March 2021

Optimizing the workload allocation in an e-fulfillment center using queueing theory

Faculty of Applied Mathematics

Department of Stochastic Operations Research

Master Thesis by:
N. Leijnse
S1431560

Supervisors University of Twente:
prof. dr. R. J. Boucherie
dr. J. C. W. van Ommeren
dr. M. Walter

Supervisors bol.com:
L. A. Botman
P. van de Ven

UNIVERSITY OF TWENTE.

Abstract

E-commerce business helps in creating trade but it heavily relies on logistical support in order to succeed. E-fulfillment, which is a commonly used term for a segment of logistics in e-commerce, is one of the biggest challenges in this sector. The main focus of research in this field is on optimizing the picking process, as this is the most labor intensive part. However, in order to maximize the performance of the e-fulfillment process as a whole, the focus should be on the distribution of the workload throughout the entire network.

In this study, we propose an approach to release pick batches to the warehouse of an e-fulfillment center, that takes into account the workload throughout the entire process. This approach is based on methods from queueing theory and involves some linear programming as well. In order to investigate whether the proposed approach performs better than the current approach, a mathematical model is formulated and for validation purposes a simulation model was created as well. The performance of the current and the proposed approach are assessed on the average throughput, sojourn time and work in progress. The results of the simulation model show that, with a significance level of 1%, the proposed approach performs significantly better than the current approach. However, the formulated mathematical model turns out to capture the logic behind the processes in the warehouse insufficiently enough to accurately determine the performance measures. Further research is required to adjust the mathematical model such that the gap between the model and practice becomes smaller.

Keywords: queueing network, work in progress, throughput, sojourn time, pick batch, batch releases

Executive summary

This research addresses the workload allocation problem in one of the warehouses of the online retailer bol.com. It is investigated how the allocation of the workload in the warehouse can be done automatically. The workload in the warehouse is controlled by the release of pick batches. By means of theoretical research, an approach for batch releases and a mathematical model are formulated. In addition, a simulation model is created for validation purposes. The performance of the proposed approach is tested against the current batch release approach.

Problem formulation

Currently, the workload allocation is done manually at bol.com by the control room. The people in the control room monitor the picking and packing areas and release new pick batches to the system. They determine how many pick batches are released and at what time. In order to help the control room in the decision-making process, the company created WES. This tool is created to provide a pick batch advice based on the number of operators and the work in progress levels at the packing stations, among some additional information. However, problems of this tool are that it can only be used for a part of the outbound process and that the generated batch advice is frequently larger than the number of waiting orders. There is clearly a need for better batching advice, which takes into consideration the workload allocation throughout the entire outbound process in the warehouse. The main research question is formulated as follows:

"How should the workload at the different work stations in the warehouse be allocated such that the overall throughput is maximized and the operating costs are minimized, while maintaining the order fulfillment score?"

Approach and methods

The outbound process in the warehouse is modelled as a multi-class open queueing network of multi-server queues. The workload in the warehouse is controlled by the release of pick batches to the system with a designated packing station, which from a queueing theory perspective corresponds to the arrival rates. These are thus the variables to be optimized. Essentially, the goal is to optimize the arrival rate of the pick batches for each type of packing station such that the throughput and utilization rate are maximized.

Working backwards through the queuing network of the warehouse, the bottleneck work center is found and the maximum aggregate arrival rate of pick batches is determined for a given utilization rate. If the bottleneck is not the packing center, it must be determined how the maximum aggregate arrival rate should be distributed over the different types of packing stations. An LP is formulated to make this decision.

Given the arrival rates, the expected number of pick batches in the system is determined. If the expected number of pick batches in the system is higher than the system capacity, above-described approach is repeated with a lower utilization rate until the arrival rates have been found that do not exceed the system capacity.

A very important assumption here is that there are always enough orders available to be released to the system for each type of packing station. This is achieved by reassigning orders to different packing stations and releasing orders that do not necessarily need to be shipped today but may also be shipped later during the week. Another LP is formulated to take care of this.

The performance of the proposed approach is tested against the current batch release approach by comparing the throughput, sojourn time, and work in progress levels as determined by the mathematical model and simulation model. T-tests are performed to validate the models and to determine whether the proposed approach performs significantly better than the current approach.

Results

In order to validate the simulation model, its results are compared to historical data of the company. Based on the results of the t-tests, the simulation model is regarded as an acceptable representation of the actual process in the warehouse. The mathematical model on the other hand, does not represent the actual process accurately enough according to the results of the t-tests. Further research should be done to diminish the gap between the model and reality.

Furthermore, the results show that the proposed approach for releasing pick batches performs significantly better than the current approach. The proposed approach results in lower and more constant work in progress levels such that the available capacity can be used more efficiently. Besides that, an increase in throughput of approximately 17% during the peak period and approximately 6% outside the peak period is expected with the proposed approach. As a result, more customer orders can be processed by the end of the day or the number of operators can be reduced. This also means that the company could reduce the interventions that put a break on the incoming customer orders. Examples of these interventions include shutting down particular shops and postponing the delivery date. Consequently, more customer orders can be accepted and fulfilled.

Limitations and recommendations

Both the simulation model and the mathematical model are a simplification of reality. A limitation is therefore that the models do not fully capture the process in the warehouse. Besides that, the models work with service time distributions of pick batches and assume that each pick batch for a designated packing station is of equal size, whereas in reality this varies a lot throughout the day and the service times are highly dependent on the number of items in a pick batch. For further research it is recommended to experiment with service time distributions that are dependent on the number of items in a pick batch and to make the sizes of the pick batches stochastic. In addition, it is recommended to research how the discrepancies between the models and reality can be further reduced.

The company is recommended to start researching how the service time distributions can be determined more accurately. After that, the logic of the proposed approach of releasing pick batches in a more timely and balanced manner can be implemented. The next step is to incorporate the logic of reassigning orders to different packing stations in order to automate this process as well.

At last, an idea for future research is to investigate how the same logic could be applied to a warehouse in which pick batches are created and coordinated from multiple picking areas. This is exactly what will happen in the Bol.com Fulfillment Center 2, which is one of the newest warehouses of the company.

Preface

In September 2020, I started my journey at bol.com. I was welcomed with open arms to perform research for my graduation assignment of the master's degree Applied Mathematics at the University of Twente. The past months have been a lot of fun and challenging in many ways. I could not have finished this thesis without the help of many people.

First of all, I would like to thank team 5P for welcoming me in their team. I definitely felt like I was part of the team and enjoyed all the jokes, walks and beautiful drawings during the checkouts. Next, I would like to thank Amir Chrigui from Ingram Micro for providing me with all the information I needed from the control room.

My special thanks goes out to Loek Botman, who served as my daily supervisor. He was always willing to help, think along, provide constructive feedback, and made me feel appreciated. In addition, I would like to thank Peter van de Ven for all his insightful questions and suggestions. He really encouraged me to dig deeper into certain topics to obtain a better understanding. Furthermore, I would like to thank Jan-Kees van Ommeren as my supervisor from the university for his constructive feedback and challenging me to think like a real mathematician. Finally, I would like to thank Richard Boucherie and Matthias Walter for being part of my graduation committee.

I look back with great pleasure at my time within team 5P and look forward to join bol.com from April on as a Business Analyst.

Nikki Leijnse

List of Figures

List of Tables

Glossary

APT	Average Processing Time of a pick batch.
ATT	Average Transportation Time of a tote from picking to packing.
BA	Batch Advice
BFC	Bol.com Fulfillment Center
BRC	Bol.com Retour Center
CATF	Current Average ToteFill in number of items.
DES	Discrete Event Simulation that provides a forecast of the future status of the orders (and batches) in BFC.
FCFS	First Come First Served
Fulfillment score	The percentage of orders delivered in time at the assigned transport carrier.
Item	A single physical unit of an international article number (EAN).
Lvb	Logistics via bol.com
Mono order	An order consisting of a single item.
Multi order	An order consisting of more than one item.
Order	Items of one customer order that are fulfilled from BFC.
Outbound line	A particular packaging line in the outbound process of BFC.
RATIO	Average processing rate, which equals the items per hour per operator.
SCV	Squared Coefficient of Variation
Tote	A blue bin used by an order picker to temporarily store the items of a pick batch being collected during a pick tour, and to transport items through BFC.
Vvb	Verzenden via bol.com (send via bol.com)
WES	Warehouse Execution Service
WIP	Work in Progress
Work center	Group of work stations that perform the same task.
Work station	Element of a work center that processes one item at a time.

Contents

Chapter 1

Introduction

All around the world e-commerce is expanding rapidly and has become an important driving force for economic development [36]. In 2016, there were 1.66 billion digital buyers who accounted for a global sales amount of 1.85 trillion US dollars. It is expected that the number of buyers will increase to 2.14 billion people with a projected revenue of 4.93 trillion US dollars in 2021 [32, 33].

E-commerce business helps in creating trade but it heavily relies on logistical support in order to succeed. Before placing an order, customers not only evaluate the product but also the delivery service. A high quality delivery service results in a satisfied customer experience, which can boost retention and consequently improve profits. It is all about getting the product to the customer at the right place, time, and cost [15]. Therefore, logistics has become the competitive element that could make the difference for online retailers.

A commonly used term for a segment of the logistics in e-commerce is e-fulfillment, which includes the picking, packing, and shipping of online customer orders [34]. A lot of research regarding e-fulfillment is focused on optimizing the order picking process, as this is the most labor intensive part. This also holds for bol.com, where multiple optimization projects have already been done in this field and several employees are continuously looking for further improvement.

However, changes in one part of the process could have a significant impact on another part. If for example the picking process has been improved and as a result the picking speed has increased, this means that the arrival rate at the sorting and packing centers also increases. The question is whether the sorting and packing centers have enough capacity to handle this increased arrival rate or that this will result in an enormous queue of items that are waiting to be sorted and packed. In order to prevent this and to maximize the throughput of the entire network, the workload allocation problem should be optimized simultaneously for the different work centers.

Currently, the workload allocation is done manually at bol.com. In this research project, it is therefore investigated how the allocation of the workload at the different work centers in the warehouse can be automated in such a manner that the overall throughput is maximized. In Chapter 2, background information is provided about the company bol.com and the logistical process in the warehouse. The problem statement is presented in Chapter 3. Next, a literature review is given in Chapter 4. A model description can be found in Chapter 5, followed by a mathematical formulation in Chapter 6. The solution approach and experimental set up are described in Chapter 7. The results are presented in Chapter 8. Finally, the limitations and recommendations can be found in Chapter 9 and the conclusions are presented in Chapter 10.

Chapter 2

Background information

The previous chapter provided a brief introduction to this research project. In this chapter some background information is provided in order to obtain a better understanding of the research environment. First, some background information about the company bol.com, at which the research takes place, is given in Section 2.1. This is followed by an overview of the logistical process at bol.com in Section 2.2. The chapter closes with a deep dive into the outbound process of the bol.com fulfillment center in Section 2.3.

2.1 Introduction to bol.com

In 1998 the German company Bertelsmann A.G. announced that they would start a global electronic bookstore with the working title "Books OnLine". One year later, the company launched bol.com in the Netherlands, since bol.nl was already taken by another company, that was not willing to sell the domain name. It was the first online bookseller in the Netherlands with an assortment of 140,000 Dutch books. Soon after, the assortment was expanded with CD's and later also with movies and television series. In 2003 Weltbild, Holtzbrinck Networkx and T-Online Venture Fund took over bol.com from Bertelsmann and started selling games and software as well. After that, the assortment kept expanding and it still is. From 2010, bol.com also started serving the Flemish part of Belgium. In the same year, director Daniel Ropers visited Silicon Valley and came back with the idea to become an open platform, such that third parties can sell their products through bol.com as well. This idea became reality in 2011. The company changed from an online retailer into an online retail platform. In 2012, the company was taken over by Ahold, who merged with Delhaize in 2016 and is now known as Ahold Delhaize. This merger created the opportunity for bol.com to expand to the French part of Belgium as well. [30]

Bol.com offers over twenty million products from four different sources:

1. Bol.com products, which are purchased by bol.com from their suppliers, stored in one of the warehouses and sent to the customers on order.
2. Plaza products, which are products from partners who use the bol.com webshop to sell their products but store and distribute the products themselves.
3. Verzenden via bol.com (*Vvb*) products, which are products from partners who use the bol.com webshop to sell their products, store the products themselves, and outsource the distribution of their products to bol.com. The partner brings the ordered packages to a collection point and then bol.com takes care of the distribution through their contracted transport carriers.
4. Logistics via bol.com (*Lvb*) products, which are products from partners who use the bol.com webshop to sell their products and outsource the storage and distribution of their products to bol.com. This process is similar to the process of bol.com products, except for the fact that the Lvb partner remains the product owner and decides how many items are sent to the warehouses of bol.com.

Daily approximately 200,000 items are distributed from the warehouses to customers in the Netherlands and Belgium. In this report an *item* is defined as a single physical unit of an international article number (EAN). Right now, bol.com operates from six different warehouses of which the bol.com fulfillment center (*BFC*) in Waalwijk is the largest and distributes 40%-50% of the total amount of distributed items. This means that on average around 90,000 items are distributed from BFC. In peak periods in the months November and December, this amount can be 2.5 times as large. The design of the warehouse was based on the demand during peak periods, which means that the stock capacity of BFC is 8.5 million items. Currently, a second bol.com fulfillment center (BFC2) is being built next to BFC. It is expected that BFC2 starts operating in April 2021.

2.2 Logistical process at bol.com

The logistical process at bol.com can be described as follows. The products of bol.com suppliers and its Lvb partners arrive at the warehouse. The inbound process starts, which includes the unloading, receiving and put away of all items. Next, the items are kept in stock until they are needed to fulfill a customer order. In that case, the outbound process starts, which includes picking and packing such that the items are ready for transport. Finally, the items are distributed to the customers. This step is outsourced to several transport carriers. If the customer is unsatisfied with an item, it can be sent back with a transport carrier as well. There is one warehouse that receives and processes the returned items, namely bol.com

retour center (*BRC*). Therefore the return flow is not part of each warehouse. In Figure 2.1, a simplified overview is given of the logistical process.

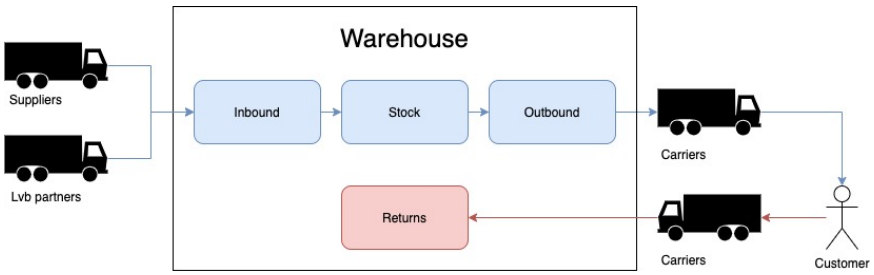


Figure 2.1: Simplified overview of the logistical process

As mentioned before, bol.com has six different warehouses and a seventh is currently being built. The current warehouses are the following:

1. BFC is the most automated warehouse, which stores small to medium size items that fit into totes. A *tote* is a blue bin used by an order picker to temporarily store the items of a pick batch being collected during a pick tour, and to transport items through the warehouse. Bol.com is the owner of this warehouse but the operations are outsourced to Ingram Micro.

2. Centraal Broekhuis is a warehouse that is used by other companies than bol.com as well. Most of the books, CDs and DVDs of bol.com are distributed from here.

3. Veerweg is the first warehouse that bol.com opened when they expanded their assortment with other products than those stored at Centraal Broekhuis. Small to large size items are stored here. Bol.com leases this building and, similar to BFC, the operations are outsourced to Ingram Micro.

4. BFC XL stores the extra large products of bol.com, such as dish washers and fridges.

5. Amsterdam Hub is a small distribution centre, which is used for same-day deliveries.

6. BRC is the warehouse that processes all items returned by customers.

In this research project the focus is put on BFC. Therefore, the information in the remainder of this report is only based on BFC and does not necessarily hold for the other warehouses too.

2.3 Outbound process BFC

The outbound process of BFC includes the batching, picking and packing of orders, such that they are ready for delivery by the transport carriers. As bol.com has multiple warehouses, it could be the case that a customer order must be fulfilled from multiple warehouses. In this report an *order* is therefore defined as the items of one customer order that are fulfilled from BFC. Besides that, bol.com distinguishes between mono and multi orders. A *mono order* is an order consisting of a single item. A *multi order* is an order consisting of more than one item. Quick reminder, an *item* is defined as a single physical unit of an international article number (EAN), so for example an order of two identical pens would be considered a multi order.

The outbound process of mono orders consists of three steps: batching, picking and packing. The outbound process of multi orders is similar but includes a sorting operation before packing. Currently, the process is driven by the packing operation. As different kinds of products require different kinds of packaging, BFC has different types of packaging lines. For example, some items can be packed by an automatic carton wrapper, whereas others require manual packing. These different packaging lines are called *outbound lines*. An overview of the different outbound lines of the mono and multi orders is given in Table 2.1.

Outbound line	**Name**	**Description**
101	Mono High Risk	High valued items
102	Mono Manual Value Added Service	Including giftwrap or wish-card
103	Mono Manual Regular	Requiring manual boxing
104	Mono Smartmailer	Small mailbox items
105	Mono Cartonwrap	Allowed to be mechanically packed
106	Multi High Risk	High valued items
107	Multi Manual Value Added Service	Including giftwrap or wish-card
108	Multi Manual Regular	Requiring manual boxing
109	Multi Automatic Sorting	Allowed to be mechanically sorted and packed

Table 2.1: Overview of outbound lines

Each of the outbound lines has a specific number and name. The items are classified based on their weight, volume and characteristics such as flammability, sharpness, or fragility. Given the classification of an item, it can then be assigned to a specific outbound line. In some cases, items can be processed on multiple outbound lines. An item that is ordinarily assigned to outbound line 105, could also

be transferred to outbound line 103 in case it gets too busy at outbound line 105 for example.

The control room is the department that is responsible for monitoring the outbound process. They need to make sure that the workload is spread over the different outbound lines and that at the end of the day the fulfillment score has been reached. The *fulfillment score* is defined as the percentage of orders delivered in time at the assigned transport carrier. The main task of the control room, in order to fulfil these responsibilities, is to determine how many orders should be batched and released for picking. This decision is based on the capacity of the outbound lines and presented in number of orders per outbound line. The control room needs to make this decision several times per day. There is no fixed time schedule for the release of new pick batches. It is up to the control room to keep track of when the capacity of the outbound lines allows for a new release of orders, and to determine how many new pick batches will then be released into the system at what time.

2.3.1 Order batching decision making process

The control room bases their initial decision, on the order batching quantities per outbound line, on the production plan. This production plan is based on the demand forecasts and states the number of packages that must be produced per outbound line per day. Based on this information and the capacity of the different stations, it is determined how much should be produced per hour at each outbound line. The order production quantity per hour is not requested to be batched at once but the control room does this in several stages, which makes it easier to intervene.

An intervention could be to switch the outbound line of a specific number of orders if the designated outbound line is not able to process all its orders. This happens frequently with outbound line 105 (mono carton wrap) as a lot of items can be processed mechanically. In this case orders are reassigned to outbound line 103 (mono manual regular). Similarly, items assigned to outbound line 109 (multi automatic sorting) could be reassigned to outbound line 108 (multi manual regular). Besides that, it is possible to send mono orders to the multi order packing lines but not the other way around. This is due to the sorting step that is required for multi orders but not for mono orders.

Another intervention is the prioritization of orders from a specific priority group. There are three priority groups, namely priority 3, priority 4 and priority 9. Priority 3 orders have to be picked as soon as possible in order to arrive in time at the customer. Priority 4 orders can be picked straightaway but would still arrive in time at the customer if they are picked at a later time. Priority 9 means that the order may not be selected, so these orders will not be planned. The priority is based on the outbound line, delivery date and cut-off times of the transport

carrier. The control room makes use of the option to prioritize orders if, for example, it is near the cut-off time of a transport carrier. The control room can only request a specific number of orders that are all high in priority. It is not possible to request for a specific order to be batched.

2.3.2 Warehouse Execution Service

To help the control room in the decision making process, the company created *WES*, which stands for Warehouse Execution Service. This service aims to transform data about the batches currently being processed in the outbound area of BFC, such that a batch advice can be generated. The idea behind this service is that the batch advice should eventually ensure a steady flow towards the picking and packing areas in order to eliminate operator standstill and thus maximize efficiency. Currently, this service is in the first iteration and includes the following:

1. A dashboard that functions as a control panel for the control room
2. Insight in buffer amounts at the picking and packing areas
3. A first version of the batch advice
4. Only the mono outbound lines
5. Connection with the service *DES* (Discrete Event Simulation), which provides a forecast of the future status of the system based on real time data

The dashboard includes an overview of picking and packing for each mono outbound line. An example of a picking overview is given in Figure 2.2.

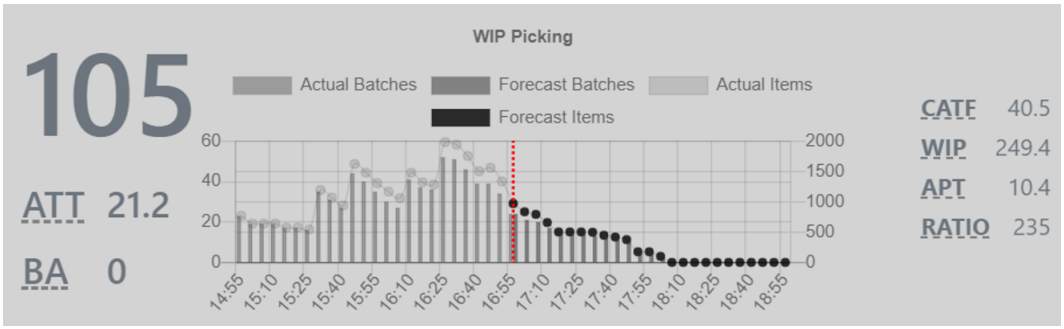


Figure 2.2: WES picking overview of an outbound line

In the graph, the bars represent the amount of batches (left axis) and the dots represent the amount of items (right axis) in the system. *ATT* stands for Average Transportation Time in minutes and equals the average time to transport a tote from picking to packing. *BA* stands for Batch Advice and is based on the data on

the right hand side of the graph. *CATF* is the Current Average ToteFill, which equals the number of items per tote. *WIP* is the Work in Progress in minutes, which equals the total workload in the system for all operators together. *APT* is the Average Processing Time in minutes, which equals the average time to pick a pick batch. *RATIO* is the average processing rate, which equals the items per hour per operator.

The control room has a similar overview for packing, as depicted in Figure 2.3.

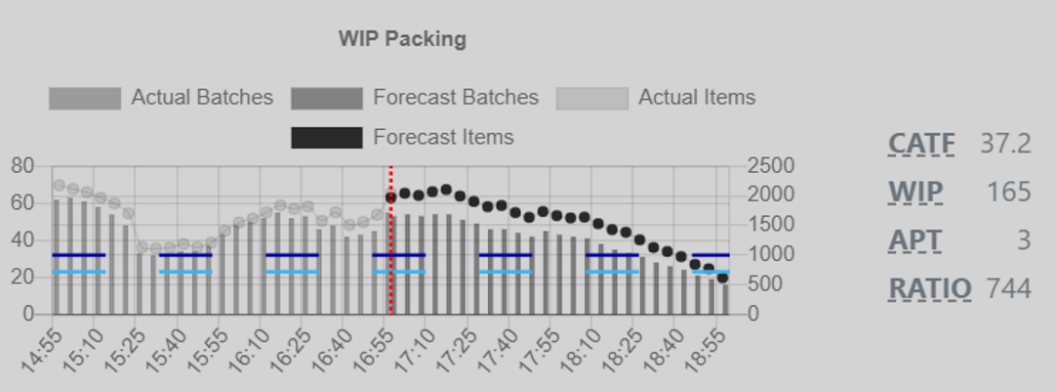


Figure 2.3: WES packing overview of an outbound line

The red line in this figure equals the current time. Everything on the left side of the red line shows the actual status in the past and everything on the right side of the red line shows the forecasts generated by DES. In order to generate a batch advice, the control room needs to fill out the fields in the manual input section of the dashboard, which is depicted in Figure 2.4 on the next page.

The control room needs to fill in the time they want to cover with the new pick batches, the number of picking operators, the number of packing operators per outbound line, and how big the packing buffer should be per outbound line. Based on this information and the real time data of the system, the upper and lower limit of the batch advice is calculated as follows:

$$Upper\ Limit = \frac{Manual\ input\ packing\ buffer\ in\ minutes}{APT\ packing} \times packers + packers \tag{2.1}$$

$$Lower\ Limit = \frac{APT\ picking + ATT}{APT\ packing} \times packers + packers \tag{2.2}$$

The upper limit is the amount of batches needed in the buffer, in order to match the buffer amount filled out by the control room in the manual input screen. The lower limit is the minimum amount of batches needed in the buffer at any time. If the amount of batches goes below this limit, it will result in idle time for the

packing operators. The reason for this is that the orders in the packing buffer are processed more quickly by the operator than new batches are delivered. The actual batch advice, which equals the amount that should be batched at this moment to fill the packing buffers up to the upper limit is then calculated as follows:

$$Batch\ Advice = (Upper\ Limit - Current\ batches\ packing) \times CATF \tag{2.3}$$

The current status of WES is that the current iteration works fine and is a sound basis for future iterations. The control room uses the tool and experiments with it but the lack of multi lines and a couple of other issues prevent the control room from using the tool as their main batching tool. These issues include a missing link with the production plan, exclusion of the stingray (buffer zone) and generated batch advice that is larger than the number of waiting orders.

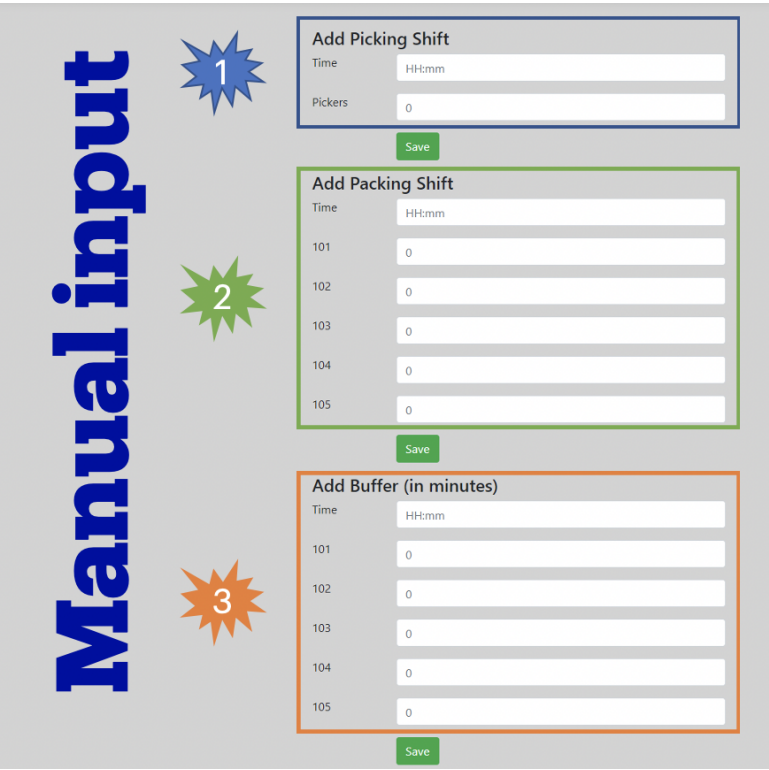


Figure 2.4: WES dashboard manual input

2.3.3 Creation of order batches

Once the control room has determined the amount of orders they want to be released for each outbound line, they put these numbers in the warehouse management system called Reflex. Subsequently, Pacman, which is a microservice implementation of the picking algorithm used by Reflex, is called upon and determines which orders will be batched and which zones in the warehouse are used for

batching. A list of picks is then forwarded to the system Blinky, which optimizes the walking routes of these picks and creates optimized pick batches. Based on the information that Pacman receives from Blinky, it creates pick batches for Reflex. These pick batches are then used by the operators in the warehouse to pick all the items. In Figure 2.5 an overview is given of the information flow between these systems.

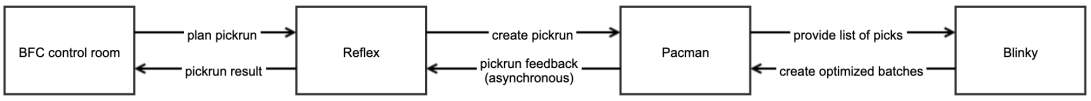


Figure 2.5: Overview data flow for pick runs

The picking algorithm implemented by Pacman requires the following input: required number of orders per outbound line, required number of orders to be transferred from one outbound line to another, required prioritization, list of unfulfilled orders, current stock levels and possible reservations. Based on this input a filtered list of unfulfilled orders is created. Next, the orders need to be sorted. This is done in the following order:

1. Outbound line
2. Priority (ascending)
3. Cut-off time (ascending)
4. Delivery date (ascending)
5. Difficulty (descending)

Once the orders have been sorted, they are allocated to a picking zone in the warehouse and pools can be created. A distinction is made between mono and multi orders in the creation of pools.

For mono orders holds that after the order with the highest priority has been selected, the picking zones in the warehouse are ranked based on availability and the zone with the highest number of potential picks is chosen. Next, orders that can be picked from the chosen zone are selected based on their priority in the sorted list and assigned to the same pool until the maximum capacity of the pool has been reached, which equals the maximum number of picks per zone. In case no unfulfilled orders can be picked from the same zone anymore and the maximum capacity has not been reached yet, the current pool is closed and a new pool is created. The process is repeated until the number of requested orders per outbound line by the control room is reached. The result is a list of pools per outbound line, where each pool consists of orders that will all be picked from the same zone in the warehouse.

For multi orders holds that each multi pool has a maximum pool size, which is a configurable number that determines the maximum amount of orders that can be included in a pool and is determined by the capacity of the outbound line. Similarly to the mono orders, the order with the highest priority is selected first. It could be the case that the items from this order must be picked from different zones in the warehouse. Subsequently, orders that can be picked from the same zones as the order with the highest priority are added to the pool, until the maximum pool size has been reached. In case the requested number of orders by the control room has not been reached yet, it becomes possible to also select orders that require a visit to an additional zone. This process is repeated until the number of requested orders per outbound line by the control room is reached. A pool of multi orders can thus contain items that are spread over multiple zones in the warehouse. Therefore, the result in this case is a list of pools per outbound line, where each pool consists of a list of zones with for each zone a list of picks.

Given the lists of picks per pool and zone, pick batches can be created. In the creation of pick batches, the physical limits of a tote such as volume and weight need to be respected. The picks are sorted and placed in totes on alphabetical location. Each tote corresponds to a pick batch. These are not the optimal pick batches, because those are created by Blinky. However, these simple pick batches serve as a fallback in case no solution was found by Blinky, such that at least a solution is provided to Reflex. Finally, the lists of picks per pool and zone are communicated to Blinky, which creates pick batches by defining and solving a vehicle routing problem with capacity constraints. Once a solution has been found, Blinky returns a list of pick batches per pool and zone. The result of Blinky is checked against the fallback solution and the result with the shortest distance is sent back to Reflex.

Chapter 3

Problem formulation

The previous chapter provided some background information to obtain a better understanding of the research environment and brought the problems in the current situation to light. This chapter describes the problem that this research is focused on. The problem statement is given in Section 3.1. This is followed by a description of the goal in Section 3.2 and the scope in Section 3.3. The chapter closes with a description of the research approach in Section 3.4.

3.1 Problem statement

Currently, the workload allocation is done manually at bol.com by the control room. In order to help the control room in the decision making process, the company created WES. However, this tool can only be used for the mono outbound lines, because the multi outbound lines and the stingray (buffer zone) are excluded. Besides that, there is a missing link with the production plan and the generated batch advice is frequently larger than the number of waiting orders. Furthermore, the calculation of the batchsize is only based on the packing centers and does not take into account the work in progress at picking and on the conveyor belts. Therefore, there is a need for better batching advice, which takes into consideration the workload allocation throughout the entire process and includes all outbound lines and the stingray.

3.2 Goal

The goal is to generate an automatic batch advice, which ensures that the workload at the different work stations in the warehouse is allocated in such a manner that the overall performance measures are optimized. From a bol.com perspective the

objective is to maximize the production (or throughput) while minimizing costs. Translating this objective to performance measures related to queueing theory, this means that the throughput should be maximized and the station/operator idle time should be minimized. In addition, the company's ultimate performance indicator, which is the order fulfillment score with a target of 99%, must remain.

3.3 Scope

Bol.com has six different warehouses but the scope of this project is limited to the Bol.com Fulfillment Center (BFC). This is the largest warehouse and distributes 40%-50% of the total amount of distributed items. The process being examined includes the following: receival of orders at the warehouse, order batching, picking, packing, and the transferal of packed orders to third party transport carriers. The transport and buffer locations between the different stations are included as well. Besides that, the capacity of the transport carriers is within the scope of this research project but the actual delivery from BFC to the customer is not within the scope. In addition, the (re)assignment of orders to outbound lines and the reallocation of operators on outbound lines is considered to be within the scope.

3.4 Research approach

The described process in the warehouse can be modelled as a multi-class open queueing network of multi-server queues. In the past decades, quite some research has been done into the application of networks of queues in computer, communication and production systems [35]. Besides that, a reasonable amount of research on this topic can be found in relation to the manufacturing and health care sector. Research about the application of queueing networks in e-fulfillment is however limited. Therefore, the methods being used in aforementioned sectors is examined, to evaluate whether (elements of) these methods can be used as input for this research project alongside the general theory behind queueing networks.

Once the process has been modelled as a multi-class open queueing network of multi-server queues and the arrival and processing times of the different stations have been retrieved through data analysis, it can be determined how the workload should be allocated such that the overall performance measures are optimized.

In order to achieve the goal as stated in Section 3.2, the following research question needs to be answered:
"How should the workload at the different work stations in the warehouse be allocated such that the overall throughput is maximized and the operating costs are minimized, while maintaining the order fulfillment score?"

This question will be answered following the subsequent sub questions:

1. What can be found in literature about modelling networks of queues?
2. What can be found in literature about solving workload allocation problems in networks of queues?
3. How can the process in the warehouse be modelled as a network of queues?
4. Which methods from literature to solve the workload allocation problem are applicable to the situation of bol.com?
5. How can it be determined which method provides the best results for bol.com?

An extensive literature research is performed to answer the first and second sub questions, the summary of this research is given in Chapter 4. Based on the information retrieved in the literature research, the process in the warehouse can be modelled as a queueing network. The model is presented in Chapter 5 and answers the third sub question. In Chapter 7, the solution approach and experimental setup are described, which provides an answer to sub questions four and five. Subsequently, the experimental results are presented in Chapter 8, followed by a discussion in Chapter 9, and the conclusions in Chapter 10. These three chapters together provide an answer to the main research question.

Chapter 4

Literature research

Queues help facilities and businesses to provide service in an organized manner. As forming a queue is a social phenomenon, it would be favorable to regulate the queue in such a way that is most beneficial for both the unit that waits and the one that provides the service. The unit waiting for service, irrespective whether it is human or otherwise, is identified as the "customer" and the unit providing the service is known as the "server". [25]

In queueing theory one analyses the mode by which a queue is formed and the service is provided. This is done by building a mathematical model of which the basic elements include the customer arrival process, service mechanism, system capacity, and queueing discipline. D. G. Kendall introduced a shorthand notation to characterize the arrival process, service times, number of servers, and capacity of the system by symbols. It is a four-part code $a/b/c/d$. The first letter specifies the inter-arrival time distribution and the second letter the service time distribution. For instance, the letter M is used for a Poisson or exponential distribution and refers to the "Markovian" or "Memory-less" property of this distribution, G stands for general distributions, D for deterministic distributions, and E_k for the Erlang distribution with scale parameter k. The third letter specifies the number of servers and the fourth and final letter specifies the capacity of the system, which includes the capacity of the queue and the customer in service. However, if the capacity is regarded as infinity, the fourth letter is often omitted. [25, 1]

The focus of research on networks of queues is primarily on performance evaluation and can be divided into three categories: exact analysis, approximation methods, and simulation and related techniques. Exact results only exist for systems with the following assumptions:

1. Poisson arrivals
2. Exponentially distributed and customer class independent service times
3. Customer class independent priority discipline

In most practical situations however, these assumptions are too restrictive for a good resemblance of reality. Therefore, researchers started to develop approximations to evaluate the performance measures. Besides that, discrete event Monte-Carlo simulation became an alternative to evaluate large queueing networks with a closer resemblance of reality. [9]

The structure of this chapter is as follows. In Section 4.1, the well known Jackson Network for exact analysis is described. This is followed by descriptions of several approximation methods for multi-class queueing models in sections 4.2 to 4.7. For each method first the characteristics and assumptions of the queueing network are described. This is followed by a description of how the performance measures such as queue length, throughput, sojourn and/or waiting times can be calculated. The chapter closes with a comparison of the discussed models in Section 4.8.

4.1 Jackson Network

A Jackson Network is a single-class open queueing network with $M \geq 1$ stations, $c_i \geq 1$ servers at station i, and exponentially distributed service times with parameter $\mu_i > 0$. All stations have a first come first served (FCFS) policy. Customers arrive from outside the network at station i according to a Poisson process with intensity γ_i. The routing of customers in the network is Markovian, which is characterized by an irreducible routing matrix P. This means that when a customer is finished at station i he will either go to station j with probability P_{ij}, or leave the network with probability $P_{i0} = 1 - \sum_{j \neq 0} P_{ij}$. [41, 23, 19]

The arrival rate at station i consists of arrivals from outside the network as well as arrivals from other stations inside the network, and can be determined with:

$$\lambda_i = \gamma_i + \sum_{j=1}^{M} \lambda_j P_{ji}, \qquad i = 1, .., M. \tag{4.1}$$

The visit ratio's of the stations are denoted by V_i, $i = 1, ..., M$. For open queueing networks holds:

$$V_i = \frac{\lambda_i}{\gamma}, \qquad i = 1, ..., M, \tag{4.2}$$

where $\gamma = \sum_i \gamma_i$.

The stationary distribution has a product-form solution, we refer the interested reader to [23, 19] for the proof. For a Jackson Network with M stations, each having c_i servers, the stationary distribution is given by:

$$\pi(n_1, ..., n_M) = \prod_{i=1}^{M} f_i(n_i), \tag{4.3}$$

with

$$f_i(n_i) = \frac{1}{G(i)} \frac{(c_i\rho_i)^{n_i}}{n_i!}, \qquad n_i < c_i \tag{4.4}$$

and

$$f_i(n_i) = \frac{1}{G(i)} \frac{c_i^{c_i}\rho_i^{n_i}}{c_i!}, \qquad n_i \geq c_i \tag{4.5}$$

where ρ_i is the utilization of station i, determined by:

$$\rho_i = \frac{\lambda_i}{c_i\mu_i}, \qquad \rho_i < 1, \tag{4.6}$$

and the normalization constant $G(i)$ is defined as:

$$G(i) = \sum_{n=0}^{c_i-1} \frac{(c_i\rho_i)^n}{n!} + \frac{(c_i\rho_i)^{c_i}}{c_i!}(1-\rho_i)^{-1}. \tag{4.7}$$

This product form solution only holds for stable networks, so when $\rho_i < 1$ for all i. [41, 23, 19]

As the probability of having n customers at station i is independent of the state of all other stations in the network (see equations 4.3 through 4.7), the performance measures may be computed for each individual station separately and can then be added up to obtain the measures for the whole network [41].

The throughput is defined as:

$$TH = \gamma = \sum_{i=1}^{M} \gamma_i. \tag{4.8}$$

The expected number of customers at station i is:

$$\mathbb{E}L_i = \frac{(c_i\rho_i)^{c_i}}{c_i!\,G(i)} \frac{\rho_i}{(1-\rho_i)^2} + c_i\rho_i. \tag{4.9}$$

Adding up the expected number of customers at each station, the total number of customers in the network is obtained:

$$\mathbb{E}L = \sum_{i=1}^{M} \mathbb{E}L_i. \tag{4.10}$$

The expected time of a customer at station i can be obtained using Little's law:

$$\mathbb{E}W_i = \frac{\mathbb{E}L_i}{\lambda_i}. \tag{4.11}$$

Finally, using the visit ratio's, the expected time in the system (sojourn time) of a customer can be calculated with:

$$\mathbb{E}W = \sum_{i=1}^{M} V_i\mathbb{E}W_i = \sum_{i=1}^{M} V_i \frac{\mathbb{E}L_i}{\lambda_i}. \tag{4.12}$$

4.2 Complete reduction method

Complete reduction methods aim to provide a way to obtain performance measures for every single customer class on both a network and a station level by reducing the multi-class network to a single-class network. There exist several variations on this approach. For example, Conway and Georganas [14] proposed a method for closed multi-class networks of FCFS queues, in which they transform the network into a network of processor-sharing queues with a hierarchy of subsystems associated with subsets of the classes. Another method, which is applicable to open multi-class queueing networks instead, is proposed by Zijm [41].

The main idea is as follows. Consider a multi-class open queueing network with M stations, R classes, general individual inter-arrival and service time distributions, and routing matrices $P^{(r)}$. The complete reduction method consists of three steps [41]:

1. Reducing the R-class open queueing network to a single-class open queueing network by aggregating the classes.
2. Analyzing the single-class open queueing network.
3. Disaggregating to obtain the performance measures per class for the given R-class open queueing network.

The first step reduces the $(4M + M^2)R + M$ input parameters to $5M + M^2$ parameters, which makes the algorithm computationally more efficient. In order to achieve this, the service times, arrival rates and the routing probabilities are aggregated.

The aggregate first and second moment of the service times at the stations are given by the weighted average of the service times of the individual job classes, where the weights are based on the arrival rates. With this first and second moment, the aggregate squared coefficient of variation (SCV) of the service time can also be determined. Similarly, the aggregate routing matrix is given by the weighted average of the routing matrix of the individual job classes, where the weights are based on the arrival rates.

The aggregate arrival rate is simply a sum over the arrival rates of the different classes. However, obtaining the aggregate SCV of the arrival process is a bit more complicated. An approximation can be obtained by taking the superposition of the R job flows, which is described by a set of linear equations. This approach is thoroughly described by Albin [3] and Whitt [38], among others.

With the aggregated input, the performance measures of the aggregated job can be approximated as in a single-class open queueing network. Finally, these results are disaggregated in order to obtain the performance measures per class.

4.3 Decomposition method

Decomposition methods aim to provide a way to deal with each station in a queueing network in isolation by approximating the parameters of each station. This is done by first computing the arrival rates exactly by means of the same traffic rate equations as for product-form networks (e.g. Jackson networks). Next, the SCV of the arrival process for each station is computed with a set of approximate formulas. If the service times are not exponentially distributed, the parameters are approximated as well. Then, given the parameters of the arrival and service time distributions, the SCV of the inter-departure times can be computed. Different variations of this approach have been suggested, among others, by Bitran and Tirupati [9], Whitt [39], Satyam et al. [29], Kim [20], and Caldentey [12]. Their methods for open multi-class queueing networks are however restricted to systems with only single-server nodes.

An alternative method for multi-class multi-server queueing networks was introduced by Rabta et al. [28]. They proposed a hybrid solution of the classical decomposition analysis and simulation techniques. The computation of the SCV of the inter-departure times by means of approximate formulas is replaced by estimation through simulation based on a set of recursive equations in this case. According to their research this results in better estimates of the performance measures than the original decomposition algorithms. Besides that, it is faster and easier to implement and leads to lower variance of the estimators than full simulation.

4.4 Fluid models

A considerable amount of literature can be found about approximating the performance measures of a queueing system by using a fluid flow model. However, most of the research is focused on single queues or networks with single-server stations and only one customer class. Models for multi-class systems have been introduced by Bertsimas, Gamarnik, and Tsitsiklis [8] and Bertsimas, Gamarnik, and Rikun [7], for example. However, these models only work with single-server stations. Bassamboo et al. [6] and Whitt [37] both introduced a multi-class multi-server model, which also includes customer abandonment. In these models, each customer class has its own buffer (waiting queue) and only visits one station. The methods are however not applicable to a network of queues and stations.

The basic idea is to model the information flow as a fluid flow and then describe it by a system of ordinary differential equations. An important note is that the method described is deterministic. Therefore, effects of random arrivals or variations in service times cannot be studied directly. The model is applicable to any processing situation that can be described by a set of flow diagrams, which show

the order of the processing steps. A flow diagram consists of boxes and links. The boxes may represent work stations, computer system components or other processing units. The links or arrows show how the information flows from one box to another. It is possible to attach percentages to some of the links to create different paths. Characteristics of the process that can be determined with this technique include the throughput, waiting time, sojourn time, and utilization of resources. These measures can be computed at a selected time interval.

4.5 BCMP theorem

BCMP stands for Baskett, Chandy, Muntz and Palacios who were the founders of the theorem. They extended the work of Gordon and Newell, who focused on product form solutions for single-class closed queueing networks, to multi-class closed queueing networks. The networks satisfying the conditions for the BCMP theorem are known as BCMP networks, which are multi-class closed queueing networks with $M \geq 1$ stations, $R \geq 1$ classes, $N^{(r)} \geq 1$ class r customers, and visit ratio's $V_i^{(r)}$. Each station i has one of the following service disciplines: first come first served ($FCFS$), last come first served preemptive resume ($LCFS$), processor sharing (PS), or ample server (AS). The service times at station i have mean value $1/\mu_i^{(r)} \geq 0$ for class-r customers. For stations with the FCFS service discipline, the service times should be exponential and class-independent but for stations with one of the other service disciplines, the service times may also be general and class-dependent. The routing through the system is Markovian, characterized by the irreducible routing matrix $P^{(r)}$ for class r. [23, 41, 5]

The BCMP theorem is only concerned with the number of customers per class in the queue but does not consider the exact sequence of these customers. The state space of this stochastic process is given by $\mathbf{S}_{BCMP} = \{(\vec{n_1}, ..., \vec{n_M}) | \sum_{i=1}^{M} n_i^{(r)} = N^{(r)}, r = 1, ..., R\}$ in which $\vec{n_i} = (n_i^{(1)}, ..., n_i^{(R)})$ and $n_i^{(r)}$ denotes the number of class r customers at station i. The vector $\vec{N} = (N^{(1)}, ..., N^{(R)})$ gives the population of the network. [23, 41]

The BCMP theorem then states the following [5]:

"The detailed Markov process, that describes the behavior of the BCMP network, has a unique stationary distribution and the aggregated stationary probabilities $\pi(\cdot)$ for the aggregate states $\vec{n} = (\vec{n_1}, ..., \vec{n_M})$ are given by:

$$\pi(\vec{n_1}, ..., \vec{n_M}) = \frac{1}{G(\vec{N})} \prod_{i=1}^{M} f_i(\vec{n_i}), \tag{4.13}$$

where the normalization constant equals

$$G(\vec{N}) = \sum_{\vec{n} \in \mathbf{S}_{BCMP}} \prod_{i=1}^{M} f_i(\vec{n_i}), \tag{4.14}$$

and for each station i the function $f_i(\vec{n_i})$ *is defined as*

$$f_i(\vec{n_i}) = \begin{cases} \frac{n_i!}{\prod_{k=1}^{n_i} min(k,c_i)} \prod_{r=1}^{R} \frac{1}{n_i^{(r)}!} \left(\frac{V_i^{(r)}}{\mu_i^{(r)}} \right)^{n_i^{(r)}} & \text{if } i \text{ is } FCFS, LCFS, PS \\ \prod_{r=1}^{R} \frac{1}{n_i^{(r)}!} \left(\frac{V_i^{(r)}}{\mu_i^{(r)}} \right)^{n_i^{(r)}} & \text{if } i \text{ is } AS \end{cases} \tag{4.15}$$

where $n_i = |\vec{n_i}| = \sum_{r=1}^{R} n_i^{(r)}$ *denotes the total number of customers at station i."*

In principle, the relevant performance measures such as the mean number of customers and waiting time at a station can be determined from the stationary distribution by means of the normalization constant. As there are too many states to calculate this within reasonable time, it is better to make use of the generalized arrival theorem first presented by Lavenberg and Reiser [21]:

"Let $C(\vec{N})$ *be a BCMP network with population* $\vec{N}$*. Denote by* $p((\vec{n_1}, ..., \vec{n_M})|\vec{N})$ *the equilibrium probability of* $C(\vec{N})$*, and by* $p_a^{(r)}((\vec{n_1}, ..., \vec{n_M})|\vec{N})$ *the equilibrium probability that, at a class-r arrival instant at an arbitrary station, the state of* $C(\vec{N})$ *is* $(\vec{n_1}, ..., \vec{n_M})$*. Then:*

$$p_a^{(r)}((\vec{n_1}, ..., \vec{n_M})|\vec{N}) = p((\vec{n_1}, ..., \vec{n_M})|\vec{N} - \vec{e_r}),$$

where $\vec{e_r}$ *is the R-dimensional unit vector with 1 at position r."*

With the generalized arrival theorem, the performance measures (which are state dependent) can then be calculated, in case all stations have a FCFS discipline, with a multi-class marginal distribution analysis [41].

4.6 Workload controlled manufacturing systems

The workload controlled manufacturing system can be modelled as a closed queueing network when we assume that there is always another customer waiting to enter the network as soon as a customer leaves. Another assumption to be made in this case is that the product form solution and generalized arrival theorem are still valid in closed queueing networks with service times that have a general distribution and are class-dependent. Under these assumptions, the performance measures can be approximated with the approximate mean value analysis. [41]

Several variants of the approximate mean value analysis have been introduced. Zhang and Down [40] introduced a numerically stable mean value analysis for single-class single-server product-form networks and extended it to multi-class networks too. Other methods for multi-class single-server queueing networks were introduced by Petriu and Woodside [27] and Eager and Sorin [17]. Methods for multi-class multi-server queueing networks have been introduced, among others, by Akyldiz and Bolch [2] and Zijm [41].

The approximate mean value analysis calculates the performance measures for each possible state of the system. As a result the computational complexity is of the order $MNR\prod_{r=1}^{R}(N^{(r)}+1)$, where M is the number of stations, R the number of classes, N the maximum number of customers in the system, and $N^{(r)}$ the maximum number of customers in the system of class r. However, the complexity can be reduced by using the generalized arrival theorem and the variable $Q_j(\vec{n})$, which is the probability that all servers at station j are busy in state $\vec{n} = (\vec{n_1}, ..., \vec{n_M})$ and is only dependent on the marginal probabilities from the previous state of the system. Using this variable, the marginal probability of having zero customers at a station in a certain state, which is normally calculated using all marginal probabilities in that state, now only requires the marginal probabilities $p_j(c_j - 1|\vec{n} - \vec{e_r})$, where $\vec{e_r}$ is the R-dimensional unit vector with 1 at position r and $r = 1, ..., R$. Consequently, the complexity is reduced to $MRc^* \prod_{r=1}^{R}(N^{(r)}+1)$ where $c^* = \max\{c_1, ..., c_M\}$, the maximum number of servers at a station.

4.7 PAC systems

PAC stands for Production Authorization Card. These cards are used in manufacturing systems and job shops to keep control over the number of jobs in the system and thus put a limit on the work in progress. In relation to queueing theory, such a system can be modeled as a closed or restricted open network. The customers then arrive in an external queue and PACs are used to control the release of customers to the system.

An extensive amount of research has been done in this field for single-class production to order systems. Several methods have been introduced, among others, by Buzacott and Shanthikumar [11], Siha [31], Avi-Itzhak and Heyman [4], Buitenhek [10], and Dallery [16]. Buitenhek [10] also presented four approximation methods for the analysis of multi-class production to order systems with universal PACs. Besides that, Perros et al. [26] proposed a generalization of the method of Dallery [16] for multi-class production to order systems with dedicated cards. They assumed that a product form solution exists, which means that the service times of each station are exponentially distributed and class-independent and the queueing discipline is first come first served. Consecutively, Zijm [41] generalized the method of Perros et al. [26] for multi-class queueing networks with stations that have general class-dependent service times.

The closed queueing network view of the production system with dedicated PACs corresponds with a multi-class closed queueing network with M general service stations and R synchronization stations, which each consist of an external queue and a card pool. Each synchronization station is dedicated to one customer class and should be numbered such that station $M + r$ is the class-r synchronization station. There are $N^{(r)}$ dedicated PACs available for customer class r. Class-r customers arrive at their own external queue and may enter the network if there is a class-r PAC available in the class-r card pool. As soon as a customer leaves the system, the PAC is returned to the class-r card pool, where it is available for use by the next customer in line. In this system the PACs are thus the main entities in the system instead of the customers.

The following assumptions are made: class-r customers arrive according to a Poisson process, service times of class-r jobs have a general distribution, the service discipline is FCFS at each station, and class-r customers circulate inside the network according to a Markovian routing matrix. The idea is to replace the synchronization stations by load-dependent servers. The service rates of the load-dependent servers should be such that the throughput of the PACs at station $M + r$ equals the class-r arrival rate.

This replacement is not trivial, because the throughput at a synchronization station is dependent on all other synchronization stations, which makes it a fixed-point problem. In order to solve this problem, Norton's theorem can be used [13]. In principle, the idea behind the theorem is to decompose a network with M stations into an equivalent network with two stations. One station is identical to the one in the original network and the other station replaces the rest of the original network by a load-dependent exponential server. The following iterative procedure can then be used to find the throughput of the synchronization stations:

1. Set the service rates for stations $M + 1$ to $M + R$ to some initial value.
2. Repeat the following steps until convergence of the service rates:
   (a) Solve the equivalent network of synchronization stations $M + r$ with an appropriate approximate mean value analysis, for all classes $r = 1, ..., R$. This yields the throughput of the synchronization stations.
   (b) Reset the service rates.

Once the service rates of the synchronization station have been obtained, the same approximate mean value analysis can be used to analyze the performance measures of the complete closed queueing network with load-dependent synchronization stations.

4.8 Comparison of models

In this section all previously introduced models are compared. First, the main characteristics of the models are examined, namely whether it is stochastic or deterministic, open or closed, single- or multi-class, and single- or multi-server. An overview of the main characteristics of the models is given in Table 4.1.

Model	**Type**	**Form**	**Class**	**Server**
Jackson Network	Stochastic	Open	Single	Multi
Complete reduction	Stochastic	Open	Multi	Multi
Decomposition	Stochastic	Open	Multi	Multi
Fluid	Deterministic	Open	Single	Multi
BCMP	Stochastic	Closed	Multi	Multi
Workload control	Stochastic	Closed	Multi	Multi
PAC	Stochastic	Closed	Multi	Multi

Table 4.1: Main characteristics of the queueing models

The output of deterministic models is fully determined by the parameter values and the initial conditions. Stochastic models, on the other hand, allow for random variables as input and therefore account for some inherent variation. In this case, the same set of parameter values and initial conditions leads to an ensemble of different outputs of the model. If one would like to take into account uncertainties of the inputs with a deterministic model, Monte Carlo simulation could be applied but this is computationally inefficient. As the BFC outbound process contains a significant amount of variation, the process is best described by a stochastic model.

Open queueing networks receive customers from an external source and after processing the customers leave the network to an external destination. Closed queueing networks have a fixed population in the network that moves between the queues and never leaves the system. If it can be assumed that there is always a new customer waiting to enter the network as soon as a customer leaves the network, an open queueing network could also be modelled as a closed queueing network without sacrificing too much accuracy. Depending on the structure of the network, it is sometimes easier to analyse the closed form of the network because of the finite set of equations. [25] In principle, the BFC outbound process is an open queueing network. However, it could also be modelled as a closed queueing network, because there are generally more than enough orders waiting to be processed.

Multi-class queueing networks allow for different customer classes with different routing matrices and service time distributions. Single-class queueing networks assume that every customer entering the system has the same routing matrix and service time distributions. The BFC outbound process is a multi-class queuing

network. For some stations the service time distributions are class independent but this does not hold for all of them. Besides that, each class has a different route through the system.

Since all models allow for multi-server stations, there is no need to compare the models on this matter. Besides the previously discussed characteristics, we are also interested in the arrival and service time distributions supported by the models and whether the performance measures are state dependent or not. An overview is given in Table 4.2. An important note here is that if a model supports general distributions, this means that the model can also be used with Poisson arrivals and exponentially distributed service times.

Model	**Arrival distribution**	**Service time distribution**	**State dependence**
Jackson Network	Poisson	Exponential	No
Complete reduction	General	General	No
Decomposition	General	General	No
Fluid	Constant	Constant	No
BCMP	n.a.	Exponential	Yes
Workload control	n.a.	General	Yes
PAC	Poisson	General	Yes

Table 4.2: Arrivals, service times and state dependence of the queueing models

An advantage of Poisson arrivals and exponentially distributed service times is that they have the "Markovian" or "Memoryless" property. The implication of this property is that if the service is ongoing at time t, the remaining service time has the same distribution as the service time itself, regardless of the start of the service time. This property makes it easier to calculate the performance measures than of models with generally distributed arrivals and service times. [25]

Furthermore, the models provide either state dependent or state independent performance measures. The state independent performance measures show the performance of the system in its steady-state (assuming that there is one). As the BFC outbound process starts and ends with an empty system and the number of operators changes between the day and the evening shift, one could argue that it would be better to use a state dependent model. However, if the focus is put on the time in between these moments for which the goal is to maintain a steady processing flow, a state independent model could be used as well.

Based on the comparison of the different characteristics of the models in this section and the fact that the arrival and service time distributions of the BFC outbound process are not Poisson or exponential, it can be concluded that if the preference is given to a state dependent model, the workload control model is the best candidate for further analysis.

One could however argue that it is not necessary to use a workload controlled model since the stingray, which functions as a buffer area in the outbound process, has a much higher capacity than is ever used. On the other hand, the reason that the maximum capacity of the stingray is never reached, is most likely due to the regulated release of workload to the system by the control room. In theory, it could be possible to exceed the capacity of the stingray if the input is no longer regulated. A significant problem of this model is however the combination of its high computational complexity and the enormous state space of the BFC outbound process. This combination results in unreasonable long computation times and makes the model unsuitable for this research project. The same holds for the other state dependent models.

Looking at the state independent models, the best candidates are the complete reduction method and the decomposition method. As described in Section 4.3, most decomposition methods are restricted to systems with single-server stations. An alternative is provided by Rabta et al. [28], which is a hybrid solution that also involves estimation of the inter-departure times through simulation. Since a simulation model is also used for validating the performance of the mathematical model, it is preferred to work with a model that does not require input or output from the simulation model itself. Therefore, the complete reduction method forms the basis for the remainder of this research.

Chapter 5

BFC outbound process as a network of queues

In the previous chapter several queueing network models have been introduced. This chapter focuses on describing the BFC outbound process as a network of queues. The outbound process consists of multiple steps, where each step can be seen as a different processing station. The full process can therefore be represented by a network that consists of a collection of nodes connected by a set of paths. Each node in the network represents a work center that consists of a number of work stations that perform a certain step in the outbound process. In Figure 5.1 this network is depicted.

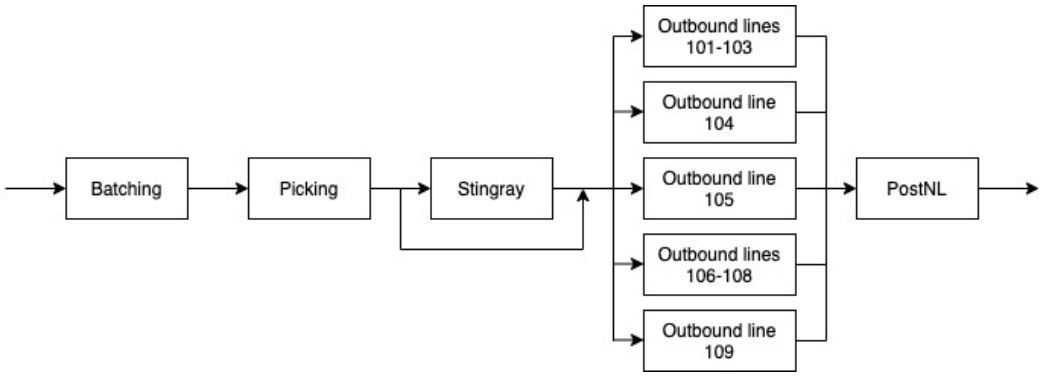


Figure 5.1: Outbound process as a network of services

The directional arrows between the nodes show the sequencing of the services. All items (customers) arrive at the batching process and leave the system at PostNL, which makes this an open network. Not all items go through the same process, therefore they can be assigned to different classes with different arrival times and routes through the network. The classes correspond to the designated outbound lines of the items. Each node can be modelled as a small queueing system and

the outbound process as a whole can be modelled as a multi-class open queueing network with multi-server queues.

As the limiting factor of the process is the number of totes in the system, we look at the network from a tote perspective. This means that one unit in the queueing network equals one tote and that the service times are expressed in the processing time of a tote. More information on an order or item level might be provided as well for a better understanding of the processes at the work centers.

The remainder of this chapter is structured as follows. In sections 5.1 to 5.5, we first zoom in at the different nodes of the network. After that, Section 5.6 describes how the transportation between the nodes should be taken into account. The chapter closes in Section 5.7 by presenting the queueing network as a whole.

5.1 Batching process

The batching process can be seen as the generator of tote arrivals. The input of the batching process is a list of orders with order items that the control rooms wants to be batched at a certain time and the output is a list of order items grouped in pick batches. These pick batches are sent to the virtual queue of the picking process, where they wait until a picking operator becomes available to collect all items of the pick batch in a tote assigned to that pick batch. Therefore, it can be said that the batching process regulates the arrival rate of totes to the system.

5.2 Picking

Once the orders have been batched, they enter the picking process of which the queueing model is depicted in Figure 5.2.

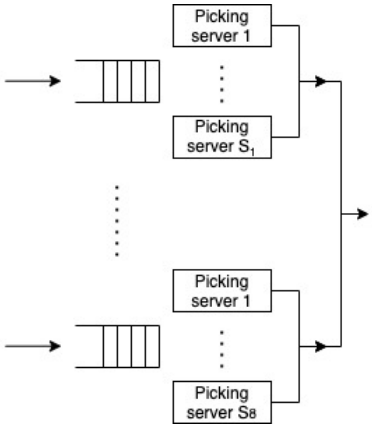


Figure 5.2: Picking process

Order picking takes place in two pick towers that each consist of four different floors. As order pickers generally do not change between pick towers or floors, this is modelled as eight multi-server queues. There are no external arrivals, so the arrival rate is fully dependent on the output of the batching step. There might still be pick batches in the queue from the previous order batching request of the control room once the new pick batch requests arrive. The capacity of the queue can be regarded as infinity, because the pick batches are virtually waiting in a database with more than enough capacity.

The current logic used for sequencing the pick batches is as follows. Each pick batch receives a weight, which is based on the cut-off time and the location of the pick batch. It holds that the lower the cut-off time and the closer the location of the pick batch to the order picker, the higher the weight. The pick batch with the highest weight is selected first. If there are more pick batches with the same weight, the pick batch is selected that has been waiting in the queue the longest.

Recently, research has been done at BFC to improve the sequencing of pick batches. The new logic determines the sequence based on the cut-off times, steady workflow and pool completion of multi-orders [18]. The main idea behind this logic is that the workload is distributed more evenly over the outbound lines and that the pool completion time is reduced. This new logic has not been implemented yet but is scheduled to be implemented in the first half of 2021.

The number of servers equals the number of order pickers that have been scheduled according to the production plan. This number may differ per day but also per shift during the day. The service time starts when an empty tote is taken to collect all items of the pick batch and ends when the full tote is put on the conveyor belt for transport to the stingray or the outbound lines.

5.3 Stingray

After a pick batch has been collected in a tote, this tote is sent either directly to the right outbound line, or to the stingray, where the tote needs to wait until it can be sent to the right outbound line. A tote is sent to the stingray for example if it needs to wait for another tote that contains the items to complete a multi order, such that the totes can be sent to the outbound line at the same time for sorting and packing. In Figure 5.3 the queueing model of the stingray is depicted.

Figure 5.3: Stingray

The arrival rate at the stingray is dependent on the picking process as well as the capacity at the sorting and packing centers. Mono orders do not have to wait

on other totes for order completion and can therefore skip the stingray if there is enough capacity at their designated outbound line. Multi orders on the other hand might have to wait on other totes for order completion. Besides that, totes can be sent to the stingray for temporary storage in case there is not enough capacity at the designated outbound line at that time.

The storage capacity of the stingray equals 4400 totes, which has hardly ever been reached and one could therefore decide to model it as infinite capacity. The queue discipline is as follows. For mono orders holds that as soon as capacity becomes available at an outbound line, a tote is sent to the outbound line to fill up the free capacity according to the priority rules (lowest cut off time, longest waiting time in stingray). For multi orders, this is a bit more complex. A distinction is made between the manual sorting center and the automatic sorter.

The following holds for the manual sorting center. As soon as capacity becomes available, it is checked whether there are still totes remaining in the stingray from a multi order pool of which other totes are already waiting at the center. If this is the case, the next tote of the pool is released. If all totes from the previous multi order pool have been released, the next multi order pool, that may be released according to the priority release rules (lowest cut-off time, longest waiting time in stingray), is selected and the first tote from this pool is released.

The automatic sorter has a north and south induct island. The islands both have a hold point from where totes are released to the buffers of the infeed workstations. Once all totes of a pool have passed the hold point, a new pool is pulled from the stingray to the hold point. The workload between the stingray and the hold points is controlled by the parameter MaxBatchRelease. When this threshold is met, no more pools are sent from the stingray to the holdpoints of the induct islands until the last tote of one of the outstanding pools passes the holdpoint at one of the two induct islands. The parameter thus regulates the workload between the stingray and the hold points and not the amount of pools that are currently on the sorter itself.

Summarized, the stingray has two functions. It serves as an extended buffer of the sorting and packing centers and as a synchronization station for pools of pick batches. The pool completion time could be regarded as the service time of the stingray.

5.4 Outbound lines

In the outbound lines a distinction is made between the mono and multi outbound lines. The mono outbound lines consist of outbound lines 101 to 105. A tote arriving at the queue of a mono outbound line either came from the stingray or directly from the picking area. The arrival rate is therefore dependent on the

output of these two centers. The multi outbound lines consist of outbound lines 106 to 109. All totes arriving at the queue of a multi outbound line came from the stingray. The arrival rate is therefore only dependent on the output of the stingray.

For all mono and multi outbound lines holds that the service time of a tote starts when the tote is taken from the queue and ends when the tote is put empty on the exit loop of the conveyor belt. Besides that, the queue discipline for the totes is first come first served but the items in the tote are randomly chosen by the sorting or packing server. In sections 5.4.1 to 5.4.5 it is briefly described how the different outbound lines can be modelled as queueing networks.

5.4.1 Outbound lines 101-103

Outbound lines 101 to 103 are located at the same spot in the warehouse and are regarded as one work center. Each packing station has one or two servers who are able to pack orders from all three outbound lines. In Figure 5.4 the corresponding queueing system is depicted.

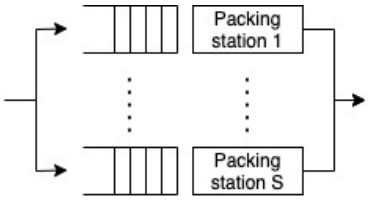


Figure 5.4: Outbound lines 101-103

Each packing station has room for two totes in the queue next to one tote being processed. In total there are 40 stations available but these are not all occupied all day nor every day. The real time capacity is dependent on the production plan and can change between shifts during the day as well.

5.4.2 Outbound line 104

Outbound line 104 consists of one machine, called the smartmailer. The process is as follows. A tote is taken from the waiting queue. All items from the tote are scanned one by one and carefully placed on the right spot on the conveyor belt of the machine. Subsequently, the machine puts the item in an envelope. The queueing system can therefore be modelled as depicted in Figure 5.5.

Figure 5.5: Outbound line 104

The smartmailer has a capacity of thirteen totes in the queue in addition to one tote being processed.

5.4.3 Outbound line 105

Outbound line 105 consists of three stations. The machines used at these stations are called carton wrappers. The process, which is similar for each station, is as follows. A tote is taken from the waiting queue. All items from the tote are scanned one by one and carefully placed on the right spot on the conveyor belt of the carton wrapper. Once an item is scanned, the machine simultaneously cuts the right amount of carton to wrap the item in. The piece of carton and the item encounter each other in another part of the machine, where the carton is then automatically folded around the item. The corresponding queueing system is depicted in Figure 5.6.

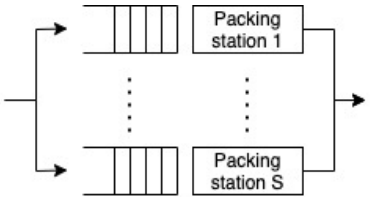


Figure 5.6: Outbound line 105

Each station has a capacity of eight totes in the queue in addition to one tote being processed. However, not all stations are operating all the time.

5.4.4 Outbound lines 106-108

Outbound lines 106-108 are located at the same spot in the warehouse and are considered one work center. There are ten stations in total and each station can process orders from all three outbound lines. At each station there is one sorting server and three packing servers. The sorting server gets a tote from the queue and puts the items in so called pigeon holes, which are storage locations for sorted orders. Once a sorted order is complete, one of the available packing servers retrieves the items from the pigeon hole and starts packing. The corresponding queueing system is depicted in Figure 5.7.

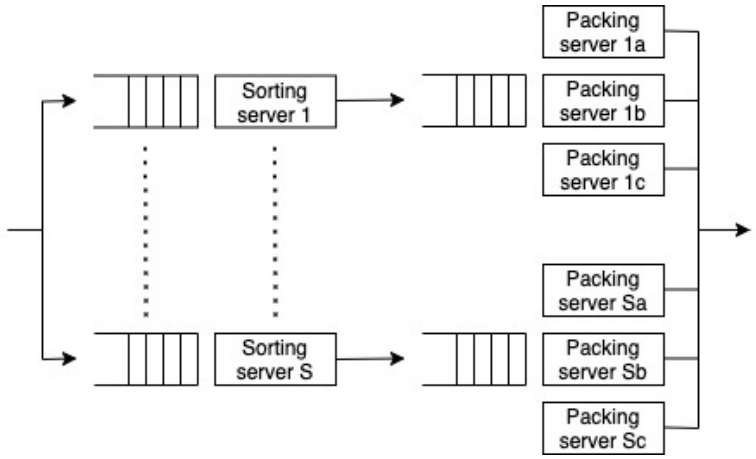


Figure 5.7: Outbound lines 106-108

Each sorting station has a capacity of five totes in the queue in addition to one tote being processed. However, not all stations are operating all the time. The capacity of the queue of each set of packing stations equals 69 orders.

5.4.5 Outbound line 109

Once the totes of a pool are released by the stingray, they arrive at the hold point of one of the induct islands. When an infeed station pushes off an emptied tote, a new tote is pulled from the hold point to the buffer of the infeed workstation. Each infeed workstation has an operator who takes an item from the tote, scans it and puts it on the conveyor belt of the automatic sorter. When an item is scanned, it is assigned to a certain chute to which the automatic sorter releases the item. In the chute the item needs to wait till all items of the multi order are collected, before the packing process is initiated. Packing is done manually by the packing servers. The packing servers receive a sign when one of the chutes in their area contains a complete multi order that is ready to be packed. In Figure 5.8 the corresponding queueing system is depicted.

Each induct island has one hold point with a capacity of nine totes, followed by eight infeed stations which each have a capacity of three totes in the queue in addition to one tote being processed. However, not all infeed stations are operating all the time.

The chutes at the packing stations each have their own green light that is turned on by the automatic sorter once the sorted multi order in the chute is complete. Generally, one packing server is responsible for five chutes. The queue discipline at the packing station is as follows. Once the packing server has finished an order it looks for a green light. If there is no green light, the packing server has to wait until the next multi order is completed in one of the chutes. If there is only one

green light, the order from the chute with the green light is taken. If multiple chutes have received a green light in the time the packing server was working on the previous order, the packing server simply takes the order from the chute with the first green light he or she saw.

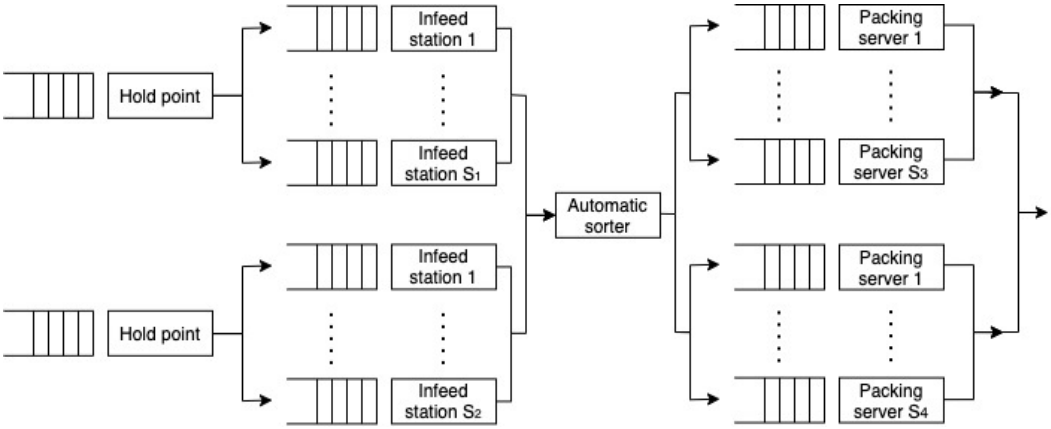


Figure 5.8: Outbound line 109

5.5 PostNL

The conveyor belts of the packing stations of the different outbound lines of bol.com are connected to conveyor belts that transfer the packages to the PostNL distribution center located on the ground floor of BFC. There is a limit on the number of packages (output) that can be sent from BFC to PostNL, such that the capacity of the sorting machine of PostNL is not exceeded. The maximum output expressed in packages per hour for the different outbound lines is given in Table 5.1.

Outbound line(s)	**Maximum output per hour in packages**
101-103	4400
104	1000
105	2200
106-108	4400
109	4400

Table 5.1: Maximum output of the outbound lines in packages per hour

In the analysis, the units of interest are the totes. The maximum output in Table 5.1 is expressed in packages instead of totes. In order to be able to take into account the maximum output, this measure should be converted to a maximum number of totes to be processed per hour at each outbound line. This is done by determining the average number of orders per tote, assuming that each order

equals one package. For mono orders this equals the average number of items per tote since each item equals an order. For multi orders this is a bit more complicated since the items of an order are usually spread over multiple totes. Therefore, for multi orders this measure is estimated by taking the average over all pools of the division of the number of orders per pool and the number of totes used per pool. The maximum output of the outbound lines can then be expressed in totes per hour by dividing the maximum output per hour in packages by the average number of orders per tote.

5.6 Transportation time

In the previous sections the individual stations are discussed. However, the transportation of totes between work centers has not been included yet. An overview of the flow through the warehouse is provided in Appendix A. Transportation takes place in the following ways:

1. From the picking area to the stingray
2. From the picking area to the sorting and packing centers (through a bypass at the stingray)
3. From the stingray to the sorting and packing centers

Under optimal circumstances, the conveyor belts are able to process 1473 totes per hour. Generally, this capacity is enough for the transport of totes from the picking area to the stingray and subsequently the sorting and packing stations. Sometimes, problems arise at the end of the day when large amounts of empty totes need to be returned, which then potentially block the way on the conveyor belts between the stingray and the sorting and packing stations. In the initial analysis this blocking problem is left out of scope.

The transportation through the system can be modelled as two stations. The first station is for the transportation between the picking area and the stingray, where the expected transportation time is independent of the customer class. The second station is for the transportation between the stingray and the sorting and packing centers, where the expected transportation time is dependent on the designated outbound line and thus the customer class.

5.7 Complete queueing network

Combining all the queueing systems of the individual nodes results in the queueing network depicted in Figure 5.9.

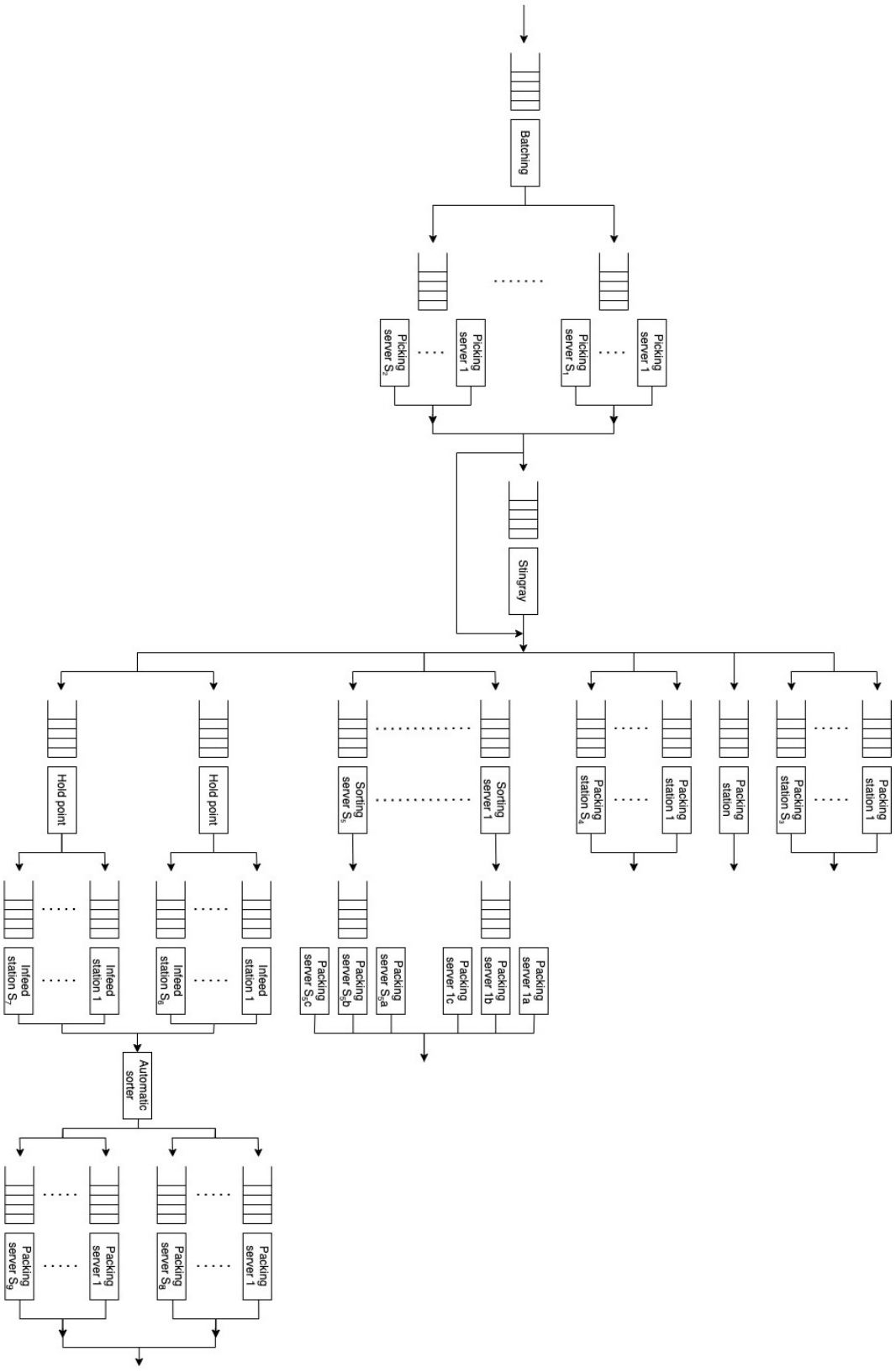


Figure 5.9: Queueing network outbound process BFC

Chapter 6

Mathematical model

In the previous chapter, the BFC outbound process is described as a network of queues. Chapter 4 describes a number of queueing network models that could be used to analyse the queueing network of the BFC outbound process and concludes in Section 4.8 that the complete reduction model is the best candidate for further analysis. In order to be able to analyse the queueing network of the BFC outbound process, including the unconventional characteristics of the stingray, some adjustments and assumptions must be made. In Section 6.1 these assumptions and model adjustments are presented. Section 6.2 describes the complete reduction algorithm as introduced by Zijm [41]. The chapter closes in Section 6.3 with a description of how the stingray logic is incorporated.

6.1 Assumptions and adjustments

In order to be able to analyse the queueing network of the BFC outbound process, several assumptions and adjustments are made.

First of all, there is no data available on the exact number of picking operators per picking area. Besides that, pick batches are released for a designated outbound line but not to a specific picking area as this is determined at a later stage by the micro service Pacman (see Section 2.3.3). Therefore, it is assumed that picking operators as well as pick batches are distributed evenly over the eight picking areas. In order to achieve this, the pick batches are created round-robin for the different picking areas. Additionally, it is assumed that picking operators are identical, do not get tired, and do not move between picking areas. In practice it may happen that a picking operator moves from one area to another, but in principle they remain in one area. Besides that, it occurs in practice that some picking areas receive more pick batches than others with an imbalance in workload as a result. However, some initiatives are set up to ensure a more even distribution of workload over the

picking areas in order to prevent these imbalances.

In addition, the waiting capacity of the picking areas is assumed to be infinite as the pick batches are waiting in a virtual queue. Besides that, the company is currently working on the implementation of the new sequencing logic for pick batches, which is thoroughly described in [18]. Therefore, in this research project the new logic is used instead of the current logic.

Furthermore, transportation times are assumed to be deterministic but different for each outbound line. In practice it may occur that a tote makes an additional round on the packing loop (see the mechanical overview of BFC in Appendix A). This is however prevented as much as possible by the strict release rules of the stingray and is regarded a rare occasion. Besides that, other errors may occur which could result in a standstill of the conveyor belts. This is considered a rare event as well but still makes this a strong assumption. In order to prevent the results to be affected by these events, as those are not of further interest in this initial analysis, the assumption is made.

Regarding the sorting and packing centers, it is assumed that operators are identical for similar work stations, do not get tired, and do not switch between work centers and stations. Besides that, it is assumed that the multi packing stations have enough capacity to process whatever is processed by the sorting stations. In practice this is also how the company makes the personnel planning. The number of operators at the multi packing stations is aligned with the number of operators at the sorting stations.

Finally, it is desired in the company that one hour of workload is present in the system before the sorting and packing centers are activated. The reason for this is to prevent idle times and this is achieved by starting the picking shift one hour earlier than the sorting and packing shift. At the end of the day, the picking operators leave one hour earlier than the sorting and packing operators. The system is emptied at the end of the day with a so called “sweep run”.

6.2 Complete reduction algorithm

The complete reduction algorithm for multi-class open queueing networks as introduced by Zijm [41] is as follows. Consider a multi-class open queueing network with M stations, R classes, general individual inter-arrival and service time distributions characterized by $\lambda_{0j}^{(r)}$ with squared coefficient of variation (SCV) $(C_{0j}^{(r)})^2$, and $\mathbb{E}S_j^{(r)}$ with SCV $(C_{sj}^{(r)})^2$ respectively, and routing matrices $P_{ij}^{(r)}$, with $i, j = 1, ..., M$ and $r = 1, ..., R$ for all parameters.

The complete reduction method consists of three steps:

1. Reducing the R-class open queueing network to a single-class open queueing network by aggregating the classes.
2. Analyzing the single-class open queueing network.
3. Disaggregating to obtain the performance measures per class for the given R-class open queueing network.

Step 1: Reduction
The aggregate first and second moment of the service time at station j, denoted by ES_j and $E(S_j)^2$ respectively, are given by the weighted average of the service times of the individual classes:

$$\mathbb{E}S_j = \frac{1}{\lambda_j} \sum_{r=1}^{R} \lambda_j^{(r)} \mathbb{E}S_j^{(r)}, \qquad j = 1, ..., M, \tag{6.1}$$

$$\mathbb{E}(S_j)^2 = \frac{1}{\lambda_j} \sum_{r=1}^{R} \lambda_j^{(r)} \mathbb{E}(S_j^{(r)})^2, \qquad j = 1, ..., M, \tag{6.2}$$

where $\lambda_j^{(r)}$ is the arrival rate of class r customers at station j:

$$\lambda_j^{(r)} = \lambda_{0j}^{(r)} + \sum_{i=1}^{M} \lambda_i^{(r)} P_{ij}^{(r)}, \qquad r = 1, ..., R, \tag{6.3}$$

and $\lambda_j = \sum_{r=1}^{R} \lambda_j^{(r)}$ is the aggregate arrival rate of customers at station j.

The aggregate SCV of the service time can then be determined with:

$$C_{sj}^2 = \frac{1}{\lambda_j (\mathbb{E}S_j)^2} \sum_{r=1}^{R} \lambda_j^{(r)} \left(\mathbb{E}S_j^{(r)} \right)^2 \left((C_{sj}^{(r)})^2 + 1 \right) - 1, \qquad j = 1, .., M. \tag{6.4}$$

The aggregate arrival rate of customers outside the network to station j is given by $\lambda_{0j} = \sum_{r=1}^{R} \lambda_{0j}^{(r)}$. The proportion of the arrival flow of station j originating from station i is given by:

$$Q_{ij} = \frac{\lambda_{ij}}{\lambda_j}, \qquad i = 0, .., M, j = 1, .., M. \tag{6.5}$$

The aggregate routing matrix, which contains the routing probabilities of the aggregate flow, is given by:

$$P_{ij} = \frac{1}{\lambda_i} \sum_{r=1}^{R} \lambda_i^{(r)} P_{ij}^{(r)}, \qquad i, j = 1, ..., M. \tag{6.6}$$

The aggregate SCV of the arrival process can then be approximated by the following set of linear equations:

$$C_{aj}^2 = a_j + \sum_{i=1}^{M} C_{ai}^2 b_{ij}, \qquad j = 1, ..., M, \tag{6.7}$$

where a_j and b_{ij} are constants depending on the input data:

$$a_j = 1 + w_j \left[(Q_{0j} C_{0j}^2 - 1) + \sum_{i=1}^{M} Q_{ij}[(1 - P_{ij}) + P_{ij}\rho_i^2 x_i] \right], \tag{6.8}$$

$$b_{ij} = w_j P_{ij} Q_{ij} (1 - \rho_i^2), \tag{6.9}$$

for which the values of the parameters ρ_i, v_j, w_j, and x_i are given by:

$$\rho_i = \frac{\lambda_i \mathbb{E}S_i}{c_i}, \tag{6.10}$$

$$v_j = \left[\sum_{i=0}^{M} Q_{ij}^2 \right]^{-1}, \tag{6.11}$$

$$w_j = [1 + 4(1 - \rho_j)^2 (v_j - 1)]^{-1}, \tag{6.12}$$

$$x_i = 1 + c_i^{-0.5} (\max[C_{si}^2, 0.2] - 1), \tag{6.13}$$

and c_i denotes the number of servers at station i.

Step 2: Analysis
Let $\gamma = \sum_{i=1}^{M} \lambda_{0i}$ and $V_i = \frac{\lambda_i}{\gamma}$, the total arrival rate from outside the network and the visit ratios of the stations respectively. The expected waiting time in the queue at workstation i is calculated by:

$$\mathbb{E}W_{Qi} = \frac{C_{ai}^2 + C_{si}^2}{2} \frac{\rho_i^{(\sqrt{2(c_i+1)}-1)}}{c_i(1 - \rho_i)} \mathbb{E}S_i, \qquad \rho_i < 1. \tag{6.14}$$

The total expected time at workstation i then equals:

$$\mathbb{E}W_i = \mathbb{E}W_{Qi} + \mathbb{E}S_i, \tag{6.15}$$

and the overall expected time in the system is determined by:

$$\mathbb{E}W = \sum_{i=1}^{M} V_i \mathbb{E}W_i. \tag{6.16}$$

Given Little's formula $\mathbb{E}L_i = \lambda_i \mathbb{E}W_i$, which provides the number of customers at workstation i, and $\mathbb{E}L = \sum_{i=1}^{M} \mathbb{E}L_i$, the number of customers in the entire system is given by:

$$\mathbb{E}L = \sum_{i=1}^{M} \lambda_i \mathbb{E}W_i = \sum_{i=1}^{M} \gamma V_i \mathbb{E}W_i = \gamma \mathbb{E}W. \tag{6.17}$$

Step 3: Disaggregation
The mean number of customers in the queue of station j is given by:

$$\mathbb{E}L_{Qj} = \mathbb{E}L_j - c_j\rho_j. \tag{6.18}$$

The time spent in the queue of station j is equal for each class, such that:

$$\mathbb{E}L_{Qj}^{(r)} = \frac{\lambda_j^{(r)}}{\lambda_j}\mathbb{E}L_{Qj}. \tag{6.19}$$

The mean number of class r customers in service at station j is given by $c_j\rho_j^{(r)} = \lambda_j^{(r)}\mathbb{E}S_j^{(r)}$, and therefore:

$$\mathbb{E}L_j^{(r)} = \mathbb{E}L_{Qj}^{(r)} + \rho_j^{(r)} = \frac{\lambda_j^{(r)}}{\lambda_j}\mathbb{E}L_{Qj} + \lambda_j^{(r)}\mathbb{E}S_j^{(r)}. \tag{6.20}$$

However, if the service times are equal for all classes at the station, then a simpler expression can be used:

$$\mathbb{E}L_j^{(r)} = \frac{\lambda_j^{(r)}}{\lambda_j}\mathbb{E}L_j. \tag{6.21}$$

The total number of customers in the entire system of a certain class is simply found by adding up the number of customers of that class per station:

$$\mathbb{E}L^{(r)} = \sum_{r=1}^{R}\mathbb{E}L_j^{(r)}. \tag{6.22}$$

The expected time in the system for customers of class r can be found by:

$$\mathbb{E}W^{(r)} = \sum_{j=1}^{M}V_j^{(r)}\mathbb{E}W_j^{(r)} = \sum_{j=1}^{M}V_j^{(r)}\frac{\mathbb{E}L_j^{(r)}}{\lambda_j^{(r)}}, \tag{6.23}$$

where $V_j^{(r)} = \lambda_j^{(r)}/\gamma^{(r)}$ and $\gamma^{(r)} = \sum_j \lambda_{0j}^{(r)}$.

6.3 Incorporating the stingray logic

In the complete reduction method, the first and second moment of the service time of a station are used in the calculations of the performance measures. As described in Section 5.3, the stingray serves as a queue for totes of all outbound lines as well as a checkpoint where totes from multi outbound lines need to wait for pool completion. The latter functionality can be seen as the service of the stingray, which can be modelled in several ways. The most conventional option is to fit a distribution to historical data and use this as the service time.

An alternative is to model the pool completion process in a similar fashion as batch formation is modelled in queueing theory, where a batch equals a complete

pool of totes. The arrival rate of a complete pool λ_P is dependent on the arrival rate of the totes λ_T of that pool. A pool of size N is ready when the N^{th} tote arrives, which means that if all pools are of equal size a new pool arrives at every N^{th} tote. The mean inter-arrival time equals [41]:

$$E(A_P) = \frac{1}{\lambda_P}, \tag{6.24}$$

where the arrival rate of pools is given by:

$$\lambda_P = \frac{\lambda_T}{N}. \tag{6.25}$$

The pool completion time is dependent on the inter-arrival time of the totes in the pool. The N^{th} tote does not have to wait before the pool is complete. The $N-1^{th}$ tote has to wait until the N^{th} tote arrives, which takes on average $\frac{1}{\lambda_T}$. The arrivals of totes are independent and identically distributed. Given A_j, the inter-arrival time between tote $i-1$ and tote i, the SCV of the pool arrival process can be determined by [41]:

$$C_P^2 = \lambda_P^2 \text{Var}\left(\sum_{i=1}^{N} A_i\right) \tag{6.26}$$

$$= \frac{\lambda_T^2}{N^2} N \ \text{Var} A_1 \tag{6.27}$$

$$= \frac{C_T^2}{N}, \tag{6.28}$$

where C_T^2 is the SCV of the arrivals of totes.

Since the pool size is not a constant, some adjustments must be made. The expected inter-arrival time of pools is given by:

$$E(A_P) = \frac{E(N)}{\lambda_T}. \tag{6.29}$$

The variance of the inter-arrival times can be determined as follows:

$$\text{Var}(A_P) = \text{Var}\left(\sum_{i=1}^{N} A_i\right). \tag{6.30}$$

The SCV of the inter-arrival time of pools is then given by:

$$C_P^2 = \frac{\lambda_T^2}{E(N)^2} \text{Var}(A_P) \tag{6.31}$$

where, since the inter-arrival times of totes A_i are independent and identically distributed, and N and A_i are independent:

$$\text{Var}(A_P) = E(N)\text{Var}(A_1) + E(A_1)^2 \text{Var}(N). \tag{6.32}$$

With this method the arrival rate at the stingray is converted from tote arrivals to pool arrivals. The service time of the stingray then equals zero. At the sorting and packing centers, totes are processed one by one. To make these centers compatible with the stingray, the arrival rate at these centers must be converted back to totes or the service times need to be converted to pool service times. The first option is a better resemblance of reality and therefore used in this research project.

Chapter 7

Approach and methods

The previous chapter described how the queueing network of the BFC outbound process can be mathematically analysed. This chapter describes the approach and methods used to find a solution to the core problem and to answer the main research question. Section 7.1 describes the solution approach and is followed by the experimental set-up in Section 7.2.

7.1 Solution approach

The main research question is: *"How should the workload at the different work stations in the warehouse be allocated such that the overall throughput is maximized and the operating costs are minimized, while maintaining the order fulfillment score?"*

The workload in the BFC outbound process is controlled by the release of pick batches per outbound line, which from a queueing theory perspective corresponds to the arrival rates. These are thus the variables to be optimized.

The key performance indicators in this research, based on the main research question, are the throughput, operating costs and order fulfillment score. The throughput can be calculated with the output of the mathematical model introduced in Chapter 6 and the simulation model.

The order fulfillment score is obtained by converting the throughput expressed in number of totes to the throughput expressed in the number of orders and dividing this by the targeted number of orders on the production plan. The conversion of the throughput is done in a similar fashion as the conversion of the maximum output per hour in packages (orders) to the maximum output per hour in totes at PostNL in Section 5.5. The production plan is a given. This means that the throughput is the only variable for this measure, which must be maximized.

As stated in Section 3.2 translating the objective to minimize operating costs to queuing theory performance measures means that the station/operator idle time should be minimized. This is achieved by maximizing the utilization rate ρ, which equals the fraction of time that the station/operator is working. However, to avoid that the queue eventually grows to infinity, it is required that $\rho < 1$.

Essentially, this means that the goal is to optimize the arrival rate of the pick batches per outbound line such that the throughput and utilization rate are maximized. This can be achieved by creating a steady state for which holds that:

1. The arrival rate of pick batches at each outbound line can not exceed the capacity of the outbound line.
2. The aggregate arrival rate of pick batches can not exceed the capacity of the bottleneck work center.
3. The expected number of pick batches (totes) in the system can not exceed the capacity of the system.

Working backwards through the queueing network it can be determined what the bottleneck work center is and therefore what the maximum aggregate arrival rate of pick batches is. The maximum arrival rate of a station is determined as follows. The utilization rate of a station is determined by $\rho = \frac{\lambda E(S)}{c}$ and in order to keep the system stable $\rho < 1$. Given the number of servers c, expected service time $E(S)$, and utilization rate ρ, the required arrival rate λ can be determined. The maximum arrival rate of a station is then determined by setting $\rho = 0.99$, and by multiplying the outcome by the number of stations in the work center, the maximum arrival rate of the work center is obtained.

If the bottleneck is not the sorting and packing center, a linear program (LP) can be formulated in order to distribute the maximum aggregate arrival rate over the different outbound lines according to some objective function. Then given the arrival rates, the expected number of pick batches (totes) in the system and corresponding throughput can be determined with the mathematical model described in Chapter 6. If it turns out that the expected number of pick batches in the system is higher than the maximum capacity, the previous steps are repeated with a lower utilization rate. This is repeated until the right arrival rates have been found that create a steady state in which the capacity of the system is not exceeded.

A very important assumption that needs to be made for this approach to work, is that there are always enough pick batches available to be released to the system. Generally, only must-go orders, which are orders that need to be shipped today, are released to the system and in that case it could happen that the order basket of an outbound line becomes empty. However, if switching orders between outbound lines is allowed and could-go orders, which are orders that could be shipped at

a later time, may be released as well, the assumption is valid. Adjusting the designated outbound line of orders is already done by the control room if the number of orders for a certain outbound line is getting low whereas there are plenty of orders for other outbound lines. Releasing could-go orders is also done already but not on a large scale yet.

The solution approach can be summarized by the flowchart depicted in Figure 7.1.

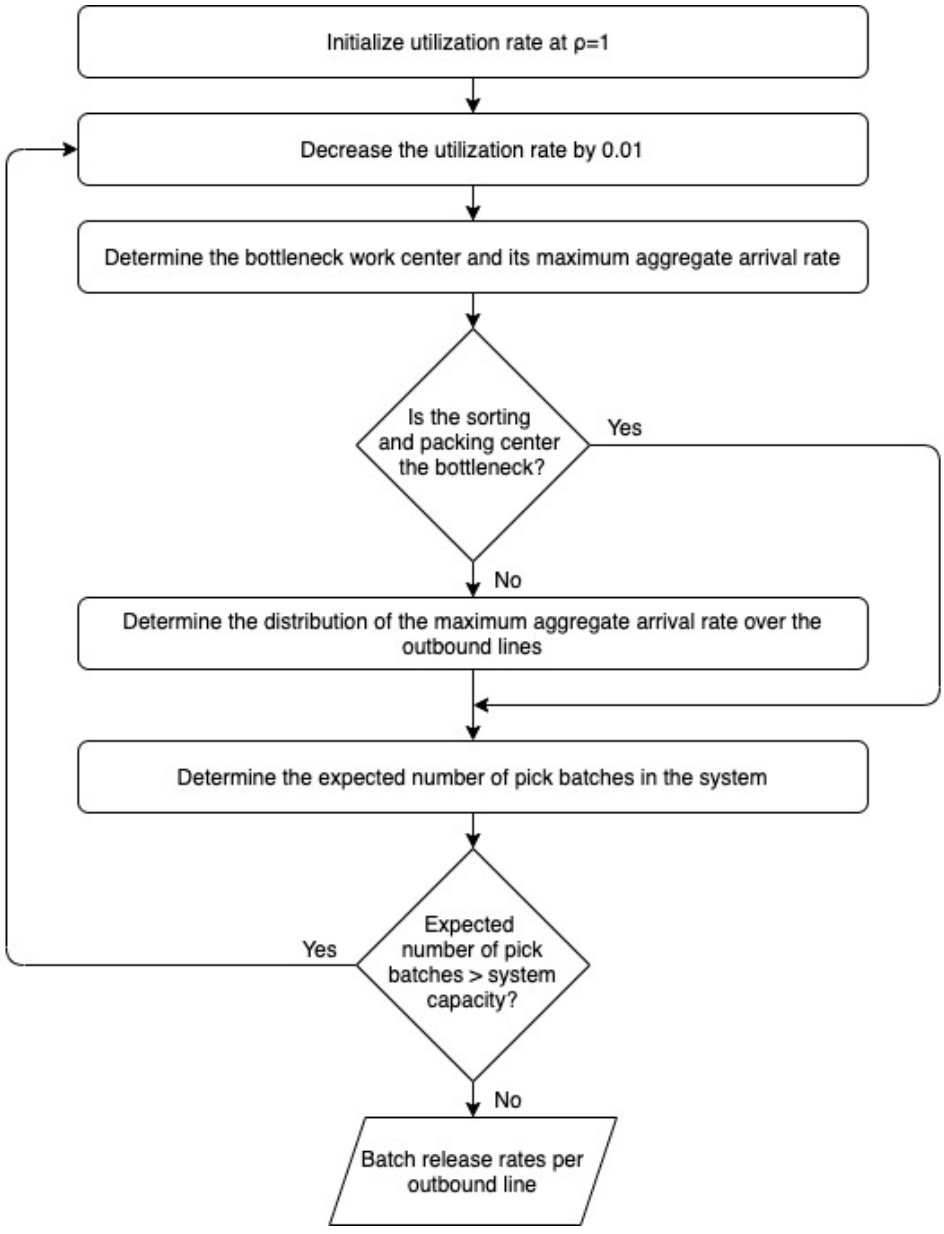


Figure 7.1: Flowchart of the solution approach

The outcome of this approach is the arrival rate, better described as the release rate, of pick batches for each designated outbound line. This holds on the assumption that the control room makes sure that there are always enough pick batches

available for every outbound line by changing the designated outbound line of orders and enabling the release of orders that could be shipped at a later time. However, this process could also be automated by using certain threshold rules or formulating an LP to make the decision in certain time intervals according to some objective function.

As the model introduced in Chapter 6 is based on several assumptions and thus results in a simplified version of the BFC outbound process, a simulation model is made as well. This simulation model is used to validate the mathematical model on both a station and system level. The simulation model itself needs to be validated as well. This is done with historical data from the company.

7.2 Experimental set-up

The experiments in this research project consist of three parts. First of all, experiments are done in order to determine how well the mathematical model introduced in Chapter 6 and the simulation model capture the BFC outbound process. Second, experiments are done to determine whether the proposed approach of determining the release rate of pick batches performs better than the current approach. At last, experiments are done to determine in what manner pick batches should be released to the system.

As explained in the previous section, an LP can be formulated to distribute the maximum aggregate arrival rate over the outbound lines in case the bottleneck is not the sorting and packing center. This LP is presented in Section 7.2.1. Besides that, an LP is formulated for changing the designated outbound lines of orders and adding could-go orders to the order basket, which is presented in Section 7.2.2. Finally, a more extensive explanation of the experiments is given in Section 7.2.3.

7.2.1 Distributing the maximum aggregate arrival rate over outbound lines

In order to determine how the maximum aggregate arrival rate should be distributed over the different outbound lines, an LP is formulated. The following notation is used:

Parameters

d	designated outbound line(s)
λ_d^{max}	maximum arrival rate of pick batches at outbound line(s) d
λ^{max}	maximum aggregate arrival rate of pick batches
W_d	weight assigned to outbound line(s) d

Decision variables

λ_d rate of pick batches sent to outbound line(s) d

The LP model is formulated as follows:

minimize

$$\sum_{d=0}^{D} W_d * (\lambda_d^{max} - \lambda_d) \tag{7.1}$$

subject to

$$\lambda_d \leq \lambda_d^{max} \tag{7.2}$$

$$\sum_{d=0}^{D} \lambda_d \leq \lambda^{max} \tag{7.3}$$

$$\lambda_d \geq 0 \tag{7.4}$$

The objective function, Equation 7.1, makes sure that the difference between the maximum arrival rate and the actual rate of pick batches of the different outbound lines is minimized. In addition, the objective function contains weights per outbound line, which could for example be based on the preference for or production target of the different outbound lines. Based on preference, one could for example favor the mechanical outbound lines over the manual outbound lines as these have a higher processing rate. Similarly, one could favor outbound lines with higher production targets. In the end it holds that the higher the weight, the smaller the difference between the maximum arrival rate and the actual rate of the pick batches for that outbound line.

The first constraint, Equation 7.2, makes sure that the rate of pick batches sent to each outbound line does not exceed the maximum arrival rate at that outbound line. The second constraint, Equation 7.3, makes sure that the sum of the rate of pick batches sent to the outbound lines does not exceed the maximum aggregate arrival rate of pick batches. The last constraint, Equation 7.4, makes sure that the rate of pick batches sent to each outbound line is non-negative.

If the maximum aggregate arrival rate of pick batches equals the sum of the maximum arrival rate of pick batches at the outbound lines, the weights in the objective function have no impact and the result will always be that the rate of pick batches sent to the outbound lines equal the maximum arrival rate of pick batches at the outbound lines. Right now, this also holds for the BFC, since the bottleneck in the outbound process is the sorting and packing stations.

Future expansions of the sorting and packing stations could however result in a different bottleneck station. Besides that, the logic depicted in Figure 7.1 could also be applied to other warehouses, where the bottleneck station might be one of the other stations in the outbound process. In these cases, it could be interesting

to investigate the impact of different weights in the objective function on the performance measures of the models.

7.2.2 Balancing order basket levels

Initially, orders are assigned to a certain outbound line. Some orders can be sent to multiple outbound lines and therefore their designated outbound line may be changed later in time. Currently, changing the designated outbound line of an order is done before the pick batches are created. Theoretically, one could also change the designated outbound line of a mono pick batch or a pool of multi pick batches at an even later stage, for example in the stingray. In this research project, the focus is put on changing the designated outbound line of orders before the creation of pick batches and it is assumed that no more changes are made after that. An overview of which orders can be sent to which outbound line is given in Table 7.1.

	Outbound 101-103	Outbound 104	Outbound 105	Outbound 106-108	Outbound 109
Orders 101-103	yes	no	no	yes	no
Orders 104	yes	yes	no	yes	no
Orders 105	yes	no	yes	yes	yes
Orders 106-108	no	no	no	yes	no
Order 109	no	no	no	no	yes

Table 7.1: Available order baskets for outbound lines

There are two reasons why changing the designated outbound line of orders could be required:

1. The order basket of an outbound line is getting empty.
2. The order basket of an outbound line is too full and can not be emptied before the end of the day.

In the first case, it should be determined where the additional orders should come from. These could come from order baskets of other outbound lines or from could-go orders. In the second case, it should be determined to which other order basket(s) the excess orders should be sent to. Since the orders are assigned to a specific outbound line not without reason, changing the outbound line of orders should be limited as much as possible. The decision of changing orders between outbound lines and where these orders should come from or should be sent to, can be formulated as an LP.

The following notation is used:

Parameters

d	designated outbound line(s)
O_d	order basket of outbound line(s) d
L_d	lower bound of the order basket of outbound line(s) d
U_d	upper bound of the order basket of outbound line(s) d
$Y_{d,p}$	Boolean: 1 if orders may be sent from order basket d to order basket p, and 0 otherwise
E_d	the number of excess orders in the order basket of outbound line(s) d
S_d	the number of orders short in the order basket of outbound line(s) d

Decision variables

$X_{d,p}$	the number of orders sent from order basket d to order basket p
Z_d	the number of could-go orders added to the order basket of outbound line(s) d

The problem can then be formulated as follows:

minimize

$$\sum_{d=0}^{D} \left(E_d - \sum_{p=0}^{P} X_{d,p} + Z_d \right) \tag{7.5}$$

subject to

$$O_d + \sum_{p=0}^{P} X_{p,d} + Z_d \leq max(O_d, U_d) \tag{7.6}$$

$$O_d - \sum_{p=0}^{P} X_{d,p} \geq min(O_d, L_d) \tag{7.7}$$

$$\sum_{p=0}^{P} X_{d,p} \leq E_d \tag{7.8}$$

$$\sum_{p=0}^{P} X_{p,d} + Z_d \geq S_d \tag{7.9}$$

$$X_{d,p} \leq X_{d,p} Y_{d,p} \tag{7.10}$$

$$X_{d,p} \geq 0 \tag{7.11}$$

$$Z_d \geq 0 \tag{7.12}$$

The objective function, Equation 7.5, tries to minimize the remaining number of excess orders and the number of could-go orders added to the order baskets of outbound line(s) with shortages. The number of excess orders equals the number of orders on top of the upper bound of the order basket. The number of shortages

equals the number of orders needed to reach the lower bound of the order basket. The objective function minimizes the could-go orders, because the preference is to fill up shortages in the orders baskets with must-go orders from other outbound lines before adding could-go orders. One could also add weights to the objective function in order to favor adding could-go orders to the order basket of one outbound line over another for example.

The first constraint, Equation 7.6, makes sure that orders are only added to the order basket if there is room. This means that orders can only be added if the current level of the order basket is lower than the upper bound of the order basket. The second constraint, Equation 7.7, makes sure that orders can only be removed from the order basket and thus added to another order basket if the current level of the order basket is higher than the lower bound of the order basket.

The third constraint, Equation 7.8, makes sure that the number of orders that is sent from outbound line d to other outbound lines does not exceed the excess amount of orders in the order basket of outbound line(s) d. The fourth constraint, Equation 7.9, makes sure that there are no shortages.

The fifth constraint, Equation 7.10, makes sure that orders are only sent from outbound line(s) d to outbound line(s) p if this switch is allowed by the Boolean $Y_{d,p}$. The final constraints, Equation 7.11 and Equation 7.12 , make sure that the number of orders switched between outbound lines and the number of could-go orders added to the order basket are non negative.

The outcome of the LP is the number of orders that should be sent from the order basket of one outbound line to the order basket of another outbound line and how many could-go orders should be added to the different order baskets. This is done to balance the order basket levels, such that the order baskets do not get too empty and can therefore not release the required amount of orders to keep the packing center of the designated outbound line busy. In that way, no processing capacity is lost. Besides that, it minimizes the amount of excess orders such that the order baskets do not get too full and can therefore not release all orders before the end of the day. The purpose of this is to reduce the loss in order fulfillment.

This problem should be solved whenever the order basket level of an outbound line falls below its lower bound. The lower and upper bound levels should be set by the company. An example for a time related upper bound is the number of orders that could still be fulfilled by the outbound line before the end of the day. In this calculation one should also take into account the expected number of orders that come in during the remainder of the day. An example for the lower bound is the number of orders that are processed within a certain time interval plus a safety margin. These are also the upper and lower bounds used in the experiments in this research project.

7.2.3 Experiments

As mentioned before, experiments are done with three different purposes:

1. Validation of the mathematical model and the simulation model
2. Comparison of the current and proposed approach of batch releases
3. Impact of different time intervals for batch releases

In these experiments, we make use of the paired t-test in order to determine the statistical significance. The paired t-test is an analysis on two populations of which the observations are collected in pairs. Each pair of observations is taken under homogeneous conditions but there may be differences in the conditions between the pairs. The test is done on the differences between each pair of observations. The underlying assumption of the paired t-test is that the differences are normally distributed, which is a reasonable assumption in many cases. However, if the underlying distribution is not normal, a moderate departure from normality will have little effect on the validity of the t-test. [24]

The following notation is used in the t-test:

μ_D	mean of the differences in observations
Δ_0	hypothesized value
H_0	null hypothesis
H_1	alternative hypothesis
T_0	test statistic
$\overline{D}$	sample average of the differences in observations
S_D	sample standard deviation of the differences in observations
n	number of paired observations
$t_{\alpha,n-1}$	t-value from the T-distribution
α	level of significance

The t-test can be performed one-sided or two-sided, depending on the conclusion to be drawn. If the objective is to make a claim that the differences between the pairs of observations are greater than or less than Δ_0, a one-sided test is appropriate. If the objective is to show that the differences between the pairs of observations are equal or unequal to Δ_0, regardless of the direction, a two-sided test should be used. [24]

The one-sided t-test consists of the following steps [24]:

1. Parameter of interest: μ_D
2. Null hypothesis: H_0: $\mu_D = \Delta_0$
3. Alternative hypothesis (select one):

   (a) H_1: $\mu_D > \Delta_0$
   (b) H_1: $\mu_D < \Delta_0$

4. Test statistic:
$$T_0 = \frac{\overline{D} - \Delta_0}{S_D/\sqrt{n}}$$

5. Reject null hypothesis if:

   (a) $T_0 > t_{\alpha,n-1}$ (if H_1: $\mu_D > \Delta_0$)
   (b) $T_0 < -t_{\alpha,n-1}$ (if H_1: $\mu_D < \Delta_0$)

The two-sided t-test is done as follows [24]:

1. Parameter of interest: μ_D
2. Null hypothesis: H_0: $\mu_D = \Delta_0$
3. Alternative hypothesis: H_1: $\mu_D \neq \Delta_0$
4. Test statistic:
$$T_0 = \frac{\overline{D} - \Delta_0}{S_D/\sqrt{n}} \tag{7.13}$$
5. Reject null hypothesis if: $T_0 > t_{\alpha/2,n-1}$ or $T_0 < -t_{\alpha/2,n-1}$

Below, it is explained for each of the three types of experiments how these are performed and which t-test is applied.

Validation of models

According to Law [22], validation is defined as the process of determining whether a (simulation) model is an accurate representation of the system, for the particular objectives of the study. The ultimate test of a model's validity is to proof that its output data closely resemble the output data that would have been expected from the actual system. This is done by developing a model of the actual system and comparing its output data with the historical data from the actual system.

Above described procedure is performed for validating the simulation model. The current pick batch release approach is incorporated by releasing exactly the same amount of pick batches at exactly the same time to the simulation model as was done by the control room during that day. The simulation model is then validated by simulating days in the past and comparing its performance to the historical data of those days. The comparison is done by performing a paired t-test on the average throughput, sojourn time, and WIP, with twelve days in the peak

period and twelve days outside the peak period. In this case, a two-sided test is chosen, because the objective is to determine whether the output of the simulation model is close enough to the historical data. Therefore, the null hypothesis is that the mean of the differences in observations is equal to zero and the alternative hypothesis is that the mean of the differences in observations is unequal to zero.

The mathematical model introduced in Chapter 6, is designed to determine the performance measures of a system in its steady state. As in the current approach, the arrival rates change continuously such that the steady state of the system is never reached, the mathematical model can not be used to determine the performance measures. As a consequence, the mathematical model can not be validated with above described procedure.

Among others, Law [22] suggests using another model developed for the same system and for a similar purpose, that has already been validated, in this case. An informal comparison can then be made between the two models to validate the output of the new model. In this research project, this approach is used by validating the mathematical model based on a comparison of its output to the output of the simulation model that incorporates the proposed approach. Again, a comparison is done by performing a two-sided t-test on the average throughput, sojourn time, and WIP, with the same twelve days in the peak period and outside the peak period. The two-sided test is chosen, because the objective is to determine whether the output of the mathematical model is close enough to the output of the simulation model in which the proposed approach has already been incorporated. It should be noted that this approach is only legitimate if the simulation model is considered an accurate enough representation of reality.

Comparison of approaches

Once the mathematical model and simulation model have been validated, these models can be used to compare the current and proposed approach of batch releases. This is done by comparing the output of the simulation model with the current approach to the output of the simulation model with the proposed approach and the mathematical model which only works with the proposed approach. These two comparisons are done by performing a paired t-test again on the average throughput, sojourn time, and WIP, with the same twelve days in the peak period and outside the peak period, as were used in the validation process. In this case however, one-sided tests are used, because the objective is to show that the proposed approach performs better than the current approach. The definition of better here, is a higher average throughput, lower average sojourn time and lower average WIP.

Besides a comparison of the average performance of the two approaches, several analyses are also done on the performance of the two batch release approaches over time. The simulation model is used to obtain the data for the two approaches.

This data includes the number of items released to the system, WIP levels, and number of items processed during the day.

Impact of time intervals

Currently, the release of pick batches takes place in time intervals. These time intervals are not of equal size but as explained in Section 2.3, the control room decides how many pick batches are released into the system and at what time. The proposed approach on the other hand, determines the required arrival rate of pick batches per outbound line at the beginning of each work shift and maintains this rate throughout the entire shift. However, it still must be decided which time interval should be used for these rates. Therefore, experiments are performed with different time intervals.

Instead of releasing pick batches in time intervals, one could also make this a continuous process by introducing production authorization cards (PACs). The question remains however, how many PACs are required in that case. As explained in Section 4.8, the state space is too large to use the conventional algorithms for calculating the required number of PACs. If the mathematical model introduced in Chapter 6 is regarded a good representation of practice, one could, as an alternative, experiment with setting the number of PACs equal to the expected number of pick batches in the system's steady state as determined by the model.

Chapter 8

Results

The previous chapter described the solution approach and experimental set-up. This chapter presents the results of the experiments. Section 8.1 shows the results of the validation of the mathematical model and simulation model. Section 8.2 compares the results of the current approach to the proposed approach of batch releases. Finally, Section 8.3 shows the results of the experiments with different time intervals for the batch releases.

8.1 Model validation

Model validation is done for both the simulation model and the mathematical model. In Section 8.1.1 the results of the validation of the simulation model are presented. This is followed by the results of the validation of the mathematical model in Section 8.1.2.

8.1.1 Validation simulation model

Validation is done separately for the days in the peak period and the days outside the peak period. As explained in Section 7.2.3, in order to validate the simulation model, the output of the simulation model with the current approach of releasing pick batches is compared to the historical data. In Table 8.1 an overview is given of the average throughput, expressed in the average number of totes processed per hour, average sojourn time in seconds, and average WIP expressed in number of totes, of twelve days during the peak period. This is followed by a similar overview of twelve days outside the peak period in Table 8.2.

Data source	Throughput	Sojourn time	WIP
Historical data	560	4684	734
Simulation model	565	3769	769

Table 8.1: Sample average of the simulation model and historical data of the days during the peak period

Data source	Throughput	Sojourn time	WIP
Historical data	355	4673	455
Simulation model	312	4690	470

Table 8.2: Sample average of the simulation model and historical data of the days outside the peak period

The results in Table 8.1 and Table 8.2 show that Little's Law does not hold for the results of the simulation model. This can be explained by the fact that the simulation starts with an empty system and ends with a system that is not empty yet. The reason for this is that the simulation model ends exactly at the end of the work schedule of the operators, whereas in practice some overtime might be required to finish the final batches.

In Table 8.3 an overview of the 99% confidence intervals of the performance measures of the days in the peak is given, followed by a similar overview of the days outside the peak in Table 8.4.

Data source	Throughput	Sojourn time	WIP
Historical data	(501, 619)	(4087, 5281)	(589, 880)
Simulation model	(504, 625)	(3352, 4186)	(625, 912)

Table 8.3: Confidence intervals of the simulation model and historical data of the days during the peak period

Data source	Throughput	Sojourn time	WIP
Historical data	(311, 398)	(3669, 5677)	(368, 541)
Simulation model	(261, 363)	(3460, 5920)	(372, 569)

Table 8.4: Confidence intervals of the simulation model and historical data of the days outside the peak period

The results in Table 8.3 and Table 8.4 show that the confidence intervals of the performance measures are large for both the historical data and the simulation

model. This could be explained by the varying number of operators between the days both within and outside the peak period. Between some days the difference in the number of operators is big, with a higher variance in the performance measures and thus a larger confidence interval as a result.

In addition, the results in Table 8.3 show only a small overlap in the confidence intervals of the sojourn time. There seems to be a significant difference for this measure between the historical data and the simulation model. On the other hand, the confidence intervals of the other performance measures show a lot of overlap between the historical data and the simulation model. For the days outside the peak, presented in Table 8.4, the confidence intervals of all performance measures overlap a lot for the historical data and the simulation model. Whether the differences in the performance measures between the historical data and the simulation model are significant, is determined by the results of the t-tests.

In Table 8.5 and Table 8.6, the results of the t-tests are given for the days in the peak period and the days outside the peak period respectively. The t-tests are performed for a significance level of 1%, since the simulation model should be an accurate representation of reality as possible.

Result	**Throughput**	**Sojourn time**	**WIP**
Test statistic	0.52	-5.60	0.77
T-value ($\alpha = 0.01$)	3.11	-3.11	3.11
Significant	no	yes	no

Table 8.5: t-tests of the simulation model and historical data of the days during the peak period

Result	**Throughput**	**Sojourn time**	**WIP**
Test statistic	-2.13	0.04	0.39
T-value ($\alpha = 0.01$)	-3.11	3.11	3.11
Significant	no	no	no

Table 8.6: t-tests of the simulation model and historical data of the days outside the peak period

The results in Table 8.6 show that the differences between the simulation model and the historical data are insignificant for all performance measures with a 1% significance level. The results in Table 8.5 show a significant difference between the simulation model and the historical data for the sojourn time with a 1% significance level, whereas the difference in the throughput and WIP is regarded insignificant.

The difference in the sojourn time could be explained by the fact that the simulation model does not take into account breaks of the operators or interruptions caused by operators from the evening shift replacing the operators from the day shift. In practice it could happen that an operator started working on sorting or packing the items in the tote before his or her break, put the process on hold during the break, and resumed the process after the break. Besides that, it could happen that operators change in the middle of processing a tote and the new operator requires some start up time. In the simulation model this is not possible since breaks are not included and operators change instantly without loss of time. Since the same holds for the simulation model used for the days outside the peak period, one would expect a similar effect on the sojourn time there but that is not the case.

Another explanation for the differences in the sojourn time could be the differences in the service times. In practice, the service times depend on the number of items in the pick batch and this number varies throughout the day. The simulation model makes use of service time distributions and generates a random variate each time a pick batch requires service at a work station. These service time distributions are fitted over all service times of that day. It is then assumed in the simulation model that each pick batch contains the same amount of items and follows this distribution. As a consequence, two pick batches that have entered the process at the same time in reality as in the simulation model end up with completely different service and waiting times. Furthermore, the functionality of the stingray in the simulation model is a simplification of reality. Therefore, the sequence of the release of totes and pools of totes to the outbound lines could be slightly different, which results in different service and waiting times of the pick batches as well.

As the number of pick batches that go through the system during the peak period is much larger than outside the peak period, the total impact of the differences in the service and waiting times is much larger than outside the peak period. In addition, an explanation for the differences in impact between the peak period and non peak period could be that the service time distributions fit better for the days outside the peak period than the days during the peak period.

Based on the results presented in Table 8.1 to Table 8.6, the simulation model is regarded as an acceptable representation of the BFC outbound process. Still, one should take into account the differences between the simulation model and the actual process and their impact on the results in further experiments. This holds especially for the sojourn time of the simulation model for the days in the peak period, where a significant difference was observed.

8.1.2 Validation mathematical model

Similarly to the simulation model, validation of the mathematical model is done separately for the days in the peak period and the days outside the peak period. As explained in Section 7.2.3, in order to validate the mathematical model, the output of the model is compared to the output of the simulation model with the proposed batch release approach. In Table 8.7 an overview is given of the average throughput, expressed in the average number of totes processed per hour, average sojourn time in seconds, and average WIP expressed in number of totes, of the twelve days during the peak period. This is followed by a similar overview of the twelve days outside the peak period in Table 8.8.

Data source	**Throughput**	**Sojourn time**	**WIP**
Simulation model	754	3263	673
Mathematical model	856	4674	1112

Table 8.7: Sample average of the simulation model and mathematical model of the days during the peak period

Data source	**Throughput**	**Sojourn time**	**WIP**
Simulation model	376	2705	244
Mathematical model	355	8952	884

Table 8.8: Sample average of the simulation model and mathematical model of the days outside the peak period

Table 8.7 and Table 8.8 show some differences in the performance measures, especially for the sojourn time and WIP, between the simulation model and the mathematical model. To obtain a better impression of the differences in the performance measures of these models, the 99% confidence intervals are presented in Table 8.9 and Table 8.10 for the days during the peak period and outside the peak period respectively.

Data source	**Throughput**	**Sojourn time**	**WIP**
Simulation model	(604, 904)	(2995, 3531)	(442, 905)
Mathematical model	(692, 1021)	(3021, 6327)	(888, 1336)

Table 8.9: Confidence intervals of the simulation model and mathematical model of the days during the peak period

Data source	Throughput	Sojourn time	WIP
Simulation model	(317, 434)	(2504, 2907)	(194, 293)
Mathematical model	(282, 429)	(5352, 12551)	(603, 1165)

Table 8.10: Confidence intervals of the simulation model and mathematical model of the days outside the peak period

The results in Table 8.9 and Table 8.10 show that the confidence intervals of the performance measures are large, especially for the mathematical model. As explained before in Section 8.1.1, this could be explained by the highly varying number of operators between the days with a high variance in the performance measures and thus a larger confidence interval as a result.

T-tests are performed in order to determine whether the differences in the performance measures, that already become apparent when looking at the confidence intervals, between the simulation model and the mathematical model are significant. The results of these t-tests are presented in Table 8.11 and Table 8.12 for the days during the peak period and outside the peak period respectively.

Result	Throughput	Sojourn time	WIP
Test statistic	6.09	2.53	4.06
T-value ($\alpha = 0.01$)	3.11	3.11	3.11
Significant	yes	no	yes

Table 8.11: t-tests of the simulation model and mathematical model of the days in the peak period

Result	Throughput	Sojourn time	WIP
Test statistic	-3.48	5.61	7.81
T-value ($\alpha = 0.01$)	-3.11	3.11	3.11
Significant	yes	yes	yes

Table 8.12: t-tests of the simulation model and mathematical model of the days outside the peak period

The results in Table 8.11 and Table 8.12 show a significant difference between the results of the mathematical model and the results of the simulation model with the proposed approach for all performance measures with a significance level of 1%, except for the sojourn time in the peak period. The difference in the sojourn time in the peak period is however regarded significant for a significance level of 5%, which has a T-value of 2.20. It can thus be concluded that the simulation model and the mathematical model provide significantly different results for the proposed

approach. This could be explained by the differences between the models. Both models contain a simplified version of the stingray, each in a different way. Besides that, the mathematical model does not take into consideration the buffer areas of the individual work stations, whereas these are included in the simulation model. Furthermore, the difference could be explained by the fact that the mathematical model determines the performance for the steady state but the results of the simulation model also include the beginning and the end of the day, which are phases in which the system is not in its steady state.

Taking a closer look at the performance of the individual stations in the simulation model as well as the mathematical model, it seems that the mathematical model underestimates the pool completion time and suggests much higher waiting times at the sorting and packing stations. In Table 8.13 and Table 8.14, the confidence intervals of the pool completion time and waiting time for the sorting and packing stations are presented.

Data source	**Pool completion time**	**Waiting time**
Simulation model	(229, 319)	(509, 624)
Mathematical model	(29, 52)	(1361, 2104)

Table 8.13: Confidence intervals of the simulation model and mathematical model of the pool completion time and waiting time of the days in the peak period

Data source	**Pool completion time**	**Waiting time**
Simulation model	(134, 176)	(365, 552)
Mathematical model	(41, 125)	(1069, 3862)

Table 8.14: Confidence intervals of the simulation model and mathematical model of the pool completion time and waiting time of the days outside the peak period

For both the peak period and the non peak period holds that there is no overlap in the confidence intervals of the pool completion time and the waiting time for the sorting and packing stations of the simulation model and the mathematical model. The significance of the differences in the results is demonstrated by the results of the t-tests presented in Table 8.15 and Table 8.16.

Result	**Pool completion time**	**Waiting time**
Test statistic	-13.52	9.11
T-value ($\alpha = 0.01$)	-3.11	3.11
Significant	yes	yes

Table 8.15: t-tests of the simulation model and mathematical model of the pool completion time and waiting time of the days in the peak period

Result	Pool completion time	Waiting time
Test statistic	-6.00	4.53
T-value ($\alpha = 0.01$)	-3.11	3.11
Significant	yes	yes

Table 8.16: t-tests of the simulation model and mathematical model of the pool completion time and waiting time of the days outside the peak period

From the results in Table 8.15 and Table 8.16 it can be concluded that there is a significant difference between the simulation model and the mathematical model regarding the pool completion time and the waiting time at the sorting and packing stations. The simulation model incorporates the logic of the stingray to only release totes or pools of totes to the sorting and packing stations if there is room in the buffers. Until that time, the totes and pools of totes need to wait in the stingray. The mathematical model on the other hand only waits for pool completion and then sends the totes immediately to their next station. This difference between the models could explain why the expected waiting times at the sorting and packing stations are higher for the mathematical model than the simulation model. However, the observed difference is very large and there should not be a significant difference between the pool completion times of the models.

Based on the results presented in this section, the mathematical model is not regarded as a good representation of the process in practice at this stage. Improvements must be made to the model to obtain a more accurate estimation of the pool completion time and the waiting times at the sorting and packing stations.

8.2 Current approach versus proposed approach

In order to determine whether the proposed approach of releasing pick batches to the system performs better than the current approach, several analyses are done. Since the mathematical model is not regarded as a good representation of the actual process, the comparison is done by only comparing the results of the simulation model with the current batch release approach and the simulation model with the proposed batch release approach. In Section 8.2.1 the average results, confidence intervals and corresponding t-tests are presented. Section 8.2.2 presents the results of a number of analyses of the performance of the two approaches over time.

8.2.1 Average results and t-tests

Similar to the validation process, the comparison of the current and proposed approach is done separately for the days during the peak period and the days outside the peak period. In Table 8.17 an overview is given of the average throughput, sojourn time and WIP of the twelve days during the peak period. This is followed by an overview of the confidence intervals of the performance measures in Table 8.18.

Data source	**Throughput**	**Sojourn time**	**WIP**
Simulation (current)	565	3769	769
Simulation (proposed)	754	3263	673

Table 8.17: Sample average of the simulation model with both approaches of the days in the peak period

Data source	**Throughput**	**Sojourn time**	**WIP**
Simulation (current)	(504, 625)	(3352, 4186)	(625, 912)
Simulation (proposed)	(604, 904)	(2995, 3531)	(442, 905)

Table 8.18: Confidence intervals of the simulation model with both approaches of the days in the peak period

Based on the results in Table 8.17 and Table 8.18, it seems that the throughput is significantly higher and the sojourn time significantly lower for the proposed approach. There seems to be a difference between the two approaches in the average WIP but the confidence intervals show a considerable overlap. In Table 8.19 the results of the t-tests are presented in order to determine whether these differences are significant or not.

Result	**Throughput**	**Sojourn time**	**WIP**
Test statistic	5.11	-2.85	-0.95
T-value ($\alpha = 0.01$)	2.72	-2.72	-2.72
Significant	yes	yes	no

Table 8.19: t-tests of the simulation model with both approaches of the days in the peak period

The results in Table 8.19 show a significant difference between the results of the simulation with the current and the proposed approach for the throughput and sojourn time with a 1% significance level, whereas the differences in the WIP are

regarded insignificant. From these results it can thus be concluded that the proposed approach of releasing pick batches results in a higher average throughput of approximately 33%. One should however take into consideration that the simulation model does not take into account breaks of the operators or interruptions caused by operators from the evening shift replacing the operators from the day shift. The control room already accounts for these effects in their approach of releasing pick batches. Each operator has three breaks that add up to one hour in total. There are two shifts during the day, which means that there are two hours during the day that an operator is not working. A production day equals 16.5 to 17 hours, depending on whether it is the peak period or not. This means that the breaks add up to approximately 12% of the production time. In practice, the increase in throughput is therefore somewhat lower than the simulation model suggests. Based on this information, the increase in the average throughput is expected to be approximately 17%.

A similar overview of the average performance of the current and proposed approach of the days outside the peak period is given in Table 8.20, followed by an overview of the confidence intervals of the performance measures in Table 8.21.

Data source	**Throughput**	**Sojourn time**	**WIP**
Simulation (current)	312	4690	470
Simulation (proposed)	376	2705	244

Table 8.20: Sample average of the simulation model with both approaches of the days outside the peak period

Data source	**Throughput**	**Sojourn time**	**WIP**
Simulation (current)	(261, 363)	(3460, 5920)	(372, 569)
Simulation (proposed)	(317, 434)	(2504, 2907)	(194, 293)

Table 8.21: Confidence intervals of the simulation model with both approaches of the days outside the peak period

The results in Table 8.20 and Table 8.21 indicate a substantially higher throughput, lower sojourn time and lower WIP. In Table 8.22 the results of the t-tests are presented in order to confirm whether these differences are significant or not.

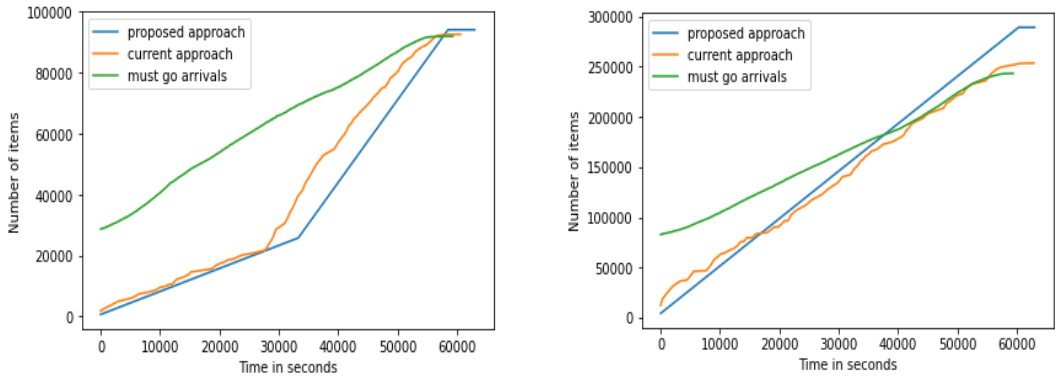
Result	Throughput	Sojourn time	WIP
Test statistic	5.68	-5.33	-7.53
T-value ($\alpha = 0.01$)	2.72	-2.72	-2.72
Significant	yes	yes	yes

Table 8.22: t-tests of the simulation model with both approaches of the days outside the peak period

The results in Table 8.22 show a significant difference between the results of the simulation model with the current and proposed approach for all performance measures with a significance level of 1%. From these results it can thus be concluded that the proposed approach of releasing pick batches results in a higher throughput, shorter sojourn times, and lower WIP levels for days outside the peak period. The increase in average throughput is approximately 21%. Again, one should take into consideration that the simulation model does not take into account breaks of the operators and interruptions during the switch of operators between the day and evening work shifts. Taking this into account, a better approximation is an expected increase in the average throughput of approximately 6%.

8.2.2 Results over time

Besides a comparison on the average performance measures, the performance of the current and proposed approach is also analyzed over time by the simulation model. In Figure 8.1, the cumulative arrival of must go items, which are items that must be shipped today, are depicted together with the cumulative number of items that are released to the system with the current and proposed approach for a day in the peak period and a day outside the peak period. The figures of the other days during the peak period and outside the peak show comparable results and can be found in Appendix B.



(a) Non peak period

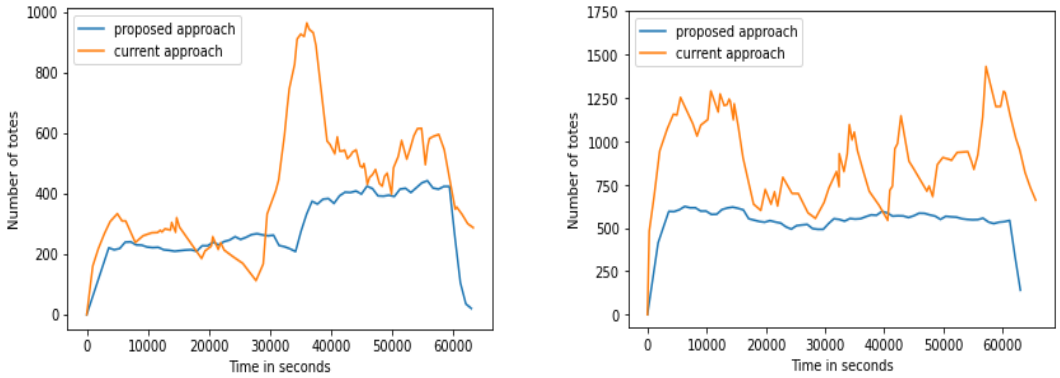
(b) Peak period

Figure 8.1: Must go items arrived and items released over time

In the non peak period, it can be seen that the number of items released to the system are fairly similar in the beginning for the current and proposed approach. In the middle of the production day, a significant increase in the release rate is observed for both the current and the proposed approach. This increase in the release rate is due to the increase of operators from the day shift to the evening shift. For the current approach this takes place one hour earlier than the proposed approach. The reason for this is that with the current approach a buffer with workload of approximately one hour is created for the stations that are opened in the evening shift. This buffer is not created in the proposed approach.

In the peak period, it can be seen that the number of items released to the system exceeds the number of must go items that have arrived in the second half of the production day. This means that not only must go items but also could go items, which are items that may also be shipped the next day instead of today, were released to the system. In the beginning of the day, the release rate is a bit higher with the current approach but soon the release rate of the proposed approach overtakes the current approach. At the end of the day a significant amount of additional items is released with the proposed approach in comparison to the current approach.

In Figure 8.2, the work in progress during the day, expressed in the number of totes in the system, is depicted of the current and the proposed approach for a day in the peak period, as well as a day outside the peak period. The figures of the other days during the peak period and outside the peak show comparable results and can be found in Appendix C.



(a) Non peak period (b) Peak period

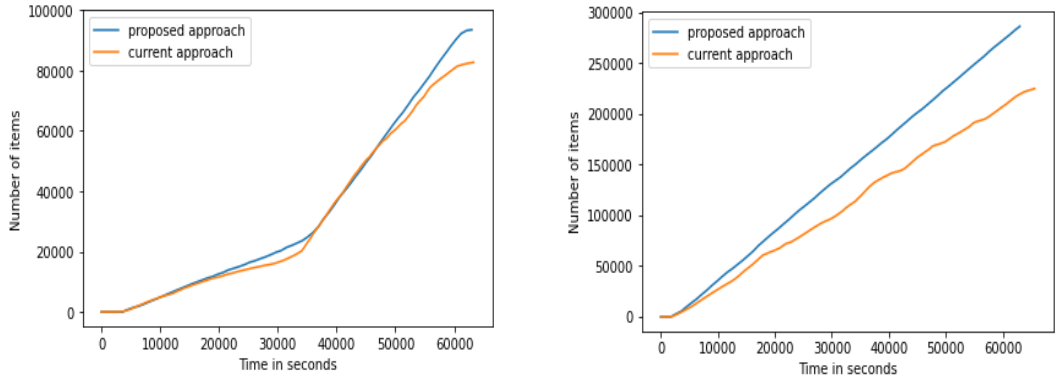
Figure 8.2: Work in progress over time

For both the peak and non peak period holds that the WIP levels are relatively stable for the proposed approach and fluctuate a lot for the current approach. In the non peak period, an increase in the WIP level can be seen at the middle of the production day, which is the moment that the day shift is replaced by the evening shift. In the same time period, a huge increase in the WIP level with the current approach is observed. This can be explained by the buffer that is created one

hour in advance of the switch between the day and evening work shifts. During that hour, the release rate of items to the system is already increased whereas the number of operators remains the same. As a result the rate in is much larger than the rate out, such that the WIP level is increased. In the peak period, the difference between the number of operators in the day shift does not differ that much from the number of operators in the evening shift, therefore the impact is much lower there. Other factors that cause the fluctuations in the WIP levels are the fluctuations in the release of items to the system and the stochastic service times.

Another observation here, is that at the end of the day the WIP levels are still fairly high for the current approach. This means that the system is not empty yet and still some time is required to process the pick batches that are still in the system. This could be explained by the fact that the simulation model starts and stops exactly at the times that the operators are hired for, whereas in practice a little overtime might be required to finish the final batches.

In Figure 8.3, the cumulative number of items processed is depicted of the current and proposed approach for a day in the peak period and a day outside the peak period. The figures of the other days during and outside the peak period show comparable results and can be found in Appendix D.



(a) Non peak period

(b) Peak period

Figure 8.3: Items processed over time

For both the peak and non peak period holds that in the end more items are processed with the proposed approach than the current approach. The difference is larger in the peak period than the non peak period. In the non peak period, the number of items processed with the current and proposed approach run almost parallel in the beginning but near the end of the day, the processing of items with the proposed approach continues steadily whereas the processing of items with the current approach attenuates. In the peak period the proposed approach outperforms the current approach right from the start of the day. The difference between the non peak period and peak period can be explained by the number of operators working. In the non peak period the number of operators is much

lower, especially in the mornings, and a lower amount of operators results in less additional processed items.

8.3 Batch release time intervals

The proposed approach provides a batch release rate for each outbound line. The question is however which time interval should be used for these rates. Therefore, experiments have been done for several time intervals in order to assess the impact. The results for the days in the peak period are presented in Table 8.23 and those for the days outside the peak period can be found in Table 8.24.

Time interval	**Throughput**	**Sojourn time**	**WIP**
10 minutes	739	3285	712
12.5 minutes	800	3053	655
15 minutes	754	3263	673
17.5 minutes	795	3186	654
20 minutes	798	3206	639
22.5 minutes	795	3419	689
25 minutes	797	3512	693

Table 8.23: Average performance per time interval of days in the peak period

Time interval	**Throughput**	**Sojourn time**	**WIP**
10 minutes	374	2563	242
12.5 minutes	207	3355	506
15 minutes	376	2705	244
17.5 minutes	201	3692	180
20 minutes	201	3793	185
22.5 minutes	200	3932	188
25 minutes	201	4131	199

Table 8.24: Average performance per time interval of days outside the peak period

The main reason for the differences in the results between the time intervals is caused by how well the batch release rates, which are based on the processing times of the sorting and packing stations, can be converted to that time interval. For example for the days in the peak period, the longest average packing time is approximately 12.5 minutes. In order to be able to release one pick batch to this station in every time interval, the time interval should be equal to or larger than 12.5 minutes. Similarly, the shortest average packing time for days in the peak

period is approximately 2.5 minutes. In that case, the time interval should ideally equal or be a multiple of 2.5 minutes.

As new orders come in all the time and the items in the warehouse are distributed randomly, it is preferred to release pick batches to the system at the latest time possible. The reasoning behind this is that more productive pick batches can be created if there are more items to choose from. By releasing the pick batches at the latest time possible, it is prevented that pick batches, that have been sent to the system already but have not been processed yet, could have been more productive if the items of later incoming orders were considered as well.

The best time interval is therefore the shortest time interval in which the rates (or multiple of the rates) of the different outbound lines fit best. These rates are highly dependent on the service time distributions of the outbound lines so these must be as accurate as possible.

Chapter 9

Discussion

The previous chapter presented the results of the experiments performed with the mathematical model and simulation model. In this chapter the limitations of the models are discussed in Section 9.1, followed by the recommendations in Section 9.2.

9.1 Limitations

One obvious limitation of the mathematical model is the simplification of the functionality of the stingray. As a result, the mathematical model underestimates the pool completion time and suggests higher waiting times at the sorting and packing stations. Another limitation of the mathematical model is that it does not account for the buffer areas of the individual work stations. As a consequence the model allows for longer queues at the individual sorting and packing stations, whereas in practice these pick batches would be waiting in the stingray until a spot becomes available in the queue of the next station.

Similarly to the mathematical model, a limitation of the simulation model is the simplification of the functionality of the stingray. The release sequencing of totes and pools of totes to the outbound lines in the simulation model does not fully correspond to the sequencing in practice. In practice some additional rules apply. As a result, totes or pools of totes might be waiting longer in the simulation model than they would have in practice and vice versa.

Furthermore, a limitation for both models is that in practice the service times are dependent on the number of items and for picking also on the distances, whereas the models assume that each pick batch for a designated outbound line is of equal size and uses a distribution fitted on the service times of all pick batches sent to that designated outbound line during the day.

9.2 Recommendations

Regarding the mathematical model, it is recommended to improve the incorporation of the stingray logic in the complete reduction method. One might be able to reduce the underestimation of the pool completion time by improving the estimation of the inter-arrival times of pick batches that belong to the same pool. Besides that, a solution may be found to take into account the buffer space of the sorting and packing stations.

With respect to the simulation model, it is recommended to formulate service times that are dependent on the number of items in the pick batch and make the size of the pick batches stochastic. The next level would be to make a connection with Pacman, such that the pick batches are created in the same manner as they would be in practice. In addition, it is recommended to improve the functionality of the stingray by trying to incorporate the full logic that is used in practice. Furthermore, the validation of the simulation model could be improved by performing the same analysis for the current approach using the actual pool sizes and the actual processing times at each station of the individual totes.

In relation to the LP introduced for distributing the arrival rate of pick batches over the outbound lines, it is recommended to examine the impact of the weights in the objective function in case the sorting and packing stations are no longer the bottleneck.

Ultimately, it is recommended that the release of pick batches is done in a more timely manner and in such quantities that the system is balanced. It is recommended that the overall release rate of pick batches is slightly lower than the processing rate of the bottleneck and the release rate of pick batches for each outbound line is slightly lower than the processing rate of that outbound line. In addition, it is recommended that the time interval of the batch releases equals the shortest time interval in which the rates or multiple of the rates of the different outbound lines fit best. Moreover, it is recommended to research how the processing rates of the outbound lines can be determined as accurately as possible, since these are highly dependent on the number of items in a pick batch which could vary a lot throughout the day.

The company is recommended to start researching how the processing rates of the outbound lines can be determined accurately. After that, the logic of the proposed approach of releasing pick batches in a more timely and balanced manner can be implemented. Next, the company could incorporate the logic of balancing the order basket levels in order to automate this process as well.

At last, an idea for future research is to investigate how the same logic could be applied to a warehouse in which pick batches are created in and coordinated from multiple areas. This is exactly what will happen in BFC2, which is the new warehouse of the company.

Chapter 10

Conclusion

The main research question of this research project is: *"How should the workload at the different work stations in the warehouse be allocated such that the overall throughput is maximized and the operating costs are minimized, while maintaining the order fulfillment score?"*

By means of theoretical research, an approach for pick batch releases and a mathematical model were formulated. In addition, a simulation model was created for validation purposes and to be able to better compare the performance of the current and proposed approach for pick batch releases.

Based on the results of the experiments in this research project, it can be concluded that the mathematical model does not represent the reality accurately enough. This is mainly caused by the discrepancies of the stingray. Therefore, future research should focus on how the model can be adjusted such that the gap between the functionality of the stingray in the model and in practice becomes smaller.

Furthermore, it can be concluded from the results of the experiments that the proposed approach for releasing pick batches to the system performs significantly better than the current approach. A big advantage of the proposed approach is the more constant WIP levels, which means that the risk of peaks in the WIP levels is lower. Therefore, less capacity needs to be reserved for the outbound process and can be used more efficiently for other processes in the warehouse.

Moreover, the proposed approach results in significantly higher throughput rates. The increase in throughput is expected to be approximately 17% during the peak period and approximately 6% outside the peak period. As a result, more customers orders can be processed by the end of the day or the number of operators can be reduced. During the peak period, the company often puts a break on the incoming customer orders by shutting down particular shops or postponing the delivery date. The increase in the throughput can reduce these kind of interventions. Consequently, more customer orders can be accepted and fulfilled.

Bibliography

[1] I. Adan and J. Resing. *Queueing Systems*. University Lecture Notes. 2015.

[2] I. F. Akyildiz and G. Bolch. "Mean Value Analysis Approximation for Multiple Server Queueing Networks". In: *Performance Evaluation* 8 (1988), pp. 77–91.

[3] S. L. Albin. "Poisson approximations for superposition arrival processes in queues". In: *Management Science* 28.2 (1982), pp. 126–137.

[4] B. Avi-Itzhak and D. P. Heyman. "Approximate queuing models for multiprogramming computer systems". In: *Operations Research* 21 (1973), pp. 1212–1230.

[5] F. Baskett et al. "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers". In: *Journal of the Association for Computing Machinery* 22.2 (1975), pp. 248–260.

[6] A. Bassamboo, J. M. Harrison, and A. Zeevi. "Dynamic Routing and Admission Control in High-Volume Service Systems: Asymptotic Analysis via Multi-Scale Fluid Limits". In: *Queueing Systems* 51 (2005), pp. 249–285.

[7] D. Bertsimas, D. Gamarnik, and A.A. Rikun. "Performance analysis of Queueing Networks via Robust Optimization". In: *Operations Research* 59.2 (2011), pp. 455–466.

[8] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis. "Stability Conditions for Multiclass Fluid Queueing Networks". In: *Transactions on automatic control* 41.11 (1996).

[9] G. R. Bitran and D. Tirupati. "Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference". In: *Management Science* 34.1 (1988), pp. 75–100.

[10] R. Buitenhek. *Performance evaluation of dual resource manufacturing systems*. Universiteit Twente, 1998. ISBN: 90-36512131.

[11] J.A. Buzacott and J.G. Shanthikumar. "Models for understanding flexible manufacturing systems". In: *AIIE Transactions* 12 (1980), pp. 339–350.

[12] R. Caldentey. "Approximations for multi-class departure processes". In: *Queueing Systems* 38 (2001), pp. 205–212.

[13] K. M. Chandy, U. Herzog, and L. Woo. "Parametric analysis of queueing networks". In: *IBM Journal of Research and Development* 19 (1975), pp. 36–42.

[14] A. E. Conway and N. D. Georganas. "Decomposition and aggregation by class in closed queueing networks". In: *IEEE Transactions on Software Engineering* SE-12.10 (1986), pp. 1025–1040.

[15] R. Cui, M. Li, and Q. Li. "Value of High-Quality Logistics: Evidence from a Clash Between SF Express and Alibaba". In: *Management Science* (2019). URL: https://doi.org/10.1287/mnsc.2019.3411.

[16] Y. Dallery. "Approximate analysis of general open queueing networks with restricted capacity". In: *Performance Evaluation* 11 (1990), pp. 209–222.

[17] D.L. Eager, D.J. Sorin, and M.K. Vernon. "AMVA techniques for high service time variability". In: *Performance Evaluation Review* 28.1 (2000), pp. 217–228.

[18] J. Eras. "Throughput time reduction in an e-fulfilment context: Design and evaluation of a job sequencing tool". In: *Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology* (2020).

[19] J. R. Jackson. "Networks of waiting lines". In: *Operations Research* 5.4 (1957), pp. 518–521.

[20] S. Kim. "Approximation of multiclass queueing networks with highly variable arrivals under deterministic routing". In: *Naval Research Logistics* 52.5 (2005), pp. 399–408.

[21] S. S. Lavenberg and M. Reiser. "Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers". In: *Journal of Applied Probability* 17.4 (1980), pp. 1048–1061.

[22] A. M. Law. *Simulation Modeling and Analysis*. McGraw-Hill Education, 2015. ISBN: 978-1-259-25438-3.

[23] J. Medhi. *Stochastic Models in Queueing Theory*. Second Edition. Academic Press, 2003. ISBN: 0124874622.

[24] D. C. Montgomery and G. C. Runger. *Applied Statistics and Probability for Engineers*. John Wiley Sons, 2011. ISBN: 978-0-470-50578-6.

[25] U. Narayan Bhat. *An Introduction to Queueing Theory*. Second Edition. Springer, 2010. ISBN: 9780817684204.

[26] H. G. Perros, Y. Dallery, and G. Pujolle. "Analysis of a queueing network model with class dependent window flow control". In: *Proceedings IEEE INFOCOM: The Conference on Computer Communications* 2 (1992), pp. 968–977.

[27] D.C. Petriu and C.M. Woodside. "Approximate mean value analysis based on Markov chain aggregation by composition". In: *Linear Algebra and Its Applications* 386 (2004), pp. 335–358.

[28] B. Rabta et al. "A hybrid analysis method for multi-class queueing networks with multi-server nodes". In: *Decision Support Systems* 54 (2013), pp. 1541–1547.

[29] K. Satyam, A. Krishnamurthy, and M. Kamath. "Solving general multi-class closed queuing networks using parametric decomposition". In: *Computers and Operations Research* 40.7 (2013), pp. 1777–1789.

[30] M. Schaeffer. *Het geheim van bol.com*. Vierde druk. Atlas Contact, 2017. ISBN: 9789047010937.

[31] S. Siha. "The pull production system: modelling and characteristics". In: *International Journal of Production Research* 32.4 (1994), pp. 933–949.

[32] Statista. *Number of digital buyers worldwide*. URL: `https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/`. (accessed: 09.09.2020).

[33] Statista. *Worldwide retail e-commerce sales*. URL: `https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/`. (accessed: 09.09.2020).

[34] J. M. Tarn et al. "E-fulfillment: the strategy and operational requirements". In: *Logistics Information Management* 16.5 (2003), pp. 350–362.

[35] H. Tijms. *Operationele analyse*. Vierde druk. Epsilon Uitgaven, 2013. ISBN: 9789050410755.

[36] A. Vatankhah Barenji et al. "Intelligent E-commerce logistics platform using hybrid agent based approach". In: *Transportation Research Part E: Logistics and Transportation Review* 126 (2019), pp. 15–31.

[37] W. Whitt. "A multi-class fluid model for a contact center with skill-based routing". In: *International Journal of Electronics and Communications* 60.2 (2006), pp. 95–102.

[38] W. Whitt. "The queueing network analyzer". In: *The Bell System Technical Journal* 62.9 (1983), pp. 2779–2815.

[39] W. Whitt. "Towards better multi-class parametric-decomposition approximations for open queueing networks". In: *Annals of Operations Research* 48.3 (1994), pp. 221–248.

[40] L. Zhang and D.G. Down. "A stable mean value analysis algorithm for closed systems with load-dependent queues". In: *ValueTools 2016 – 10th EAI International Conference on Performance Evaluation Methodologies and Tools* (2017), pp. 178–181.

[41] W. H. M. Zijm. *Manufacturing and logistic systems analysis, planning and control*. University Lecture Notes. 2012.