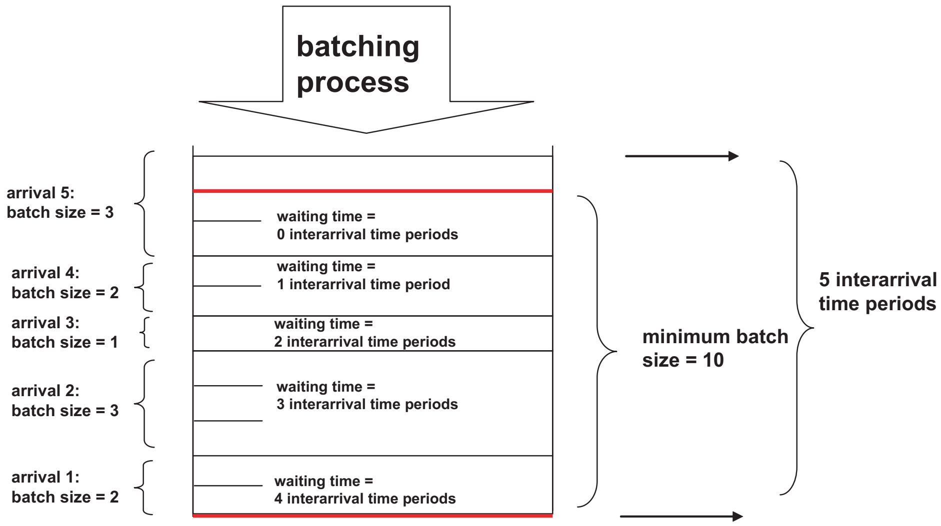

Figure 12: Example of a batch forming process with $k = 5$ arrivals to fill a batch of at least size $n = 10$. Illustrated is the waiting time of arriving customer orders as multiples of the interarrival time.

The size of the first order is x, and Equation 14 therefore delivers the expected size of the first customer order. However, because Y is iid, $E(Y|N_a = k)$ represents also the expected size of all customer orders from 1 up to $k-1$.

The quotient $E(Y|N_a = k)/n$ is the probability that an arriving order belongs to the ith arrival with $i = 1, ..., k-1$ and $k > 1$. We determine the waiting time distribution of a customer order when $N_a = k$ arrivals are required to form a batch by

$$P(W_b = i|N_a = k) = \begin{cases} \sum_{j=1}^{k-1} \{\vec{a}^{(k-j)\otimes}\}_{k-j} \left(\frac{E(Y|N_a=k)}{n}\right) & \text{if } \ i \geq 1 \\ 1 - \sum_{i=1} P(W = i|N_a = k) & \text{if } i = 0. \end{cases} \quad (15)$$

The law of total probability leads to the waiting time distribution, given as follows

$$w_{b,i} = \sum_{k=1}^{k_{max}} P(W = i|N_a = k)P(N_a = k) \quad i = 0, 1, ..., \quad (16)$$

where $P(N_a = k)$ is given by Equation 10.

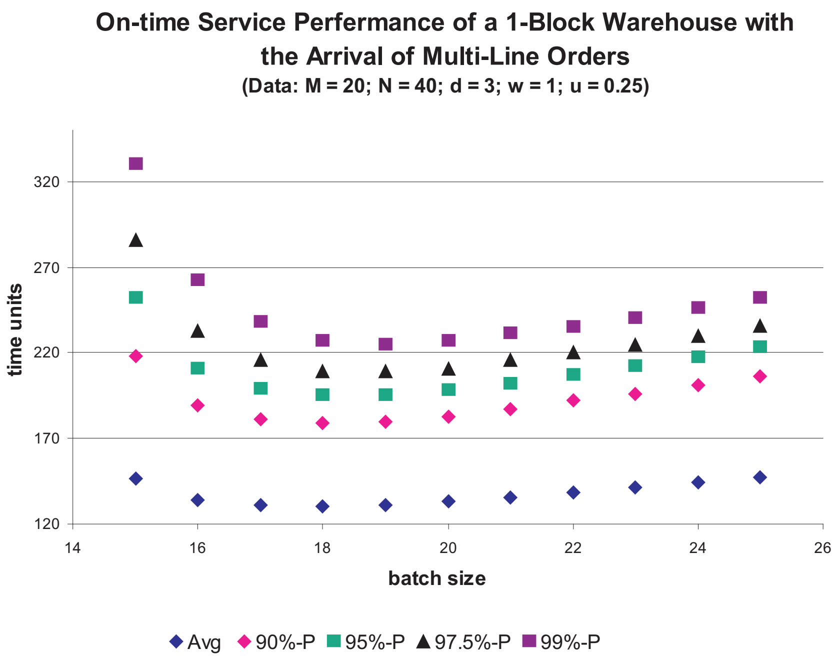

Figure 13: Optimal batch size in a 1-block warehouse for the multi-order line case.

Given $d_{b,i}$, $y_{b,i}$, $w_{b,i}$, and the travel time model introduced in Section 2, the throughput time distribution for the multi-line order case can be computed using the results in Section 3.4.

4.2. Throughput Time Analysis for Multi-Line Orders

The system behavior for the multi-line order case is similar to that for the single-line case. As before, we can easily compute the optimal batch size or calculate the expected service level for a given set of parameters. In an example (see Figure 13), we assume $M = 20, N = 40, d = 3, w = 1, n = 12, u = 0.25$, the interarrival time distribution has pmf $\vec{a} = \{0.0, 0.010, 0.020, 0.025, 0.050, 0.070, 0.050, 0.090, 0.080, 0.070, 0.060, 0.055, 0.055, 0.050, 0.050, 0.045, 0.040, 0.035, 0.030, 0.030, 0.030, 0.025, 0.020, 0.020, 0.015, 0.015, 0.010\}$, and the distribution of number of order lines in a customer order has pmf $\vec{y} = \{0.0, 0.4, 0.3, 0.1, 0.1, 0.1\}$, where the first element is y_0. As in the single-line case, the variability of the system has a strong effect on the throughput time, and therefore on the system's performance. The service level of the picking system is considerably more sensitive to an increasing process variability than is the mean throughput time.

5. Conclusion

The contributions of the paper to the literature are the following: First, we introduced with discrete time queues a new methodology to model picking processes in a warehouse. Discrete time queues have the attractive feature of admitting empirical data for interarrival times or for the number of lines in an order, which makes them easier to implement in practice. Second, the models produce throughput times distributions of picking orders in a warehouse, but require only a stationary arrival stream of orders. The models provide a fast way to predict, for example, what percentage of orders can be processed in a specific lead time or, equivalently, what lead time is required to promise a specific level of service. Third, we showed that experiments with prototypical, single-picker warehouse systems suggest that systems designed with high service level as the goal can be different than those designed to optimize mean throughput, albeit not significantly. For example, the batch size that minimizes the 95th percentile of throughput time is slightly larger than the batch size that minimizes the mean throughput time. High percentiles of throughput time are also more sensitive to changes in processing rates and variances than are mean throughput times.

Picking time distributions are also necessary to properly establish the length of picking waves, in which batches of orders are released to workers simultaneously. In practice, warehouse managers sometimes assess the mean time to process a batch and add to this time an experience-based "safety buffer" to account for variability in the processing time. Our models might be used to establish these times more scientifically.

Role of the funding source

This research was sponsored by the German Research Foundation grant SCHL 1870/1-1. The sponsor played no role in the design or execution of the research, nor in the writing of the paper. The authors are thankful for the support.

References

Ackroyd, H. M., 1980. Computing the waiting time distribution for the G/G/1 queue by signal processing methods. IEEE Transactions on Communications 38 (1), 52–58.

Chew, E. P., Tang, L. C., 1999. Travel time analysis for general item location assignment in a rectangular warehouse. European Journal of Operational Research 112, 582–597.

de Koster, R., van der Poort, E., Wolters, M., 1999. Efficient order batching methods in warehouses. International Journal of Production Research 37, 1479–1504.

Gademann, N., van de Velde, S., 2005. Order batching to minimize total travel time in a parallel-aisle warehouse. IIE Transactions 37, 63–75.

Grassmann, W. K., Jain, J. L., 1989. Numerical solutions of the waiting time distribution and idle time distribution of the arithmetic GI/G/1 queue. Operations Research 37 (1), 141–150.

Hall, R. W., 1993. Distance approximations for routing manual pickers in a warehouse. IIE Transactions 25 (76-87).

Hasslinger, G., Klein, T., 1999. Breitband-ISDN und ATM-Netze. Teubner.

Hübner, F., Tran-Gia, P., 1995. Discrete-time analysis of cell spacing in ATM systems. Telecommunications Systems Vol. 3, pp. 379-395.

Jarvis, J. M., McDowell, E. D., 1991. Optimal product layout in an order picking warehouse. IIE Transactions 23, 93–102.

Jewkes, E., Lee, C., Vickson, R., 2004. Product location, allocation and server home base location for an order picking line with multiple servers. Computers & Operations Research 31, 623–626.

Johnson, N. L., Kotz, S., 1977. Urn Models and their Application. John Wiley, New York.

Karmarkar, U. S., Kekre, S., Kekre, S., 1985. Lotsizing in multi-item multi machine job shops. IIE Transactions 17 (3), 290–297.

Kunder, R., Gudehus, H., 1975. Mittlere Wegzeiten beim eindimensionalen Kommissionieren. Zeitschrift für Operations Research 19, B53–B72.

Le-Duc, T., de Koster, R., 2007. Travel time estimation and order batching in a 2-block warehouse. European Journal of Operational Research 176, 374–388.

Petersen, C. G., 2000. An evaluation of order picking policies for mail order companies. Production and Operations Management 9, 319–335.

Petersen, C. G., Aase, G., Heiser, D. R., 2004. Improving order picking performance through the implementation of class based storage. International Journal of Physical Distribution and Logistics Management 34, 534 – 544.

Roodbergen, K. J., Sharp, G. P., Vis, I. F., 2008. Designing the layout structure of manual order picking areas in warehouses. IIE Transactions 40, 1032–1045.

Roodbergen, K. J., Vis, I. F., 2006. A model for warehouse layout. IIE Transactions 38, 799–811.

Rosenwein, M. B., 1996. A comparision of heuristics for the problem of batching orders for warehouse selection. International Journal of Production Research 34, 657–664.

Schleyer, M., 2007. Discrete time analysis of batch processes in material flow systems. Dissertation, Institute for Material Handling and Logistics, University of Karlsruhe.

Schleyer, M., 2010. An analytical method for the calculation of the number of units at the arrival instant in a discrete time G/G/1-queueing system with batch arrivals. OR Spectrum(Online first).

Schleyer, M., Furmans, K., 2007. An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals. OR Spectrum 29 (4), 745–764.

Tran-Gia, P., 1996. Analytische Leistungsbewertung verteilter Systeme. Springer.

Van Nieuwenhuyse, I., de Koster, R., 2009. Evaluating order throughput time in 2-block warehouses with time window batching. International Journal of Production Economics 121, 654–664.

6. Appendix A

Following is an overview of the method of Grassmann and Jain (1989) to calculate the waiting time distribution of a discrete time G/G/1 queue. The basic assumptions are as follows: Orders arrive to the system and are served in the order of arrival. The waiting time of the nth order is W^n and the service time is T^n. The time elapsing between two order arrivals is A^n. We denote $C^n = T^n - A^n$ and name it the working balance of the nth order. It is assumed that $E(C^n) < 0$ and that C^n are independent and identical distributed random variables, which can only take on discrete time values. Furthermore, the following variables are defined:

a_i probability that the time between two orders takes on i time units, $i = 1, 2, ..., max(a)$,

t_i probability that the service time takes on i time units, $i = 1, 2, ..., max(t)$,

w_i probability that the waiting time takes on i time units, $i = 0, 1, 2, ...$,

c_i probability that the working balance takes on i time units, $i = min(c), ..., -2, -1, 0, 1, 2, ...max(c)$.

Then,

$$c_i = \sum_{j=1}^{max(t)} t_j a_{j-i} \qquad i = min(c), ..., -2, -1, 0, 1, 2, ..., max(c). \tag{17}$$

By Lindley's equation in discrete form, w_i is given by

$$w_i = \begin{cases} \sum_{j=0}^{\infty} w_j \cdot c_{i-j} & \forall i = 0, 1, 2, ... \\ 0 & \forall i < 0. \end{cases} \tag{18}$$

Grassmann and Jain (1989) present three algorithms to solve Equation 18, which is based on the Wiener-Hopf factorization. They show the convergence of Algorithm 1, which includes the following steps:

1. Initialize $\beta_i^0 = 0, i = 1, 2, ..., -min(c)$ and $\alpha_i^0 = 0, i = 1, 2, ..., max(c)$
2. For m = 0,1,2,...
   (a)

$$\beta_i^{m+1} = c_{-j} + \sum_{i=1}^{\infty} \frac{\alpha_i^m \beta_{i+j}^m}{(1 - \beta_0^m)} \; j = 0, 1, \ldots, -min(c) \tag{19}$$

(b)

$$\alpha_i^{m+1} = c_j + \sum_{i=1}^{\infty} \frac{\alpha_{i+j}^m \beta_i^m}{(1-\beta_0^m)} \; j = 1, \ldots, max(c) \tag{20}$$

3. Iterate until $\max(|\alpha_i^m - \alpha_i^{m+1}|) < \epsilon$
4. It follows:

$$w_0 = 1 - \frac{\sum_{i=1}^{max(c)} \alpha_i}{1-\beta_0} \tag{21}$$

$$w_i = \frac{\sum_{i=1}^{max(c)} w_{i-j}\alpha_i}{1-\beta_0} \tag{22}$$

5. β corresponds to the idle time distribution.

7. Appendix B

i	a_i								
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0056	0.0000	0.0000	0.0388	0.0614	0.0837	0.1010	0.1167	0.0460
2	0.0033	0.0200	0.0200	0.0200	0.0169	0.0138	0.0113	0.0098	0.0500
3	0.0015	0.0268	0.0879	0.0200	0.0146	0.0100	0.0100	0.0073	0.3169
4	0.2215	0.3051	0.2830	0.2913	0.2844	0.2774	0.2717	0.2680	0.0354
5	0.3261	0.1992	0.1656	0.1829	0.1752	0.1674	0.1611	0.1568	0.0100
6	0.2755	0.1780	0.1432	0.1617	0.1540	0.1462	0.1399	0.1353	0.0196
7	0.1478	0.1724	0.1467	0.1586	0.1518	0.1448	0.1390	0.1345	0.4348
8	0.0020	0.0986	0.1324	0.0832	0.0779	0.0724	0.0680	0.0637	0.0086
9	0.0035		0.0212	0.0000	0.0000	0.0000	0.0000	0.0000	0.0103
10	0.0132			0.0435	0.0638	0.0842	0.0879	0.0852	0.0108
11							0.0032	0.0017	0.0101
12							0.0070	0.0211	0.0475
mean	5.4000	5.4000	5.4000	5.4000	5.4000	5.4000	5.4000	5.4000	5.4000
SCV	0.0500	0.0750	0.1000	0.1250	0.1500	0.1750	0.2000	0.2250	0.2500

Table 2: Interarrival time distributions, a_i, which have the same mean value but a varying SCV. These data are used to study the impact of the variability of the order arrival process on throughput time (Figure 11).

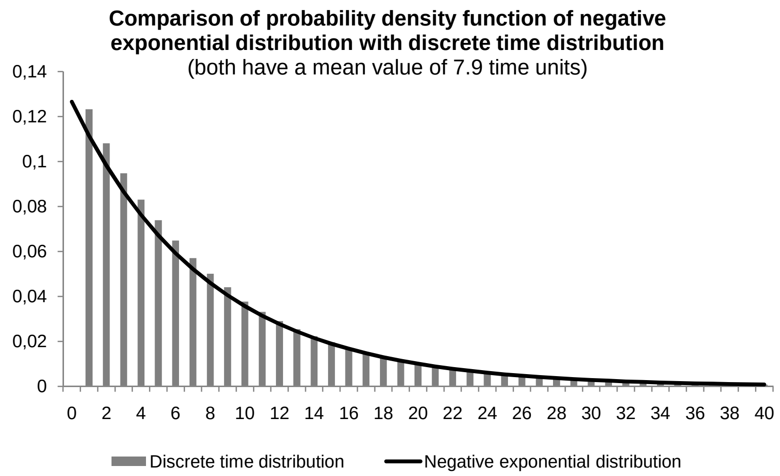

Figure 14: Continuous and discrete time interarrival time distributions with the same mean.