The effect on fulfillment time of the M-Division warehouse design approach

Ferretti I.*, Das S.**

**Department of Mechanical and Industrial Engineering, University of Brescia, Via Branze, 38 25123 – Brescia – Italy (ivan.ferretti@unibs.it)*

***New Jersey Institute of Technology, Newark, NJ – USA (das@njit.edu)*

Abstract: In last years, we are facing on the diffusion of on-line orders, a process further accelerated by the spread of the COVID-19 virus. A fundamental lever in the use of the e-Commerce is the continuous reduction of the fulfillment time. In order to obtain these improvements, the storage method has been revolutionized. In fact, we are talking about Online order fulfillment warehouses (F-Warehouses) which, unlike traditional warehouses, have a very large number of small bin locations, an explosive storage policy, where an incoming bulk is separated into small lots stocked in any bin throughout the warehouse, and commingled bin storage. The divided warehouse (M-Division) with explosive storage is a novel design, where the warehouse is divided in M-zones managed by specific pickers. In this paper we model the fulfillment process as an N server queuing model with uniform service times and compare the fulfillment time of the traditional warehouse with the M-Division warehouse. In particular, an analytical solution assigns items to the M-Division warehouse in order to minimize the fulfillment time. The results show that the M-Division approach permits a reduction of the fulfillment time compared with the traditional one.

Keywords: e-Commerce, F-warehouses, Queueing Theory

1. Introduction

The fulfilment and logistics systems of an online retailer typically involve one or more fulfilment centers and an associated parcel delivery network. In the case of Amazon, though, significant parts of the network are in-house operations. In particular, the core of Amazon is the software organization which provides complex algorithms and optimization programs that run the daily operations of the fulfilment centres. Onal et al. (2017) was one of the first to provide detailed insights into the operational flows within an Amazon fulfilment centre. They documented how sophisticated flow control models leverage new logistics and operational models to ensure fast fulfilment. They present a new paradigm in the operational design and control of warehouses, identifying six specific operational differentiators:

- explosive storage policy - incoming bulk inventory is exploded into a large number of small lots which are then dispersed to storage locations throughout the warehouse;
- very large number of beehive storage locations - storage is organized into small library style bins (13 cubic feet);
- bins with commingled items - multiple items are simultaneously stored in an unorganized way in the same bin;
- immediate fulfilment objective - customer orders arrive continuously throughout the day and the goal is for same-day shipment;
- short picking routes with single unit picks - most orders are only for a single unit and the pick list retrieves several different items within a short pick zone;
- high transactions volumes and total digital control - there is a much higher rate of store/pick movements per unit shipment, and all movements are modelled and instructed by a central controller.

Together these differentiators uniquely describe a new approach to fulfilling online orders, and those characteristics are identical for all the Internet Fulfilment Warehouses (IFW). Moreover, it should be noted that each of the first three differentiators represents a radical change from the traditional storage management approach. In table 1 are reported IFW operational differences (Zhang et al. 2020).

Table 1: IFW operational differences

Attribute	F-Warehouse	Distribution Center
Stock locations (bins)	2,000,000	10,000
Items (SKUs) Stocked	100,000	6,000
Orders Fullfilled/days	40,000	4,000
Inventory data records	20 Milion	10,000

In selecting an online retailer, two important criteria from a consumer perspective are price and fulfilment time. Fulfilment time is defined as the interval between a customer’s order placement and delivery to the customers location. In particular, faster fulfilment is a critical driver of success in online retail and is motivating customers to shift to online buying. In fact, a key component of IFWs is customer order picking with the objective of immediate fulfilment. Recent papers use a location closeness partitioning algorithm to make a pick batch decision, while others derive an order picking throughput model with multiple pickers and aisle congestion. Rao and Adil (2017) developed analytical travel distance models for storage policies under across-aisle, within-aisle and a newly hybrid product placement schemes. De Koster et al. (2007) suggest that order picking for online demand is different and is akin to stochastic optimization problems. The storage location assignment problem (SLAP) is described by Gu et al. (2007) as assigning zones and locations (M-Division) to incoming bulk to reduce material handling costs and improve space utilization. They found that the three common SLAP criteria are turnover-based assignment, class-based assignment and cube-per-order index. Xiao and Zheng (2010) studied a correlated storage location assignment problem where relationships are determined from a bill of materials. The SLAP class of problems is closely aligned to explosive storage in IFWs. The common location assignment policies are identified as random storage, closest open location storage, dedicated storage, full turnover storage, and class-based storage. Several studies have confirmed the advantages of a random policy, though, not conclusively. Petersen compared class-based storage to random storage. The results show that class-based storage when compared to random storage results in shorter picker travel distances. A solution is presented by Pang and Chan (2017), with a data mining-based random SLAP solution by extracting and analyzing the association relationships between different products in customer orders. Ho and Sarma (2009), on the other hand, modeled free form storage, where items are stocked in multiple locations, and showed that it can improve picker travel time. Another solution considered a warehouse arranged in multiple cells, and any arriving bulk could be split and stocked in one or more cells, with two possible policy, one based on chaining and the other on picker workload balancing. Weidinger and Boysen (2018) also considered a warehouse where incoming bulk was exploded and stored in multiple locations. They segmented the warehouse into several parts each of which has a defined delivery point where pickers deposit the items. The problem is formulated as a MIP and then initially solved by a p-center search procedure. They developed an adaptive binary search heuristic algorithm and compared it to a random storage assignment. Weidinger and Boysen (2018) used a simulation model to confirm that their approach decreased picker travel distances as the number of measuring points increased.

In a shared storage policy multiple SKUs are stored in the same location thus increasing space utilization. Some research showed that a duration-of-stay-based shared policy is optimal under an assumption of perfectly balanced inputs and outputs. Cormier and Gunn (1992) states that shared storage policies offer excellent potential for travel time and rack size reductions. Commercial warehouse management systems apply three well-known methods: S-shape strategy, Return strategy and Zig-zag strategy. Aisle layout is also an associated problem since cleverly designed aisles could reduce pick travel distance. Various researchers propose several such designs. However, they pointed out that IFWs typically use a classical rectangular row layout. Won and Olafsson (2005) consider the joint objective of low picking time and faster customer response time. They proposed a batching and order picking solution using simple heuristics. Tsai, Liou, and Huang (2008) propose a batch picking model that considers also an earliness and tardiness penalty with a focus on quick-response. Traditional warehouses store a SKU either in a set of contiguous dedicated or random locations or slots. In either case, the number of storage locations for a specific SKU is few (<10) to preserve the bulk. In contrast, in an IFW the incoming bulk is immediately broken into unit SKUs upon arrival. The exploded units are then dispersed to bins throughout the warehouse. The bins could be random or prescribed by a rule, and each bin could receive several units of the same SKU. Onal et al. (2017) describe this as an explosive storage policy: the bulk SKU is exploded into E storage lots of one or more units such that no lot contains more than 10% of the received quantity. The lots are then stored in non-contiguous bins anywhere in the warehouse. Let $i \in N$ be the unique SKUs stored in the warehouse, E_i the explosion factor and V_i the current total warehouse inventory for i, and L_i the number of unique bins where it is stocked. Then we introduce:

$$\text{Explosion Ratio for product } i = \psi_i = \frac{L_i}{V_i}$$

$$\text{Warehouse Explosion Ratio} = \psi_0 = \frac{\sum_{i \in N} L_i}{\sum_{i \in N} V_i}$$

L_i is not simply equal to E_i. Since bulk batches are arriving at some interval, every explosion will send the lots to both existing and new locations. At the same time fulfilment is occurring, thus L_i is changing constantly and Ψ_i is time variant. The warehouse explosion ratio is then an inventory weighted function. An explosive strategy will significantly improve fulfilment time and IFWs should operate between Ψ_0 = 0.1 and 0.8 where the picking opportunities are maximized; further explosion generates only redundant opportunities.

In this paper we model the fulfillment process as an N server queuing model with uniform service times and compare the fulfillment time of the traditional warehouse with the M-Division warehouse in order to evaluate differences. In particular, the focus of the research is only on the warehouse activity. The reminder of the paper is structured as follows. The problem definition, notation and assumptions are in Section 2. The development of the model is in Section 3. The application of the model and simulation is introduced in Sections 4. A summary of the paper, its main findings, and future developments of the present work are in Section 5.

2. Problem and assumptions

It is clear that the concept of fast fulfillment requires that an incoming sales order be immediately collected and processed. If the latter is collected within an hour, then it can be packed and shipped within hours of receiving it. In practice we have that the speed with which an item is picked is a function of how much it is stored near a picker: the greater the distance between the operator and the object to be picked, the greater the time needed to complete the task. The proximity of the article to a selector can be described by a probability model which is a function of the storage policy adopted by the system. As we have seen, one of the key features of Internet Warehouse Fulfillments is the adoption of an explosive storage policy, which divides the incoming supplies into many small batches and then distributes them throughout the warehouse and this consequently increases this probability. In fact, the problem presented here deals with the storage policies adopted, comparing the traditional approach, with a single department store that processes all the orders, and the more recent one, which divides the orders among several locations. By expanding the number of storage locations, the probability of quick fulfillment for a given item increases, and therefore the goal is reached more easily. To demonstrate this concept, we present two examples:

- 16-location storage policy;
- 64-location storage policy.

In the first example we find that each picker has immediate access to four positions, if there is only one article it is available in all 16 positions and the probability of rapid evasion of the same is equal to 100%. Furthermore, as the number of products increases, this probability decreases and, as the number of locations increases, it increases. This organization shows strong limitations as already with 8 products, and even by exploiting 4 positions, the probability of immediate evasion is equal to 50% and with 16 objects it further decreases to 25%.

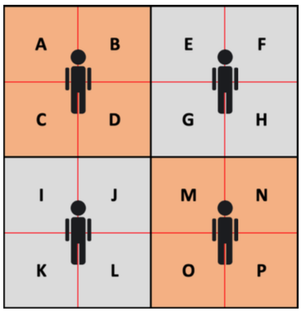

Figure 1: Graphic representation of a 4x4 warehouse.

By decreasing the size of the cells and arriving at an 8x8 matrix arrangement, the number of locations to which the individual pickers have immediate access increases to 16. Even with 16 items each can be stored in four different locations and you get a 100% fast withdrawal probability for all items. As the number of stocked items increases and therefore expands the number of stocking positions, a probability of rapid fulfillment of 100% is maintained. It can be seen by comparing in the Table 2 the two configurations that the performances in the second case are considerably improved.

Table 2: Probability of quick evasion.

Stock Items	Case	
	4x4	8x8
1	1	1
2	1	1
4	1	1
8	1/2	1
16	1/4	1

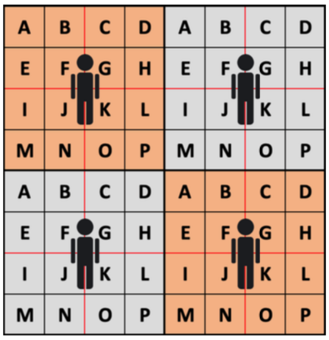

Figure 2: Graphic representation of a 8x8 warehouse.

However, the fact that orders are processed from different locations is not a guarantee of a satisfactory order fulfillment time for the competitive and commercial purposes of an IFW. Therefore, a further problem arises of assigning items in the warehouses, having to choose which destination to assign a certain object and in what quantity.

The objective of this research is to compare the traditional warehouse configuration, where in a unique partition are stocked all the SKUs and where the pickers work at the same time in the same area, with the IFW configuration, where the SKUs are assigned to specific partitions (*M*-division) and only one picker works in a specific partition at the same time. In the specific, the final goal is to compare the two configurations in terms of fulfillment time and saturation of the pickers. The fulfillment time is calculated as the difference between the arrival time of one order and the delivery time of the same order. For simplicity, the order requests only one SKU and the order strategy is “order picking”. In the next figures the process flow of the two configurations considered are shown.

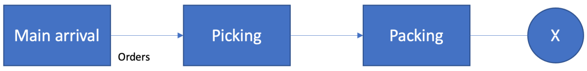

Figure 3: Process flow of the Traditional warehouse configuration.

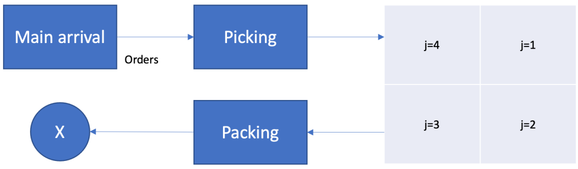

Figure 4: Process flow of the IFW configuration (with *M*=4).

In the case of the traditional configuration, after the arrival of one order, the pickers start the process and pick-up the related SKU in the warehouse. After that, the SKU is moved to the packing area where another operator prepares the pack for the delivery. In the case of IFW configuration, when the order arrives, before it is assigned to a specific partition, then the specific picker start the process and pick-up the related SKU in the partition of his/her competence. After that, the SKU is moved to the packing area where another operator prepares the pack for the delivery. The packaging is not considered in this study for simplicity.

Summarizing, we assume that:

- Item demand is exponential.
- Picking time is uniform and a function of partitions in case of IFW configuration.
- Pickers cannot work in other partitions.
- Each order is only for a single item.

3. Model

In order to solve the problem, we define an open queueing network with exponential arrival time and uniform service time. In the figure 5 and 6, we show the schemas of the open queuing network in the case of traditional configuration and IFW configuration. We model the two networks with an open Jackson networks with exponential arrival rate and finite number of servers. The network for the traditional warehouse configuration is simply a node with *M* servers, while for the IFW configuration, we have parallel nodes with one server.

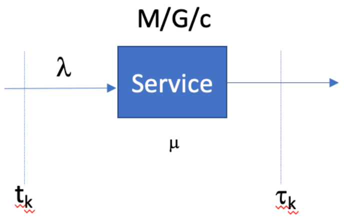

Figure 5: Open Jackson Networks for Traditional warehouse configuration.

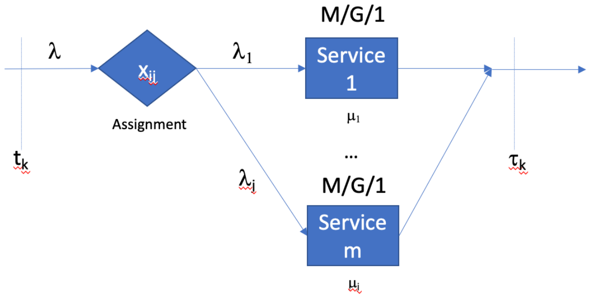

Figure 6: Open Jackson Networks for IFW configuration.

In the case of traditional configuration, we model the warehouse with *M* picking servers. So, the picking activity is modeled with a M/G/c queueing model, where the number of resources $c = M$. The server has a specific uniform service distribution. Following, we show the notation and the model for the traditional warehouse:

- $i = 1,\ldots,N$: number of Items
- *M:* number of pickers
- *k*: number of orders
- t_k : arrival time
- τ_k: delivery time
- λ = average arrival rate (orders)
- μ = average service rate (picking)
- σ = service standard deviation (picking);

In order to evaluate the fulfillment time, it is necessary to calculate the difference between the delivery time τ_k and the arrival time t_k. By using the queueing theory formulation this difference in the case of traditional warehouse is equal to:

$$\tau_k - t_k = \frac{C^2+1}{2}\frac{\rho(\rho M)^{a_M}\pi_0}{\lambda M!\,(1-\rho)^2} \quad (1)$$

where:

$$\rho = \frac{\lambda}{M\mu} \quad (2)$$

$$\pi_0 = \left(\sum_{h=1}^{a_M-1}\frac{(M\rho)^h}{h!} + \frac{(M\rho)^{a_M}}{M!\,(1-\rho)}\right)^{-1} \quad (3)$$

$$C = \frac{\sigma}{\mu} \quad (4)$$

In the case of IFW configuration, we model the warehouse with an open Jackson network with *M* parallel picking servers. So, the *M* picking activities are modeled with M/G/1 queueing model. The server has a specific uniform service distribution. In particular, we have to consider that the items are distributed, with respect to the demand, with a discrete distribution. For simplicity we consider in the present work a homogeneous distribution of the items equal to *1/N,* where N is the number of items considered. The presence of *M* warehouses leads to the assignment of the picking task to a specific picker. In order to model this condition, we introduce the variable x_{ji} that is the assignment of the picking task for item *i* to the warehouse or picker *j*. Following, we show the notation and the model for the IFW:

- $i = 1,\ldots,N$: number of Items;
- $j = 1,\ldots,M$: warehouse partitions;
- p_i: items distribution;
- x_{ji}: assignment of the picking task for item *i* to the warehouse *j;*
- *k*: number of orders;
- t_k: arrival time;
- τ_k: delivery time;
- λ = average arrival rate (orders);
- μ_j = average service rate (picking);
- σ_j = service standard deviation (picking);

In the case of IFW configuration the fulfillment time is equal to:

$$\tau_k - t_k = \sum_{j}^{M} \left(\sum\nolimits_{i}^{N} p_i x_{ji}\right)\left(\frac{\lambda_j^2 \sigma_j^2 + \rho_j^2}{2\lambda_j(1-\rho_j)} + \frac{1}{\mu_j}\right) (5)$$

In the case of IFW configuration, in order to solve the assignment problem, it is necessary to introduce the following constraints:

$$\lambda_j = \lambda \sum\nolimits_i^N p_i x_{ji} \text{ with } j = 1,\ldots,M \quad (6)$$

$$\sum\nolimits_j^M x_{ji} = 1 \text{ with } i = 1,\ldots,N \quad (7)$$

$$\lambda_j/\mu_j < 1 \text{ with } j = 1,\ldots,M \quad (8)$$

$$x_{ji} > 0 \; with \; j=1,\ldots,M \; and \; i=1,\ldots,N \quad (9)$$

where the constraint (6) represents the application of the Jackson Theorem for the calculation of the arrival rate for every warehouse, in the constraint (7), the sum of the assignment variable is equal to one and greater than zero (9). In constraint (8), the saturation index for every warehouse has to be lesser than 1 in order to have the convergence of the queuing system. We validate the analytical model by using a simulative model coded in Anylogic®. The problem is solved using the software LINGO®.

4. Parameters setting and application

In this section we explain the value for every parameter and the scenarios simulated. In order to carry out the simulations, a series of simplifying hypotheses are made. Among these various are in common between the two models created and analyzed:

- The distribution of demand is known and exponential.
- The demand for the different items is unknown, but each type of order is assigned the same probability of being generated.
- Each order contains one and only one item.
- The warehouses have no size constraints and can store all types of items.

In the specific, we model the items demand as exponential distribution in order to have a random behavior of the items arrival in the system. Moreover, in order to simplify the case study, we assume that each order considers only one item and the warehouses can store all types of items.

In the following table, we list the different parameters with the reference value:

Table 3: Parameters and settings

Parameter	Value
N	40
λ	
μ	8
μ_j	4,5
σ	0,57
σ_j	0,86
p_i	2.5%
M	3

In particular, *N* is the number of items considered, μ is the average service rate in the case of traditional configuration, while μj is the average service rate in the case of IFW. The value of μj is lower than μ because the physical area for every picker in the case of IFW is lower (we assume a reduction of about 50%). *Pi* is the probability distribution of the items with respect to the demand; in this case we assume a probability equal to 1/N.

In the case of the IFW configuration, the definition of the solution leads to the minimization of the average fulfillment time by optimizing the assignment variable x_{ij}.

5. Results

In this section we compare the results obtained from the two configurations under examination to understand which of the solutions offers performance best suited to the final purpose with respect to the fulfilment time and the saturation of the pickers. For the simulations of the two configurations an exponential distribution at the input was used with mean values λ ranging from 3.5 to 6 and with increases of 0.5, thus reaching a total of 6 distinct scenarios.

As shown in figure 7, the IFW configuration finds sense of application even with much higher order generation frequencies, thus ensuring operation even with much higher workloads. The traditional configuration on the other hand does not allow for satisfactory performance below the generation Mean Value of 3.5, further underlining how it is not suitable for use in online sales. Looking at the detail, we can also observe that the IFW configuration tends to the horizontal asymptote much faster than the other. The gap between the second simulation and the last (Mean Value 4.0 and 6.0), in the fast fulfillment model, is only 0.62 units, while in the first case this difference is 1.45.

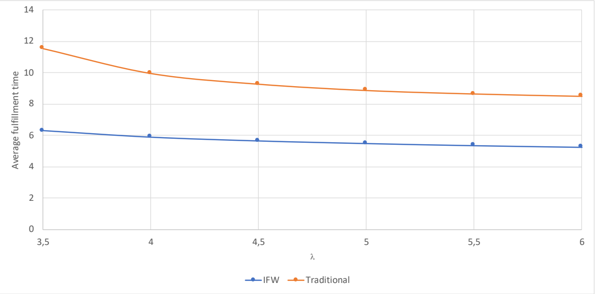

Figure 7: average fulfillment time comparison.

Finally, by comparing the average saturation in figure 8, we show that the IFW has an average saturation for all the scenarios lower than the traditional configuration.

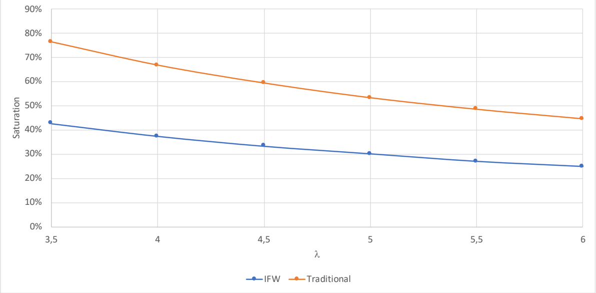

Figure 8: average saturation comparison.

So, we can conclude that in case of IFW configuration by increasing the number of warehouses, and optimizing the distribution of products among them, the average order fulfillment time of the system decreases, albeit at the expense of operator saturation increasing.

6. Conclusions and future research

This study sought to answer the question: "Does the new approach adopted by IFWs actually perform better than the traditional one?". For this purpose, an optimization model was conducted in order to verify and compare the effectiveness of a traditional approach to the more recent one adopted by Internet Fulfillment Warehouse - IFW. These models were tested with different parameters, varying the average generation time in the exponential distribution in order to compare the two configurations with respect to the fulfilment time and the pickers saturation.

The results obtained from these simulations show that the IFW significantly performs better than the traditional configuration in terms of fulfilment time, but not in terms of pickers saturation. In fact, in IFW configuration, the emphasis is placed on fulfillment time, as the final customer expects the items to be delivered in ever shorter times as an essential condition. From the other hand, the traditional configuration maximizes in the same scenarios the saturation of the pickers, despite to the increasing of the fulfilment time. To further validate the results obtained, the models themselves should be gradually expanded to reflect more concrete operating conditions, considering factors such as the packaging activity, "last mile delivery time", item procurement time and presence of orders with multiple items.

Reference

G. Cormier, E. A. Gunn. 1992 A review of warehouse models. European Journal of Operational Research, vol. 58, issue 1, 3-13.

De Koster, R., Le-Duc, T., Roodbergen, K.J., 2007. Design and control of warehouse order picking: a literature review. Eur. J. Oper. Res. 182 (2), 481–501.

Gu, J., Goetschalckx, M., McGinnis, L.F., 2007. Research on warehouse operation: a comprehensive review. Eur. J. Oper. Res. 177 (1), 1–21.

Ho, S.S., Sarma, S., 2009. The fragmented warehouse: location assignment for multi-item picking. IEEE 2nd Int. Logist. Ind. Inf. LINDI 2009 1–6.

Onal, S., Zhang, J. and Das, S. (2017), "Modelling and performance eval- uation of explosive storage policies in internet fulfilment warehouses", International Journal of Production Research, Vol. 55 No. 20.

Pang, K., Chan, H., 2017. Data mining-based algorithm for storage location assignment in a randomised warehouse. Int. J. Prod. Res. 55 (14), 4035–4052.

Petersen, C.G., Aase, G., Heiser, D.R., 2004. Improving order-picking performance through the implementation of class-based storage. Int. J. Phys. Distribution Logist. Manag. 2004 34 (7), 534–544.

Rao, S. S., and G. K. Adil. 2017."Analytical Models for a New Turnover-based Hybrid Storage Policy in Unit-Load Warehouses." International Journal of Production Research 55 (2): 327–346.

C.Y. Tsai, J.J.H. Liou, T.M. Huang 2008 Using a multiple-GA method to solve the batch picking problem: Considering travel distance and order due time. International Journal of Production Research 46(22):6533-6555.

Weidinger, F., Boysen, N., 2018. Scattered storage: how to distribute stock keeping units all around a mixed-shelves warehouse. Transp. Sci. 52 (6), 1412–1427.

J. Won, S. Olafsson 2005 Joint order batching and order picking in warehouse operations International Journal of Production Research Vol. 43

Xiao, J., Zheng, L., 2010. A correlated storage location assignment problem in a single- block-multi-aisles warehouse considering BOM information. Int. J. Prod. Res. 48 (5), 1321–1338.

Zhang, J., Onal, S. and Das, S. (2020), "The dynamic stocking location problem Dispersing inventory in fulfillment warehouses with explosive storage", International Journal of Production Economics 224.

Zhang, J., Onal, S., Das, R., Helminsky, A. and Das, S. (2019), "Fulfil- ment time performance of online retailers - an empirical analysis", In- ternational Journal of Retail and Distribution Management, Vol. 47 No. 5.