Structured production planning and control using a push-pull hybrid approach

Paul Maarleveld

5-11-2015

UNIVERSITEIT TWENTE.

Author

P.J. Maarleveld

Supervisory committee

First supervisor Dr. Ir. A. Al Hanbali

Second supervisor Dr. P.C. Schuur

Company supervisor Drs. X

Management summary

Problem & objective

The main problem for production planning at Company X is the high degree of variability. Company X is a supplier for Original Equipment Manufacturers and as such, it does not have its own products. Company X (engineers and) manufactures customer specific products on demand (pull-controlled). The market is characterized by short delivery times. Demand and product-mix are difficult to predict, if not impossible. It operates in an environment of "unknown unknowns". Not only is demand unknown, neither is the product which will be manufactured. Order priority can change rapidly, causing waiting time for others. As a result, the internal lead time increases and becomes uncontrolled. The production schedule is frequently infeasible. The main research question is:

"How could the planning function at Company X be improved so as to support the company's objective of controlling and reducing internal lead time?"

Approach

With information gathered during preliminary conversations, we define several problem areas pertaining to planning and control. These problems are translated into research questions. Based on the conclusions from a literature review and analysis of the current situation at Company X, we identify the current network of television screens as an opportunity to focus ourselves in the direction of a solution. The solution approach is adapted to fit the characteristics of Company X and tested using a simulation study to prove its favorable effect.

Findings

The current sales order lead time is X days on average, whereas the actual internal lead time is 13.8 days. 57% of the internal lead time consists of waiting time. Using queueing theory, we conclude that we cannot explain, predict or control how and when internal lead time increases. Nor is there a typical or average situation of demand and product mix. We are forced to accept that we cannot completely control our scheduled production using the current *Material Requirements Planning* (MRP) method (push-controlled). We find a gap between research and practice. Because neither a pure push nor pure pull approach is fitting, we come to understand that a hybrid push-pull approach is needed. None of the hybrid approaches described in the literature proves particularly useful. We identify four things to prevent the lead time from increasing uncontrollably:

1. A method to roughly assess required versus available capacity (push);
2. A method to schedule the production orders (push);
3. A flexible material control system to achieve an efficient order flow (pull);
4. A mechanism to guard the scheduled production order lead time between departments (pull).

The first two requirements are met by using the current MRP system, if applied correctly. We propose using the existing network of television screens in the factory to link capacity groups. On the screens we apply an order release mechanism based on available capacity authorization at the capacity groups. Implementation of this system requires no extra investment, except for the time to reprogram the computer system running the television screens.

The solution provides us with an approach to meet requirements 3 and 4. The order release mechanism based on available capacity in the next capacity group, prioritizes the orders in queue based on a pull strategy. We work on production orders for which we know capacity is available downstream, ensuring an efficient order flow (3). Meanwhile, the solution has the ability to guard the MRP schedule by limiting the allowable workload in a capacity group (4).

This approach is tested using a simulation study and the result provides us with a proof of concept. The new approach outperforms the current approach; it decreases the lead time by 15% and reduces its variability. The ability to maintain the MRP schedule is increased, while the throughput remains the same. Unfortunately, because no such thing as a typical or average situation of demand and product mix exists, the hypothetical simulation model is not suitable to predict how much reduction in lead time is to be expected if implemented. This is expected however and not the intended objective of the simulation study. Case studies in literature show lead time reduction between 22 and 70%.

This new approach is not a stand-alone solution to the problems at Company X. We identify the possibility to increase the chances of success by dividing the complex problem of planning into several planning levels. This also helps to make the planning function more transparent and understandable. A hierarchical framework is used to distinguish between a strategic, tactical and operational level, each with its own objectives, responsibilities and tools.

Recommendations

We recommend to implement the new approach as an operational tool in a hierarchical planning framework. It should be combined with a balanced workload over the capacity groups using MRP on a tactical level. We recommend to plan against 80% capacity on a tactical level to account for uncertainty. On an operational level we work at 100%. This will improve lead time, decrease *Work In Progress* (WIP) inventory in production, while maintaining throughput. Any unused safety capacity we recommend is used to fulfill short term customer orders, solve backlog or work ahead of schedule to free up capacity at a later time. On a strategic level, we recommend to strengthen the tactical and operational performance by facilitating with the right resources and clear objectives. For example, well-trained and well-maintained multifunctional employees to promote the exchange of flexible capacity.

The focus of this research is on production planning and control. For manufacturers of customer specific products, an important part of the process is of course the design of a product and the process of preparing it for release into production. We recommend to invest time and resources in researching the possibilities to improve the engineering and pre-production phase; control how much time is spend and to what end. Do we engineer until customer satisfaction only, or do we devote extra time to improve manufacturability? Related to this topic; invest time and resources into the possibility to convert initial work into repeat business as quickly as possible. Repeat business accounts for steady income at relatively low control and cost. Effort to turn an initial order stream into an easily manufactured, easily controlled order flow could pay off. We can think of design for manufacturing initiatives, to reduce the variance between products and improve overall quality. Or attempt to strengthen the customer relationship through (mutually) beneficial design changes and become first choice supplier.

Preface

After I finished my bachelor of applied science at the Hogeschool Utrecht, it felt as if I had not yet learned as much as I had hoped for. Looking back on my period at the University of Twente, I really have the feeling I challenged myself to go the extra mile and be as much as I can be. I feel very proud and satisfied to present to you the result of my efforts over the last nine months. A period during which I experienced ups and downs, but in the end came out stronger and wiser. I could not have done it without the support of some people I would like the thank.

First of all, I would like to thank the people at Company X for their collaboration and interest in my research. I would especially like to thank Jasper for his excellent guidance and the way he has supported me in achieving this great result. I sincerely hope that the outcome of this research will help Company X in its future endeavors.

Second of all, I would like to thank Ahmad and Peter for their efforts, valuable insights and pleasant cooperation.

Thirdly, I would like to thank my parents and my sisters. I am grateful for their love and support. I feel very fortunate to have had the possibilities my parents gave me.

Last but not least, I would like to thank Ruth for everything she has done for me and her believe in me during the good times and the bad.

Paul Maarleveld

Location X, November 2015

Table of content

List of abbreviations

BOM	Bill Of Materials
CODP	Customer Order Decoupling Point
ERP	Enterprise Resource Planning
ETO	Engineer-To-Order
FCFS	First Come First Serve
FCS	Finite Capacity Scheduling
HIHS	Horizontally Integrated Hybrid System
IT	Information Technology
JIT	Just In Time
MPS	Master Production Schedule
MRP	Material Requirements Planning
MTO	Make-To-Order
MTS	Make-To-Stock
OEM	Original Equipment Manufacturer
OPT	Optimized Production Technology
POLCA	Paired-cells of Overlapping Loops with Capacity Authorization
RCCP	Rough Cut Capacity Planning
TPS	Toyota Production System
VIHS	Vertically Integrated Hybrid System
WIP	Work In Progress

1. Introduction

1.1 Personal motivation

This thesis is written as part of my master program Industrial Engineering and Management to obtain the degree of Master of Science at the University of Twente. The research is conducted for the planning department at Company X over a period of nine months and focuses on introducing a new, structured method of production planning and control.

1.2 Company background

The research assignment is performed at Company X. Together with Company Y it forms the XY Group. Company X has been a production company since 19XX with a turnover of XX million euro in 20XX. Company X distinguishes itself as a quality supplier and partner of Original Equipment Manufacturers (OEM). The company has a varied customer base for whom it manufactures housings and frames for automated systems, such as baggage claims at airports, automated sorting systems for logistics operations and assembly systems for material handling.

Production is located at a plant in Location X, where around 100 people are employed. The company does not produce its own products. The products are customer specific, either designed by the customer itself or engineered by Company X's engineering department. Collaboration is important, from product creation to product realization, logistical support, assembly and maintenance. As such, the company has a diversified product mix, where no product is the same.

1.3 Origin of the problem

Company X operates in a difficult market with a lot of competition. Especially at this time of decreasing repeat business, it is important to attract new customers. Short delivery times at controllable cost are a strong competitive advantage. To this end, the management of Company X has set up a number of objectives:

- Decreased internal lead times;
- Increased efficiency in fulfilling orders with less Work In Progress (WIP) inventory and reduced unnecessary movement of materials;
- Increased effectiveness (increased delivery reliability and quality);
- Decreased operational cost.

Because production and planning are closely linked, Company X is looking for someone who is able to take a critical look at the planning process and come up with concrete solutions how the planning department can support the production process. Highly fluctuating demand makes it difficult to match demand with available capacity. Nevertheless, we are looking for a method to control and reduce the internal lead time, and increase the chances of a feasible schedule. Moreover, the intention is to make the planning process more transparent, more understandable and easier. Formulating a clear planning method with broad support throughout the organization, contributing to the achievement of the objectives of Company X.

Another challenge is the quantification of improvement initiatives. Validating them beforehand is currently a qualitative process for Company X. Quantitative validation of the impact is performed after the initiative has been implemented, without a sound, quantified justification of any investments beforehand. Company X realizes the importance of quantifiable solutions and is therefore interested in exploring possibilities to validate the impact of new solutions quantitatively beforehand instead. For the problems to be addressed in the remainder of this thesis, it is strongly

desired to quantify any solutions as well. Subsequently, Company X is able to compare the current situation and any possible solution to assess its impact.

1.3.2 Current planning process

The planning function of Company X can be described by the two different order flow patterns it distinguishes: so-called *initial orders* and *repeat orders*. The organization maintains a pull strategy, meaning that no production starts if there is no demand. This does not necessarily precludes building a product before it is due for delivery. In some cases, Company X works with framework contracts and therefore does have some insight into upcoming demand. In these cases it may decide to manufacture some parts upfront, awaiting final assembly. This is called the customer order decoupling point (CODP). After this CODP the manufacturing process is a pull strategy (demand-driven), before the CODP it is a push strategy (forecast-driven).

1.3.3 Software and underlying methods

Company X uses an ERP system from software manufacturer Company Z for its business. The underlying method that is used to generate production schedules is MRP. To produce an article, Company X uses numbered production orders. For every production order, a bill of materials (BOM) is generated specifying required components and materials, a standard routing which dictates the required processing steps and resources and a calculation of required production time. This information has been entered into the system by the engineering department at the time it was an initial order.

The deadline is the day upon which the final product should be ready for shipment to the customer. The overnight run of the MRP proceeds to schedule all steps in the routing and create new production orders for the parts in the BOM. It repeats this until all required parts for the finished product are scheduled. MRP works backwards in time, scheduling processes based on a standard off-set lead time on required resources, against infinite capacity.

1.4 Detailed problem description

Company X has a desire to reduce the lead time of its production orders to five work days. The final objective is to be able to produce 80% of its sales orders within this time. It should be noted that this pertains to internal lead time, from the moment production is started until the time of order expedition. It does not necessarily mean that five work days will become the external lead time with which Company X delivers its product, i.e., it is not the time from order acceptance until order expedition. The lead time of raw materials and other purchase components is still a limiting factor for the external lead time, as is the engineering process for initial orders. This aspect of the lead time falls outside the scope of this research, as this falls under the responsibilities of the purchasing department and engineering department respectively.

To accomplish this objective a number of problems need to be solved. These can roughly be categorized as problems related to knowledge, capacity and culture. These problems will now be discussed in more detail.

1.4.1 Knowledge

One of the main issues with the planning procedure is the fact that it depends on a single person to perform the task. This position is currently filled by someone with 30 years of experience at different positions within the company, thereby gathering intimate knowledge on virtually all business processes. First of all, this type of knowledge inside the mind of a single person makes an organization vulnerable. Secondly, this is the experience that is preferably captured in an automated system such that it is able to make a more intelligent schedule. Experience on, for example, sales and

forecasts is something that ideally would be translated into a sound Master Production Schedule (MPS) and smart parameter settings in an MRP system.

Another problem related to knowledge, is the fact that there is limited knowledge about planning as a function. The organization does not have a clear understanding of what the function and objective of a well-structured planning process is and how it can add value to the business.

1.4.2 Capacity problems

The main problem concerning capacity is related to order intake. When accepting a new order, there is not enough insight into and attention for the impact on current sales orders. The lack of overview results in more accepted work than free capacity to allocate. This manifests itself in different ways and situations.

The departments can be divided into two capacity restrained categories; first category is the manual labor constraint department where capacity is determined by the number of people. The other category is capacity restriction by machine. Each machine has a limited amount of work it can perform in an amount of time. According to the Director of Operations, machine capacity mainly forms the bottleneck for Company X. The foremen of all departments have autonomy and responsibility when it comes to organizing the work force. This means they can exchange workers with other departments when necessary but also hire temporary workers if needed. This should mean that manual labor capacity should not have to be an issue. That being said, asking around to verify turned up some issues in practice. First of all, qualified replacement is not readily available for some functions. For instance when it comes to programming the machines at the sheet metal center, for this type of work a skilled programmer is required. Secondly, it is difficult for the departments to predict when a peak in workload is coming due to uncertainty in demand. In reality, it is not always possible to organize extra labor in reaction to an upcoming capacity shortage.

Furthermore, the notion that Company X only manufactures on demand leads to a false idea that forecasts can be omitted as a restricting factor in the capacity allocation. This is problematic, as some customers have framework contracts with forecasts. As such, Company X has forecasts for the required amount of capacity to manufacture these items. This reserved capacity cannot be allocated to another customer without considering the consequences. However, this is what happens in the current situation where forecasted production capacity is ignored in favor of another customer.

The added difficulty in the planning process is the fact that a denied delivery date is not perceived to be a real option, especially for some of the (potentially) important, high revenue customers. These orders need to be fulfilled in a short amount of time, which causes disruptions in the production process. A strong desire from management is to look for possibilities to be flexible in capacity. This flexibility most likely comes down to adding extra machines as it is believed these form the most important bottleneck as previously stated.

In practice, the manufacturing activities are often not performed as planned. Sometimes because of unforeseen circumstances, sometimes due to unbalanced workload. Fact of the matter is that too often scheduled production planning is not feasible or is disturbed by short notice developments.

1.4.3 Cultural problems

Several problems in the current situation can be classified as cultural or behavioral issues, where the current mindset causes problems. The first cultural problems are related to the forecasts mentioned in the previous subsection 1.4.2. It happens on a regular basis that customers do not transform their forecasted order in a sales order. In this situation the forecast must be manually

moved in time in the system, a process the planner performs. It could be argued that this formation of forecast to sales order is a process the sales department should monitor, thereby freeing up the time for the planner to perform more value-adding tasks. A different problem occurs when the customer does call upon the forecasted order, however in a different quantity than agreed upon. Within ERP a minimal lot size is often required when starting a production, cost price of the item is based on this lot size. If the planner decides to overrule the system and start production for an actual smaller demand, economies of scale are partially lost, the actual cost price will be higher and the margin will decrease. This causes friction with the pre-production department, responsible for the pricing. To counteract this, the sales department uses graduated prices (Dutch: *Staffelprijzen*). On the other hand, if the planner decides to respect the lot size, a part of the items could end up in inventory. This contradicts the objective of Company X to manufacture on demand according to the Lean philosophy[1].

A second cultural problem stems from the imagined flexibility in production capacity. This method to cope with fluctuating capacity demand is perceived to require no further control policy. The first manifestation of this lack of control occurs when an increased workload calls upon the deployment of the flexibility. Temporarily ramping up production or solving backorders requires the foremen to actively communicate with one another in order to use excess capacity from departments. The foremen have the responsibility to do so but the feeling is that not all foremen are equally assertive. Conversations with the production employees on the other hand, point out that the support for this is perhaps insufficient. This lack of support is related to the second manifestation of the lack of control.

In the past, flexible teams have been attempted, workers have been trained in various processes besides their own. After these trainings, however, the workers predominantly worked in their original department on familiar processes, most likely because lack of control has not urged them to maintain the newly acquired skills. Furthermore, without proper control the skill sets of the individual employees have not been formally documented. This makes it difficult for the foremen to know where to acquire the additional capacity required to finish the work, or guarantee the quality of output.

Lastly, because of the management choice to monitor and actively pursue machine efficiency, production employees believe that idle machines should be avoided. This focus on machine efficiency (contradicting demand-driven production) counteracts interchanges between departments as it would cause efficiency to drop when an employee supports another department. Subsequently, one machine could be working ahead of schedule while others are behind.

1.5 Research question

Based on the problems which have been discussed in the previous paragraphs, the main research question of this thesis is:

"How could the planning function at Company X be improved so as to support the company's objective of controlling and reducing internal lead time?"

A number of sub questions have been devised to analyze the main problem in a more structured and detailed manner:

1. What has been written in academic literature regarding the problems facing Company X?

[1] In 2013 Company X started a Lean transformation initiative. See also section 2.3.

a. Can we explain and predict lead times?
b. Which problems related to ERP/MRP and capacity management are known?
c. What are well-known problems related to ERP and Lean (pull strategies in general)?
d. What is the objective of a well-defined planning function?
e. What should a structured planning function look like?

2. What does the production process look like and what is the current performance?
    a. How is the production process structured and what do the production departments look like?
    b. What is the current performance of the departments and Company X in general, and how is performance measured?
    c. Which performance problems and production characteristics form the largest obstruction towards completing Company X's objectives?
3. How could Company X implement an approach to control and reduce lead time?
4. How could Company X structure its planning function?

1.6 Research scope

This research focuses on the planning function itself by clearly defining its objective and added value to the overall performance of the company, and how it can be improved by devising a structured approach. Thereby supporting Company X's objectives as were introduced in the origin of the problem (section 1.3).

1.6.1 Inclusions

Firstly, a part of Company X's production process consists of outsourcing. To be able to control the throughput time of an order, outsourcing must be considered in the planning process. These production steps and their influence will be taken into account. Secondly, order intake has a significant influence on capacity and occupation of the resources and is a source of variability. To be able to create a realistic, feasible production schedule, it is necessary to take order intake into account in this research. Thirdly, for new projects and products the processing steps and times are often not known with certainty. This also has a strong influence on the planning process and its ability to generate feasible schedules. Therefore project and manufacturing preparation will be taken into account in this research.

1.6.2 Exclusions

Decisions regarding processing steps to outsource and where to outsource them, will not be included, as is the purchasing of raw materials and the logistics of distributing the final product to Company X's customers. Due to time constraints, the implementation of the solution falls outside the scope of this thesis as well.

1.7 Deliverables

- Report – a detailed study and analysis of the problem
- Solution method – a description of the possible solution and adaptation to fit Company X
- Proof of concept – a simulation model to test the solution
- Planning framework – a structured function, describing what the objective of production planning is and how decisions should be made

1.8 Approach

We start the research with a series of preliminary conversations with various people in the organization. With the information from these conversations, we define several problem areas pertaining to planning and control. Based on these problems we formulate a main research question

and multiple sub questions. A literature review is conducted to construct a framework of academic background information regarding the production planning and control problems facing Company X. We continue with a detailed description and analysis of the current production process and performance. With the information from the literature research and analysis of Company X, we focus ourselves on the direction of a solution and adapt it to fit the characteristics of Company X. A simulation study is used to prove the benefits of the solution approach.

1.9 Structure of the report

In chapter 2 we present a literature review to answer the first sub question. Chapter 3 describes the current production and planning situation at Company X, as well as the current performance. In Chapter 4 and 5 we respectively present a solution possibility and a proof of concept based on a simulation study. Chapter 6 answers the sub question how to structure the planning function at Company X. Finally, chapter 7 presents the conclusions on all sub questions and answer the main research question. Chapter 7 ends with possibilities for future research.

2. Theoretical framework

This chapter answers the first research question, "*What has been written in academic literature regarding the problems facing Company X?*". Throughout meetings with various people within the organization several topics of interest have come up pertaining to manufacturing methods in relationship with a controlled planning method and the problems that arise when the business strategy and manufacturing concepts are misaligned. These topics are:

- Explaining or predicting lead time and the occurrence of waiting time;
- Shortcomings of ERP/MRP systems;
- ERP in combination with Lean;
- Pull vs. Push and hybrid approaches;
- Structured planning.

This chapter presents a literature study in which these topics have been researched. Section 2.1 is a review of lead time control and prediction. Sections 2.2 through 2.4 are focused on planning and control approaches. The process is aimed at identifying opportunities and problems in scientific research and connect this to the practical situation at Company X. The objective is to find either common ground or gaps between science and practice. This common ground, or lack thereof, may serve as a basis to start looking for solutions. Section 2.6 is the last section of this chapter and addresses the theory behind a structured and well-defined planning function.

2.1 Lead times

Before any analysis can be carried out, the definition of lead time and waiting time is clearly stated here. We define the lead time of any production order as the time elapsed between the moment at which the order becomes available to be processed and the moment at which the order is completed. This time period includes actual time spent working on the product, as well as time spent waiting before capacity is available for processing. Waiting time is therefore defined as the time elapsed between the moment an order becomes available to be processed and the time processing is started. Waiting time usually accounts for a large portion of the total manufacturing lead time, estimates vary between 70 and 80% (Subba Rao, 1992) and it is therefore most interesting to understand how waiting time can best be decreased by understanding the characteristics of the business process which cause it to originate. Research has found a relationship between characteristics of manufacturing, lead times and the production performance. Chin (2009) and Subba Rao (1992) both apply queueing theory to explain the causes of waiting time.

Subba Rao (1992) explicitly mentions capacity planning and control as an area of the manufacturing process where queueing theory can be used to gain an understanding. The manufacturing process consists of a number of work centers (machines, departments) where work is performed. At each center, work arrives from the previous work center, and after processing is completed it moves on to the next. The simplest example is an assembly line, such as the one shown in Figure 2.1. Jobs arrive at station 1 for the first process step and move through the system to be finished at station *N*. Each station has a queue where products are allowed to be stored.

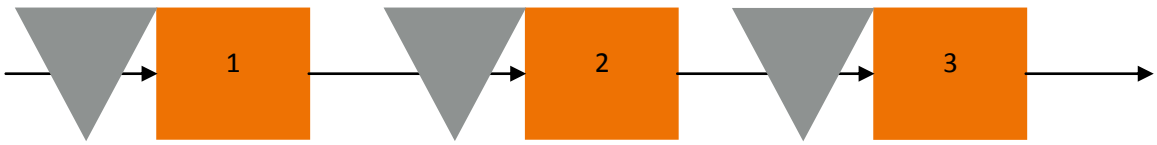


Figure 2.1: Example of a typical, simple assembly line process flow

This assembly system can be expressed as a serial system of work centers, expressed in terms of Kendall's notation, using three factors; *A/S/c*. *A* denotes the arrival process, *S* describes the service (processing) time and *c* is the number of servers at each work center (Winston, 2004). In the case of most manufacturing environments (and Company X is no exception) the arrival and service process do not follow any of the well-known and regularly used distribution functions. Therefore a general distribution (*G*) is used, such that the work center is expressed as a G/G/1 queuing model (for sake of simplicity the assumption is that each work center has one resource).

From queuing theory it is known that waiting times are influenced by three factors; variability (*V*), utilization or traffic intensity (*U*) and process time (*T*). The following formula known as Kingman's formula is used to explain and predict the occurring waiting time in a steady-state *G/G/1* system with a first come first serve (FCFS) priority:

$$W_q = V\ x\ U\ x\ T \approx \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu}$$

C_a is the coefficient of variation in the inter-arrival time, C_s is the coefficient of variation in process time, ρ is the utilization and $1/\mu$ is the average processing time. Utilization (ρ), also known as the traffic intensity, is the ratio $\frac{\lambda}{\mu}$, where λ is the number of arriving orders per unit of time and μ is the number of processed orders per unit of time. The term *U* predicts the average number of orders in queue and shows that if ρ approaches 1, this mean explodes (increases towards infinity). It is important to realize that these relationships are based on averages, an instantaneous arrival rate may exceed the service rate. Production orders will back up in queue, but as long as the arrival rate decreases afterwards the queue will not explode. On average however, the service rate must always exceed the arrival rate to avoid the system from overflowing.

The relationship *V x U x T* shows that the occurrence and magnitude of waiting time can be explained and influenced by analyzing and improving the parameters mentioned:

- Decreasing variability;
- Decreasing ρ;
- Decreasing $1/\mu$ (or conversely increase μ, the number of orders processed).

Shorter lead times can be attained by investigating the root causes that degrade each factor (Chin, 2009).

For most manufacturing organizations, the simple line assembly representation is not suitable. The process is better described as the network of diverging (station 1) and converging (station N) product flows in Figure 2.2. This complicates the situation somewhat, but for an organization manufacturing its own products queuing theory might still be directly applied by modeling it as a series of queues. The products, their structures and routings are deterministic (demand could be variable). As such the diverging and converging points in the network are known, and the proceeding station for each component is known. Therefore the arrival process for station *N* is the sum of the output of station 1 and 2, where the ratio of output from station 1 directly to station *N* (bypassing station 2) could be, for example, $\mu/3$.

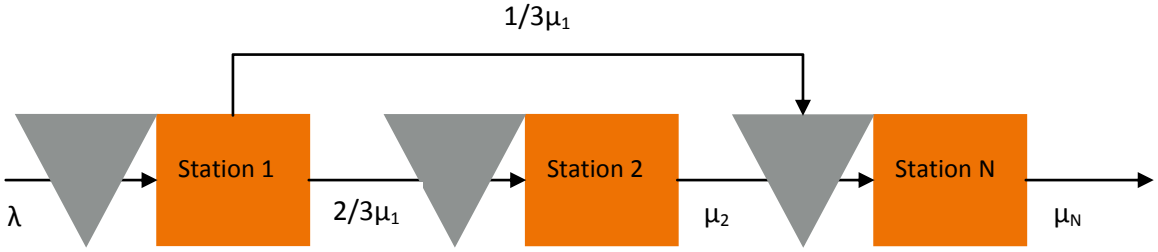


Figure 2.2: Example of a typical job shop environment

The fact that Company X is an MTO/ETO business, means that each product is different than the other in some way or another. Therefore it is unsure where any diverging and converging processing steps are, if any, and which ratio of the total output of a work center proceeds to the next (Figure 2.3). For instance, after being cut out of sheet metal, the components may go to any of the other departments in any kind of ratio. For initial products none of these relationships are known in advance. As such, the traffic pattern is complicated and uncertain. The situation at Company X can therefore not be modeled as a system of queues directly. The effects of variability, utilization and processing time should still apply though and therefore these are analyzed in the next section, because controlling WIP, assigning priorities and varying capacity over work centers could improve manufacturing efficiency (Subba Rao, 1992).

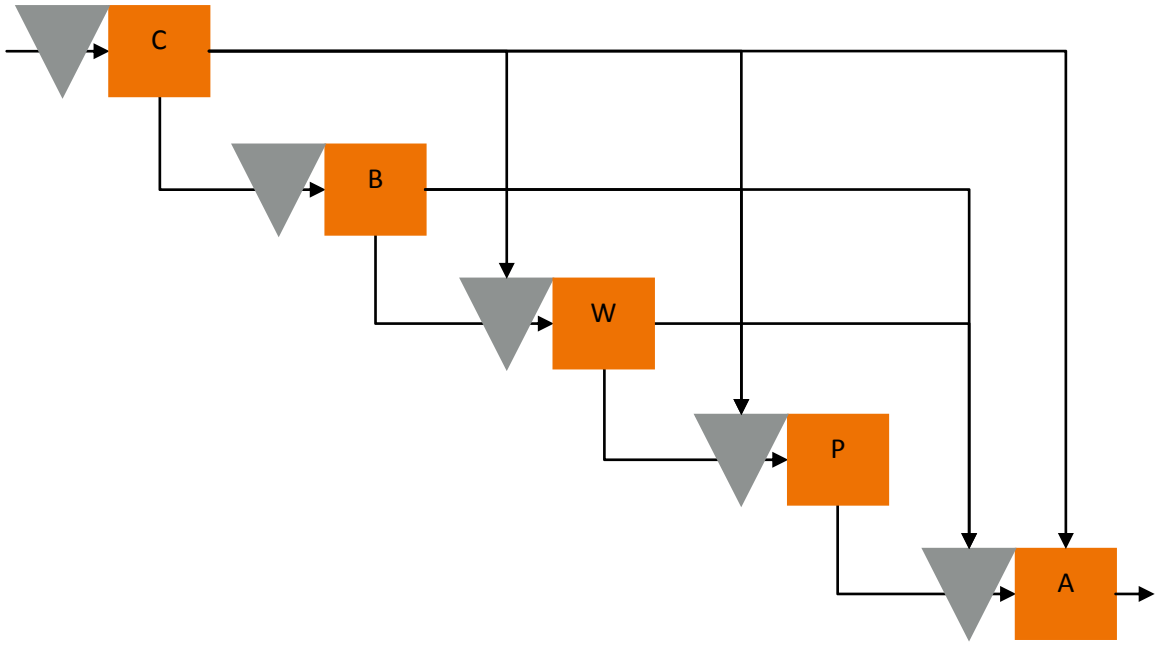


Figure 2.3: Order flow at Company X (C=Cutting, B=Bending, W=Welding, P=Painting, A=Assembling)

2.2 Shortcomings of Enterprise Resource Planning systems

2.2.1. The system

In short, an Enterprise Resource Planning (ERP) system is a (usually) commercial software package that enables the integration of transactions oriented data and processes throughout an organization (Markus, Axline, Petrie, & Tanis, 2000). In most cases, the ERP system works with Materials Requirement Planning (MRP) as its planning system. MRP is designed to translate predetermined demand or sales orders for finished products into a manufacturing schedule: *which*

materials are required, *when* are they required and *how many* are required (Murthy & Ma, 1991). The four major inputs for the system are:

1. The Master Production Schedule (MPS);
2. The product structure called Bill of Materials (BOM);
3. The resource requirements;
4. The inventory status.

Orders for finished products are obtained from the MPS; the required quantity and delivery dates for different products. By working through the levels of the product structures using the BOM, the finished product is divided into a hierarchy of required subassemblies, components and materials. All the requirements to meet the orders (several finished products might use the same component) are then aggregated into gross requirements in each time period. Using information about current inventory and work in progress, the MRP system computes the net requirements. The requirements are then offset in time to establish the dates by which actions (manufacturing or purchasing) must be initiated such that the components and subassemblies are timely available at different stages of manufacturing and the orders for finished products are met by the times specified in the MPS (Murthy & Ma, 1991).

ERP merely extends the system architecture of MRP with other business functions, without adding actual intelligent planning (Zijm, 1999). Otherwise put, an ERP system is mainly designed for transaction bookkeeping purposes. In its original intent it is good at monitoring events and bookkeeping trails of the events, but it is not meant to help the decision-making process (Moon & Phatak, 2005).

2.2.2 The problems

The previous section explained the basic principle behind MRP. It is this apparent procedural simplicity that makes MRP such a widely used system (Zijm, 1999). For some criteria it indeed outperforms other methods (e.g. the ability of order tracking and a high product variety), making it the best choice (Plenert, (1999)). However, a closer look reveals a number of issues. These can be divided into modeling errors (or wrongful assumptions), parameter errors and flexibility.

Before the MRP system works properly, the modeling of the real world situation and its underlying assumptions must be met (Euwe & Wortmann, 1997; Moon & Phatak, 2005; Zijm, 1999). Two of the most important assumptions are usually not met in reality; a deterministic character of the business processes (Moon & Phatak, 2005; Zijm, 1999) and an infinite production capacity (Bott & Ritzman, 1983; Euwe & Wortmann, 1997; Zijm, 1999). Simultaneous workplace occupation is often not considered when multiple orders require the same resource (Moon & Phatak, 2005).

To some extent the stochastic character of the business processes could be accounted for using the systems parameters. Modern MRP systems require many parameters to be set by the user regarding off-set lead times, lot sizes, set-up times and safety stocks. Unfortunately, the system does not provide any help in setting these. This creates a situation where increasing uncertainty is buffered with sufficiently long off-set lead times which tend to grow longer and longer (Zijm, 1999) the so-called lead time syndrome. Plenert (1999) argues that because lead times are labor based and performance is mostly measured by labor efficiency as well, MRP becomes a labor efficiency oriented system, using lead times as a way to maximize efficiency by building buffers of work in progress. He points out that this abuse of the parameters causes MRP systems to work improperly, rather than through design flaws. Either way, lead times tend to get inflated, according to Plenert (1999) to the point where 90 – 95% is waiting time. He mentions focusing on Finite Capacity Scheduling (FCS).

Various FCS tools have become extremely popular enhancements to the MRP process, but they present a new challenge. MRP needs to be running correctly and efficiently for these tools to be effective since they add another layer of complexity to the process (Plenert, (1999)).

Lastly, flexibility is a problem encountered in the literature. Euwe and Wortman (1997) describe a lack of alternative plans. Planners have several scenarios, e.g., lot splitting, subcontracting, overwork and such. This is not modeled in MRP, it generates only one plan. Nor does MRP generate alternative production schedules in case some materials or parts do not become available as planned (wrong quantities, inferior quality, machine break-down) (Moon & Phatak, 2005; Zijm, 1999). Moreover, Zijm (1999) points out that the system is inflexible in make-to-order organizations, requiring a rather detailed knowledge about which resources, materials and parts are needed when accepting customer orders. In this type of organization these conditions are often not fulfilled. Plenert (1999) claims MRP is in fact flexible in comparison with Just in Time (JIT) or Optimized Production Technology (OPT). First of all, because MRP handles multiple products. Second of all, scheduling flexibility allows a variety of products to be scheduled via any number of routings (Plenert, (1999)). The lack of explanation of Plenert (1999) leaves room for discussion about what this flexibility is or how it can be used to our advantage. MRP does not necessarily allocate lower level production to higher level assemblies, meaning that an assembly order is free to pull an item from any production routing that makes the required item available. This is after the fact however, when a part did not become available through its usual routing. Another interpretation could be about scheduling flexibility, considering multiple alternative routings beforehand. However, he offers no insight if programming multiple routings offers the planner a decision, or if MRP is able to make an intelligent decision on its own.

2.3 ERP and Lean manufacturing

In 20XX, Company X has started a Lean transformation and implementation. Based on the early impression of characteristics of Company X a decision to introduce Lean does not seem to be obvious. This section is meant to find the differences and similarities between Lean and ERP (applied at Company X).

In 1977, Sugimori *et al* published the first English paper on the Toyota Production System (TPS), a major precursor of the Lean philosophy. In their paper they are critical towards complicated IT systems as a means of organizing production logistics. It introduces unnecessary cost, over-production and uncertainty (in Riezebos, Klinkenberg, & Hicks, 2009). At the same time, in the Western industrialized world, IT and advanced automation were seen as the way to gain competitive advantage (Riezebos et al, 2009). One of the main contradictions is centralized (ERP) and decentralized (Lean) planning (Powell, Alfnes, Strandhagen, & Dreyer, 2012; Riezebos et al, 2009). According to Huoy (2005) Lean authors also advocate decentralized information processing and communication. Furthermore, combining the two was something that until recently seemed inappropriate because of the inherent contradiction between the push strategy of ERP (unlimited workload) and the pull strategy of Lean (limited workload) (Olhager & Östlund, 1990; Riezebos et al, 2009). The pull strategy pursuing ‘zero inventory’, uncovering any production problems or inefficiencies and solving them. The push strategy pursuing ‘zero risk’, being independent of production in case of product defects and machine break-downs (Villa & Watanabe, 1993). Needless to say, these two approaches have been regarded as incompatible for a long time. As a result, extensive information can be found in literature about either ERP implementation or Lean implementation, but applying both has not been well-documented.

Recent study has indicated that this point of view might be a misconception (Steger-Jensen & Hvolby, 2008) and ERP systems might in fact be used to support Lean practices (Powell, Alfnes,

Strandhagen, & Dreyer, 2013). Powell et al (2012) have performed a literature review where they have identified 15 areas where a well-configured ERP system could support Lean, see Table 2.1. These are evaluated through the usage of the fundamental principles of Lean identified by Womack and Jones (1996); *value, value stream, flow, pull and perfection* (in Powell et al, 2012).

Table 2.1: 15 points for ERP to support Lean (source: Powell et al, 2012)

No	Principle	An ERP system for Lean should:
1	Value	Support customer relationship management
2		Automate necessary non-value adding activities (e.g. back flushing)
3	Value stream	Enable process-modeling to support standard work processes
4		Provide a source for easy-to-find product drawings and standard work instructions
5		Support information sharing across the supply chain
6	Flow	Create synchronized and streamlined data flow (internal & external)
7		Support line balancing
8		Support demand leveling
9		Support orderless rate-based planning (e.g. takt-time)
10		Provide decision support for shop floor decision making
11	Pull	Support Kanban control
12		Support production leveling (Heijunka)
13		Support JIT procurement
14	Perfection	Provide a system to support root-cause analysis and for the logging and follow-up of quality problems
15		Provide highly visual and transparent operational measures (e.g. real time status against plan)

This extensive list of elements applies to the situation where the character of the production is equal to the traditional Lean assumption: high-volume production of low-variability products. For typical make-to-order and engineer-to-order organizations these assumptions are quite the opposite: high-variability products, manufactured in low volumes. Neither the TPS nor the pure takt-time control principle is deemed appropriate (Portioli-Staudacher & Tandardini, 2012; Slomp, Bokhorst, & Germs, 2009).

Applying Lean within its boundaries is proven to be successful and there is no need to combine it with ERP or any complex IT system for that matter. It is clear that the preferred method involves putting in place a simplified information management system (Houy, 2005). Outside of these boundaries the story changes. ERP might in fact play an important role to support Lean transformation when there is high variability in demand and products and low production volumes (Steger-Jensen & Hvolby, 2008). ERP implementation is observed to be a catalyst for successful Lean implementation and Lean initiatives proved to be helpful in developing business processes for ERP (Powell et al, 2013). Nauhria, Wadhwa and Pandey (2009) recognize effective data management as a key for successful implementation of Lean Six Sigma. (The Six Sigma methodology is a quality tool that emphasizes reducing the number of errors in a process (McKenzie, 2009)). For succesful analysis of product defects and improvement of quality, reliable and accurate data is required. The views of authors of normative works on Lean, is that the management of information must be transparent and intuitive (Houy, 2005). For instance, Lean uses Kanban as a simple way to prevent a workplace from being overloaded. The number of Kanbans to be allowed at a work station is defined by the following equation (Villa & Watanabe, 1993):

$$No.Kanbans = \frac{(production\ rate)*(lead\ time+time\ for\ safety)}{(number\ of\ parts\ carried\ by\ a\ pallet)}$$

It is not difficult to see that in case of high-variability production rates (typical for high-variability products), this equation becomes rather unstable and therefore useless. The number of required Kanbans would continuously change. Lean, through its philosophy, lacks the ability to manage a high-variability environment and related information complexity. If an organization wishes to implement Lean in such an environment, ERP is an essential system to manage the large quantities of information.

2.4 Push vs. pull and hybrid approaches

We have seen that MRP and Lean, or push and pull in general could go together. Villa and Watanabe (1993) claim they are in fact complementary and several researches have shown that hybrid approaches outperform pure push or pull strategies. Wang and Xu (1997) for example, simulate a 45-stage production system and compare several (hybrid) control strategies. They present four manufacturing systems; (i) single-material, serial process (ii) multi-material, serial process (iii) multi-part process and assembly system (iv) multi-part multi-component process and assembly system. The results indicate that the recommended hybrid push-pull strategy results in lower average cost. (It should be noted that the experimental set-up simulates a demand for just one product.)

2.4.1 Typology

We use the distinction made by Corry and Kozan (2004) to make a typology for push-pull hybrid approaches. At the end of this section we conclude whether or not the approaches treated here are applicable for Company X. Corry and Kozan (2004) separate push-pull integrations in vertical and horizontal ones. Vertical being a different policy at different levels of planning, i.e., MRP plant-wide and Kanban between work centers. Horizontally applies push for some work centers and pull for the remaining. Gerathy and Heavey (2005) call them vertically integrated hybrid systems (VIHS) and horizontally integrated hybrid systems (HIHS). HIHS are more widely applied, of which Wang and Xu (1997) is an example.

Olhager and Östlund (1990) distinguish horizontal push-pull integration based on three different concepts: (i) CODP (ii) bottleneck resource and (iii) product structure. CODP is the point at which a product becomes customer specific. In other words, it is the point at which the value added to the product is customer specific. From this point of view, it is a logic boundary between push and pull. Every process step before the CODP is pushed through the system based on forecast, these are usually generic parts and subassemblies. Because the value is added mostly after the CODP, production afterwards is demand driven (pulled).

The next concept Olhager and Östlund (1990) discuss is the bottleneck resource boundary which is also discussed by Villa and Watanabe (1993). A stable bottleneck (meaning it does not move from resource to resource) usually means that the lead time after it is stable as well, therefore this part of the process can be treated as a single planning unit and a standard lead time can be assumed. The implicit assumption being that there is sufficient overcapacity at the remaining work stations, this is reasonable since we established that the preceding work station was the capacity restricted bottleneck. The processes succeeding the bottleneck are controlled using a push strategy. The processes preceding the bottleneck should serve it with parts making sure that it keeps busy, implying a pull strategy. Some WIP inventory might be required at the boundaries between the strategies, or safety time in case of customer specific parts. The authors note that it becomes

considerably more complex when there are multiple or moving bottlenecks (Olhager & Östlund, 1990).

The third and final concept Olhager and Östlund (1990) mention is a push-pull boundary based on the product structure. Products with more complex BOMs usually have a distinguishable path that is more critical than others in terms of lead time. This concept states that the critical path components are pushed through the organization up until the point of convergence, i.e., assembly. Meanwhile the required uncritical components needed are pulled along. The capacity of the critical path serves as the restrictive factor.

The horizontal push-pull integrations all have some fundamental problem for application at Company X. Integration based on CODP and product structure seems to be less promising because all products are customer specific. The CODP is in the engineering phase. There are very few, if any, common parts that could be controlled with a push strategy. A push-pull boundary based on the complex BOM products also seems unlikely. For Company X the product mix changes rapidly. This means that we would continuously be redefining which BOMs are categorized as complex and which components determine critical paths to be pushed and which components to pull along. A horizontal push-pull integration based on a bottleneck is an interesting approach when a bottleneck resource will not change. We cannot state with certainty that a bottleneck at Company X will remain the bottleneck over time. For instance, if the product mix changes, capacity requirement for another processing step could cause a shift.

The VIHS is only limitedly applied in practice. Karmarkar (1986) introduced a hybrid approach based on the vertical integration. MRP controls the overall production facility, where component requirements are calculated based on the MPS. Production cells use Kanban as control strategy, the number of Kanbans is based on the MRP information. Furthermore, because MRP works with off-set lead times based on averages lead time, they tend to be inflated. Therefore cell supervisors get control over the number of active Kanban cards on the floor, releasing them only when appropriate. Karmarkar (1986) notes that this system is most appropriate for repetitive batch manufacturing and not appropriate for a customer order oriented system. Kanban cards are not suitable to identify and allocate parts to a specific order. This VIHS is therefore not suitable for Company X. A promising vertically integrated approach is called POLCA, it is designed specifically for manufacturers of a highly variable product mix with small batches. It is introduced in the next section.

2.4.2 POLCA

Paired-cell Overlapping Loops of Cards with Authorization (POLCA) is a method to cope with the highly variable demand for capacity due to a large product mix (Krishnamurthy & Suri, 2009). It is a vertically integrated hybrid system with a push-pull approach. The push part refers to the MRP scheduled product-specific authorization, pull refers to the capacity driven authorization. Authorization meaning when a production order is released for manufacturing. The pull principle uses *polca* cards, somewhat similar to the KANBAN system. The KANBAN signals the requirement for a particular product downstream and upstream production of this part is triggered. It is a product-specific material control system. The *polca* card signals the requirement of *some* product. It is a signal that capacity is available downstream. This is defined as a product-anonymous material control system. A specific kind of product-anonymous material control is called route-specific (Riezebos, 2010). Which product should be produced upstream all depends upon which downstream department signals available capacity. Only products which have that department as the next step in the routing are considered, hence the term route-specific. The intention is to start processing a production order only if we know with certainty that once finished, there is capacity available at the

next work station. We let the department pull work its way when it can handle it, thereby avoiding unnecessary waiting time.

The concept of signaling capacity availability and triggering the right production order to be processed is controlled using capacity control loops. Within a loop between two production cells, a number of *polca* cards circulate. If and only if a *polca* card is available to accompany a production order is it allowed to enter the queue of the department. These control loops are actually based on queuing theory, limiting the amount of WIP to make sure that the lead time does not increase.

Each production cell is in at least two loops, one with its immediate predecessor and one with its immediate successor (Figure 2.4). The first and final processing step in a routing are the exception. When a production order is supposed to be processed in cell *A*, then *B*, it requires a $polca_{AB}$ card (Figure 2.5). An absence of this card at cell *A* tells us that cell *B* does not have available capacity in the near future and the order remains in queue Q_{AB}. As soon as a production order leaves production cell *B*, its $polca_{AB}$ is removed and transported back to cell *A* to signal available capacity. Because the lead time is waiting time and processing time, we only send it back after the processing is finished. Any order authorized to be processed by a production cell within two loops, requires two *polca* cards. Production orders waiting in Q_{BC} for instance, still have $polca_{AB}$ and also require $polca_{BC}$.

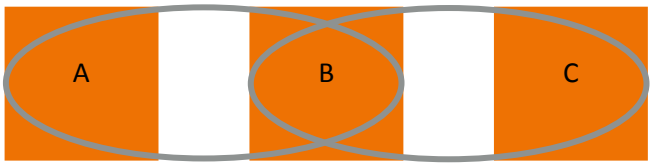


Figure 2.4: Example of paired production cells

Intermediate authorizations can be incorporated as well, Vandaele *et al* (2008) specifically mention that assembly systems can benefit from this feature. The dynamic and changing situation which occurs when multiple components are required to arrive at the assembling cell in synchronicity, can benefit from extra control.

The objective of POLCA is balancing the workload of each production cell. To avoid overloading with WIP and as a result an increased lead time. The amount of cards $N_{A/B}$ in circulation is calculated based on a modification of Little's Law (Krishnamurthy & Suri, 2009):

$$N_{A/B} = [L_A + L_B] \cdot [\frac{NUM_{A,B}}{D}]$$

Where L_A and L_B are the estimated average lead time during planning period *D*, *W* in Little's Law. $NUM_{A,B}$ is the total number of jobs which go from cell A to B during this same period, λ in Little's Law. The total number of cards in the loop is the total number of customers in the system *L* in Little's Law. The basis of the POLCA material control system is therefore queuing theory. By limiting the number of customers in the system with the use of *polca* cards, we implicitly limit the utilization of a production cell. Utilization between 30 and 80% performs best according to Riezebos (Polca scanningtool). Lower than 30% leads to underutilization, while more than 80% leads to congestion.

Table 2.2 shows an overview of different case studies and the reduction in lead time which is achieved. Based on these results, we expect that the concept of POLCA can definitely contribute to the objective of Company X to reduce internal lead time.

2.5 Structured planning

Every manufacturer offers a product to its customer and in order to deliver, it relies on resources with a limited capacity. Furthermore, it is likely to offer multiple products to multiple customers simultaneously. A complex flow of information and materials originates. The production planning and control function of an organization is burdened with the task of managing this flow, looking for a balance between timely delivery and cost efficient use of the required resources. The objective would be a planning procedure which enables this balanced work flow, in a such a way that the products are delivered on time at a profitable price, while the process of planning itself is efficient as well. Ultimately, the objective for planning and control is to contribute to the main business objective to achieve long-term profitability. This is a large, complex problem and to structure this a hierarchical planning framework can be applied.

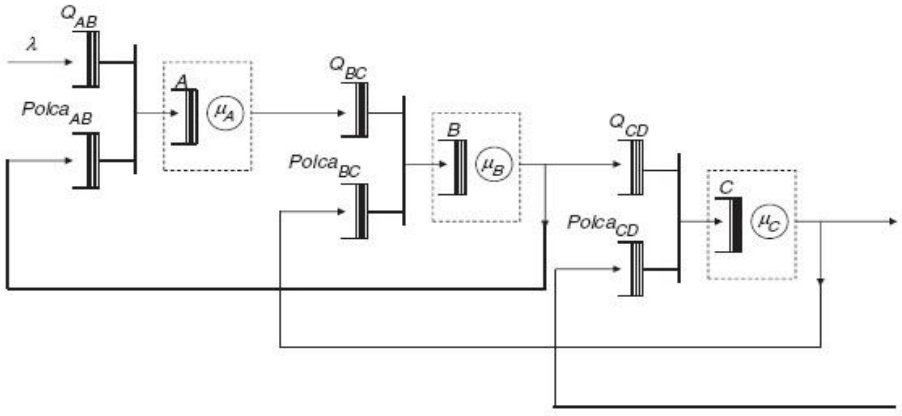


Figure 2.5: Example of POLCA control loops (Riezebos, 2010)

Table 2.2: Lead time reduction based on case studies

Case study	Lead time reduction	Author(s)
Bosch Hinges	70%	(Riezebos,2010)
Spicer	50%-60%	(Vandaele et al, 2008)
Olsen Engineering	22%-68%	(Krishnamurthy & Suri, 2009)
Rockwell Automation	25%	(Krishnamurthy & Suri, 2009)

Perhaps the best description of a hierarchical planning approach is dividing the large, complex planning problem into less complex subproblems. These subproblems are then solved sequentially and the decisions made on higher levels form restrictions for subsequent levels (Gelders & Van Wassenhoven, 1982; Moscoso, Fransoo, & Fischer, 2010). As we move down the levels, we generally decrease the time we take into consideration (planning horizon) and increase the level of detailed information (disaggregation of data). This is a natural approach, as more detailed and accurate information becomes available as time goes along (Tenhiälä, 2011; Zijm, 1999). By decomposition, the complexity of the planning process is decreased, giving the opportunity to control the problem at the lowest possible level. This is aimed at increasing the possibility to react in a timely manner, using locally avaliable information as much as possible (Moscoso et al, 2010). At every planning level, we consider what a useful aggregation of data would be, i.e., the required level of detail. Several types of aggregation are used. (i) Products (aggregate forecasts for product type). (ii) Production stages (several sequential operations may be considered as an aggregate operation). (iii) Capacities (capacities of similar parallel machines may be summed). (iv) Time (the horizon is divided into time buckets) (Gelders & Van Wassenhoven, 1982). An easily understandable distinction

between the various levels is strategic, tactical and operational (Hans, Herroelen, Leus, & Wullink, 2007; Liberatore & Miller, 1985).

2.5.1 Strategic

At a strategic level, the planning horizon is long term (usually a few years). This level includes decision related to the market, such as what product mix to manufacture. It also includes long term investments, such as the location of manufacturing facilities and the purchase of machines and hiring employees (Hans, Resource loading by Branch-and-Price Techniques (ch.1 of Ph.D. thesis), 2001). Furthermore decisions related to the logistic concept and manufacturing controls, i.e., make-to-stock, dedicated production lines etc.

2.5.2 Tactical

At this level, decisions are made regarding the MPS given the available resources (the strategic decisions impose restrictions), thereby creating a balanced workload. The tactical level has a medium term planning horizon usually up to a year, but the exact planning horizon and time buckets should be chosen to match accurate and useful (aggregated) data. If accurate and useful data about forecasts does not extend six months, efforts to control beyond this point would be a waste. Furthermore, detailed information about actual processing times is generally not available. Therefore detailed daily planning would also be a waste of effort. Planning in time buckets of a week would be more appropriate. This process, often referred to as Rough Cut Capacity Planning (RCCP) (Hans et al, 2007; Jonsson & Matsson, 2002; Tenhiälä, 2011), aims at minimizing the total throughput time of all orders given the available resources and respecting the delivery dates. The process includes order intake as well. Many organizations have a tendency to accept as many orders as they can possibly acquire. In the process, organizations promise a delivery date that is as early as possible, without sufficiently assessing the impact on resource capacity. This may lead to serious overload of resources, having a devastating effect on the delivery performance and the profitability of the production system as a whole (Hans et al, 2007). Through the process of an RCCP, we create a capacitated MPS; forecast and sales orders are input, required resources over (aggregated) work centers [departments] are considered, finalized due dates are given (Frank, Neumann, & Schwindt, 1997). Capacity is still somewhat flexible at this level, the planning horizon is long enough to detect shortages and temporarily increase it. A good RCCP leads to improved schedules at an operational level (Hans, Resource loading by Branch-and-Price Techniques (ch.1 of Ph.D. thesis), 2001).

2.5.3 Operational

At this stage we know which orders have been accepted and therefore must be manufactured before a certain due date (a capacitated MPS forms the constraints for the operational planning). Furthermore, at this stage we know precisely which manufacturing processing steps are required (BOM) and we have a more accurate estimation of the required capacity (processing time). At this point the planning horizon is further decreased to short-term, e.g., a week. The challenge is to create a daily production schedule where all jobs are assigned to a work center, minimizing the total throughput time of all jobs within the planning horizon.

The ‘higher’ planning levels require less detailed (product) information as opposed to the ‘lower’, more sophisticated methods. The availability, accuracy or necessity of the level of data determines what a suitable planning method is (Tenhiälä, 2011). This means that not every manufacturing organization should control its production at the same levels by applying the same methods. For instance, for an organization with a low level of product variation and high degree of repetition in production processes, detailed job scheduling could be unnecessary.

2.5.4 Integration

Moscoso et al (2010) are critical towards hierarchical, decentralized (insufficiently integrated) planning, stating it causes instability in the production plans. The causes of the instability, MRP nervousness and the lead time syndrome (see section 2.2.2). They have combined the two under the denominator 'planning bullwhip', after its better-known sibling in supply chain management. The supply chain bullwhip effect is an increasing level of variability in demand, going upstream in the supply chain. The effect is caused by decentralized demand information and optimization of subproblems rather than taking end-user demand into account. This causes suboptimization in the supply chain, because each level considers a different scope. While the supply chain bullwhip propagates horizontally along the supply chain, the effects of the planning bullwhip propagate vertically along planning hierarchies, both top-down as well as bottom-up, (sub)optimizing a planning subproblem. For example, an ETO/MTO organization must adhere to the delivery dates previously agreed upon. This could mean smaller lot sizes and quicker product change-over than desired from a strictly machine efficiency point-of-view. However, detailed scheduling based on machine efficiency at the cost of missed delivery dates would be highly undesirable and damaging to the continuity of the organization.

As the example demonstrates, consistency between objectives is a problem between planning levels because objectives potentially conflict. It is not enough to have a good aggregate planning procedure and a good detailed planning procedure, they should be carefully integrated. This integration enforces consistency in the prioritization of objectives when decisions are made at different levels (Gelders & Van Wassenhoven, 1982). Hierarchical planning activities must be integrated to ensure coordination between production levels. Decisions must be consistent to avoid sub optimization (Liberatore & Miller, 1985). When making a trade-off, one should keep the objectives of earlier decisions in mind.

2.6 Summary

For an ETO/MTO organization like Company X, both a pure push (MRP) or pull (Lean) approach causes problems. MRP is too much of a deterministic, inflexible approach. The resulting production schedules are likely to be infeasible because the reality is too variable. At the same time, the flexibility that a successful ETO/MTO organization should inherently have, i.e., consider alternative solutions such as outsourcing and overwork, are not supported in MRP. On the other hand, Lean lacks the ability to manage the highly variable product and demand behavior and complex information associated with it. This is one of the strong points of MRP. MRP and Lean strategies have conflicting objectives, making it difficult to combine them. MRP is a centralized control system where machine efficiency is a leading objective. Independency of production problems in case of product defects or machine break-down, "zero-risk" through enough WIP. Lean is a decentralized control system where WIP and lead time reduction is a leading objective. "Zero-inventory" to uncover production problems and inefficiencies to be solved. A promising alternative for a vertically integrated approach is POLCA, as it is designed to deal with the variability of ETO/MTO organizations. The method focuses on product-anonymous capacity availability. Several case studies show lead time reduction between 22 and 70%.

Production planning is a large and complex problem. Therefore we can divide it into more manageable sub problems and solve these sequentially. We distinguish a strategic, tactical and operational level. Moving down the levels, the planning horizon decreases while the level of detail increases. Higher levels require less detailed information as opposed to the lower, more sophisticated methods. The choice of planning method depends on the availability or necessity of the level of detail. Regardless of a push, pull or hybrid approach and which planning method is used,

integration is the most important to make it successful. This integration refers to the different levels and their objectives, as well as planning as a function. The function must be embedded within the organization and processes should be integrated with the other functionalities. To make the planning function easier to understand and create more support, we can focus on the level at which planning and an employee's activities intersect.

We can use queuing theory to predict lead time, explain why waiting time occurs and how it increases. In the next chapter we will apply this theory and try to understand the current performance of Company X.

3. Production process and performance analysis

This chapter answers the second research question, "*What does the production process look like and what is the current performance like?*" Firstly, in section 3.2 the production process is described in detail. Section 3.3 discusses the current performance overall and of the specific departments. In section 3.4 and 3.5 we try to understand which factors influence the lead time performance and how it might be controlled. Finally in section 3.6 we combine the literature review and analysis of the current situation and draw our conclusions.

The data that is used to analyze the current performance of Company X has been collected between January 2014 and March 2015. The data used for the analysis is collected by Company X itself. Processing times of production orders are registered using hand-held scanners connected to the ERP system. The clocked times are imported in a business intelligence package (Qlick View) and made available for this analysis.

3.1 Manufacturing system typology

A manufacturer can be described using a manufacturing system typology along two dimensions. A distinction can be made between different logistic product/market combinations and based on the internal structure of the organization (Zijm, 1999). In the case of Company X the product/market dimension is a combination of make-to-order (MTO) and engineer-to-order (ETO). MTO organizations produce based on demand of the customer, as opposed to production for inventory to be sold at a later date (make-to-stock (MTS)). Engineer-to-order is characterized by a manufacturing process, starting at the design phase. In such a case, the finished product is engineered according to the customer's (technical) requirements. Company X both co-engineers products with the customer and manufactures products for which the design is supplied by the customer.

The internal manufacturing structure is based on the *job shop*. Company X has several departments in which combinations of man and machine are able to perform a manufacturing process. The factory layout is functional, the equipment is arranged by function such that all similar operations are performed in one area of the plant (Carlsson, 1989), a so-called department. All products move through the factory and the predefined routing in the ERP system determines which processing steps are required and upon which machines they are performed. These departments are:

- Cutting (sheet metal center)
- Bending (pressing, punching, drilling)
- Welding
- Painting (powder coating)
- Assembling

(The first department is Engineering and Pre-production, but because these departments are not included in the scope of this thesis they are omitted here.)

3.2 Production process

The type of manufacturing orders flowing through the factory can be distinguished into two types; initial orders and repeat orders. Figure 3.6 shows the processing steps to complete an initial order. This process starts with engineering, this can be as simple as checking the technical feasibility of the customer's own design, making sure it can in fact be manufactured. It can also mean completely designing a product based on customer requirements. Furthermore, this step includes

devising the routing for the product and calculating the required processing time for each processing step. These are important input variables for a solid planning. If the product is designed and released by the engineering department, the products move on to the programming department of the sheet metal center. For the initial order, the parts are programmed to be cut out of sheet metal. For first time repeat orders a new concept has been introduced: the programmers perform a process which is called “nesting” of the product (*Figure 3.7)*. The final product is made out of different components, possibly from differing materials. All components of the same raw material are combined into a nest in a way to use the material as efficiently as possible, i.e., with minimum waste of raw material. The nests contain the amount of components required to produce an amount of finished products that fits within the standard shipping container of the particular product. The idea here is that a customer can order a multiple of the shipping quantity, and the cutting department simply cuts the required amount of nests to build the final number. It should be noted that this does not always work out as planned in practice. Some customers, especially some of the bigger companies, are not always willing to comply with this set-up. The distinction between initial orders and repeat orders is based on the fact that nesting requires an extra process for the programmers. For a first and presumed last time production, this extra effort is not worth it. Furthermore, not all products are transformed into nests just yet. If the order is a repeat order, the engineering and programming steps are omitted (with the exception of a first time repeat where a nest could be required). The rest of the processing steps remain the same.

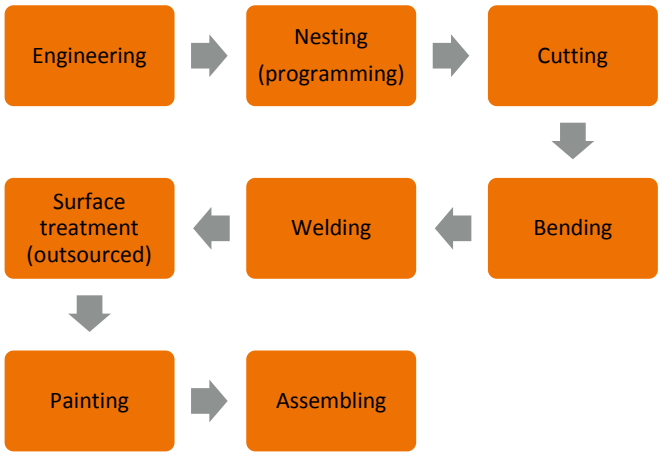


Figure 3.6: Schematic of all processing steps related to manufacturing

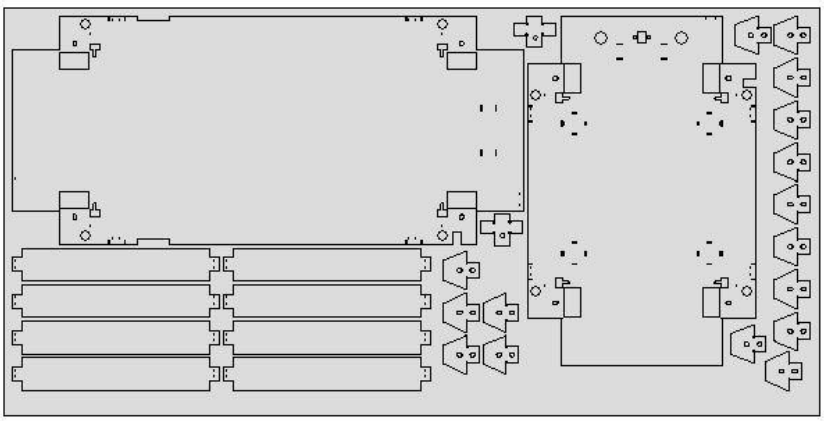


Figure 3.7: Example of a nest

3.2.1 Cutting department

At the cutting department, all parts required to manufacture a finished product are cut out of sheet metal, with the exception of purchase parts. The department uses a CNC laser cutter. The capacity for this department is mainly limited by the machine. The department works two shifts a day, six days a week. The machine can work autonomously, it is combined with a storage facility equipped with an automatic storage and retrieval system for raw materials and parts which have been cut. The cutting department is the only autonomous system, able to function without the (constant) presence of an operator.

3.2.2 Bending department

After being cut from the sheet metal, the parts go to the bending department where an individual order might follow different routings through various processing steps. These routings may include any of the following steps: bending, drilling, punching, deburring. In a typical routing, the sheet metal part is bent into a certain shape (a profile) using a press brake. A punch press provides any necessary recesses and a drill is used to create drill holes for assembly. Afterwards, the parts are deburred to get rid of sharp edges and such.

The department uses robot press brakes which operate autonomously. The department has manual press brakes as well, which are man operated, and several smaller pieces of equipment such as punch presses and drills, all of which are manually operated. 12 people are employed at this department, the capacity of this department is mainly limited by manpower.

3.2.3 Welding department

The welding department is the first converging process step, where multiple parts are combined into a subassembly. The department is divided into two separate ones, welding and stainless steel welding. This is to avoid contamination of the stainless steel with small metal particles. The department employs 18 people and has one welding robot. Capacity is limited by manpower. The welding robot is only useful when there are larger batches of relatively simple products, i.e., no corners which are hard to reach for the robot. Furthermore, it requires the presence of an operator and must be programmed.

3.2.4 Painting department

The painting department has one paint booth where products are powder coated. Before the products can be coated, the welds first need to be sanded to create a smooth surface. The department employs seven people. At this department the booth is the limiting factor on capacity. After the painting process is finished, extra manual labor can be beneficial though. To finish the process the painted parts need to dry hanging. This process can be sped up with more manpower, thereby freeing up the paint booth for a new product. Seven people are permanently employed at the department.

3.2.5 Assembling department

The assembling department is the second converging step in the process, where all subassemblies are assembled into a final product. This department is limited in capacity by manpower. It employs 10 people.

3.3 Current performance

The average lead time of a sales order is X workdays with a coefficient of variation of 0.3. This average is based on X sales orders in the period between week 2 of 2014 and the week 12 of 2015. The average weekly number of sales orders is X, again with a coefficient of variation of 0.3. Figure X shows a graphical representation of the average weekly sales order lead time, along with the number

of corresponding sales orders. It clearly shows a decline in the lead time, but also shows that the number of orders declines as well. In general lead times tend to increase if the number of orders increases. Assuming the decrease in lead times is an effect of improvement initiatives could be a false inference based on this representation. Excluding the outsourcing process (this takes around 7 work days in general), the average internal lead time is 13.8 workdays. There is a large variability in average lead times and the number of sales orders.

Over the period January 2014 until March 2015, on average #1XXX production orders have been processed in a week factory wide. This required an average of X hours (Figure X). During this period, on average X hours of actual attendance have been clocked. This means an overall efficiency of XX%. Week 52 (the Christmas Holiday period) is omitted in this calculation, as it is a clear outlier in the data. The number of processed orders for instance, is three standard deviations below average.

Table 3.3 shows the average lead time and waiting time for a sales order and also divided into the different production departments. The lead time is defined as the time lapse between the first moment of order availability, i.e., the production order is WIP in queue, and the moment the production order is registered as completed. The waiting time is defined as the time lapse between the first moment of order availability and the time actual work on the production order is started. The total production lead time of all departments is 13.8 workdays on average with a coefficient of variation of 0.5. It should be noted that the average sales order lead time of X workdays is Y workdays more than the combined average lead time of the departments. This discrepancy may have any of the following causes:

- Outsourcing of surface treatment, about seven workdays in general;
- Total lead time is based on sales orders, possibly with multiple delivery dates which could obscure the sales order lead time data;
- Quality control has not been included in the analysis of department lead times;
- During the manufacturing process, a production order could potentially have to wait on a purchase part.

Table 3.3: Summary of average lead time and waiting time (in days) for a sales order and divided into production orders (departments)

	Lead time μ	Lead time σ	Lead time $\frac{\sigma}{\mu}$	Waiting time μ	Waiting time σ	Waiting time $\frac{\sigma}{\mu}$	$\frac{waiting\ time}{lead\ time}$
Sales order	*X*		0.3				
Cutting	3.0	2.4	0.8	1.9	2.3	1.2	63%
Bending	3.0	2.7	0.9	1.9	2.7	1.4	63%
Welding	2.9	3.1	1.1	1.6	2.8	1.8	55%
Painting	2.2	1.7	0.8	1.0	1.6	1.6	45%
Assembling	2.7	4.4	1.6	1.4	3.9	2.8	52%
Total	13.8	6.7	0.5	7.8	6.2	0.8	57%

Table 3.4: Average processing time (minutes) per order, expected (VoCa) and actual (NaCa)

	μ expected	μ actual	$\frac{actual}{expected}$
Cutting	42	35	83%
Bending	78	74	95%
Welding	148	156	105%
Painting	30	23	77%
Assembling	157	150	96%

If Company X wishes to be successful in reducing the lead time, the waiting time is probably the most important to address Table 3.3 shows that on average, 7.8 workdays is waiting time, about 57%. The actual production time (lead time excluding the waiting time) is probably less likely to be reduced. As Table 3.4 shows, the processing times are fairly accurately estimated. Experience learns us that when the percentage of initial work increases, the odds of mismatches between estimated and actual processing times increases. Summed over all departments, the average total production time for a finished product is just over 7.5 hours. Including some extra time for the paint job to dry, in theory a finished product could be finished within two days. However, this implicitly assumes that the succeeding production department is idle and able to process the production order immediately (no WIP). This is an unrealistic assumption in practice and for this reason the offset lead time parameter for the MRP system is deliberately chosen to be one day. This means that each department's processes are scheduled on a separate day, which includes some safety time for each of them. Based on this MRP characteristic, the expectation is that reducing the effective production time is not possible without omitting the one day offset time, i.e., the safety time. Omitting it would result in a need for such critically hour-to-hour production schedules that the feasibility is most likely difficult to ensure. Moreover, the regulatory pressure this entails for a planning function would be very time consuming and costly. We may be able to achieve considerable positive effect on lead time by clustering different processing steps into one and thereby cutting out multiple days of off-set lead time. In fact, reducing total lead time to 5 days might be more dependent upon reducing this number of unique processing steps. This is however, neither a responsibility of the planning function, nor does it help controlling the MRP schedule.

From a planning point of view, the most effective course of action would be to reduce the waiting times before the production orders are processed by the departments. Section 3.4 attempts to understand how the lead time and waiting time could possibly be reduced, creating a better work flow through the factory.

Table 3.3 shows that the waiting time for the cutting and bending departments account for 63% of the total lead time. Based on this the conclusion could be drawn that these departments are a bottleneck. Taking into consideration that the cutting department is the first department in the production process, an uncapacitated workload and lack of clear prioritization could explain why the waiting times are as high as they are. If production orders are scheduled and released without capacity limitation, the repercussions would be most apparent at the first processing step. While on average the waiting time is the longest, the difference is small. Experience learns us that the quickly shifting product-mix can cause a shift in workload as well. This uncertainty causes a shifting bottleneck. In our literature research we already encountered this. Olhager & Östlund (1990) concluded that solution approaches involving a bottleneck become complex when there are multiple (cutting and bending) or moving bottlenecks.

3.4 Lead and waiting time explained

3.4.1 Explaining lead time quantitatively

From queuing theory the parameters variance (in arrival and processing rate), utilization ρ and processing time $1/\mu$ are known as explanatory variables for waiting time. In this section this hypothesis is tested on the data collected at Company X (Figure 3.8).

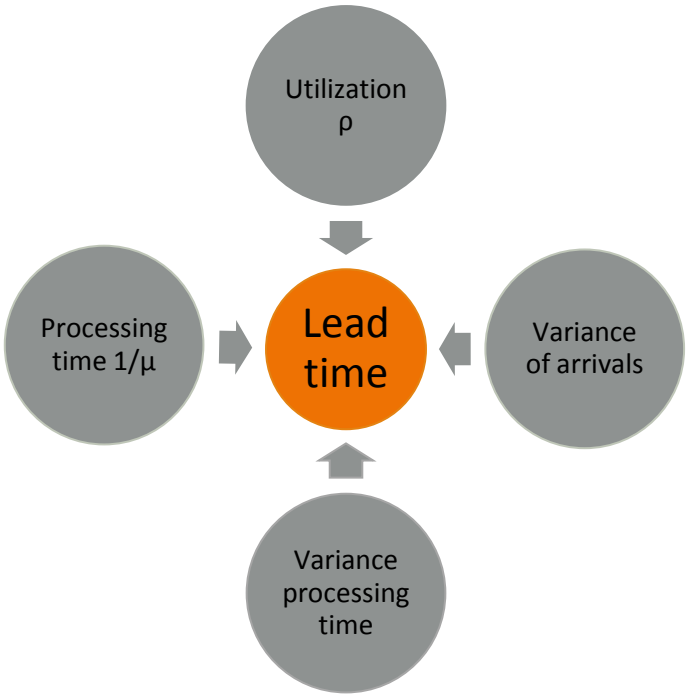


Figure 3.8: Explanatory variables affecting the waiting time

For Company X the arrival rate λ can be expressed as the daily number of orders becoming available to start, the process rate μ can be expressed as the daily number of finished orders. The utilization or traffic intensity $\rho = \frac{\lambda}{\mu}$ can therefore be expressed as $\frac{daily\ orders\ available\ to\ start}{daily\ orders\ finished}$. The explanatory variable variance in the arrival process is calculated as daily arrival rate λ compared to the average over the data set. The explanatory variable variance in process rate is calculated as the daily orders finished compared to the average over the data set.

N.B. because ρ is a ratio around 1, the other variables are relatively large numbers. Consequently their coefficients in the linear regression model would be very small, implying the variable is of little influence. To avoid any confusion, the number of products and processing times is rescaled using the natural logarithmic function.

This explanation for lead time is tested using a multiple linear regression analysis. While these variables seem to have an effect, the result unfortunately does not indicate a very strong relationship, R^2 is 0.44. According to this number, 56% of the variation cannot be explained by the included variables. Therefore it does not provide definitive proof to exclude that increasing lead time might be coincidental or (partly) dependent upon different factors. We cannot conclude that the variables from queuing theory can give us a satisfactory explanation of the situation at Company X.

Four things should be kept in mind while over thinking the result, which provide some insight into the remaining, unexplained variance in the lead time:

- The appropriate data has only been registered since the end of 2014. The first recorded data was during the Christmas period and did not represent the normal activity. Therefore it has been omitted. As such, this analysis is based on a period of just 57 days.
- This data has been registered during a period where the demand has been relatively low. Therefore it is possible that all registered data on waiting time is skewed because the pressure is off. Because some processes are performed with manual labor, output can be

increased when more work is available, quite simply people work harder is such a situation. To confirm the theories 1 and 2, the analysis should be carried out again at a later date when there is more data available and if possible during a period where order intake is higher. (A preliminary analyses based on a second, larger data set revealed even less explanation though.)

- The remaining variance may be (partly) due to the flexibility in capacity. When the workload increases extremely, extra capacity is acquired through the employees with the ability to support another department. This too would suppress the waiting times. The flexibility is not planned in advance and therefore useful data is not easily obtained in hindsight, meaning that it is difficult to incorporate it as an explanatory variable in the regression analysis.
- The final and perhaps most important aspect of the current situation to keep in mind is the fact that robust, feasible production schedules are virtually non-existent at the moment. Available capacity is not considered for order intake. Company X is able to manage to deliver its products only through the use of rush orders for customers it deems as most important and adding an extra shift without formally registering this extra capacity. Such trickery may well help to solve the short term problems, but it obstructs and is very harmful to an organized production.

The multiple regression analysis gives the following equation to predict the daily lead time per department and considers the effect on lead time if one of the variables is changed:

$$Y = 0.1X_1 + 1.1X_2 + 0.6X_3 - 0.4X_4 + 2$$

Where

Y = Lead time

$X_1 = \rho = \frac{\lambda}{\mu} = \frac{daily\ orders\ avallable\ to\ start}{daily\ orders\ finished}$

X_2 = Deviation of daily λ from average arrival rate λ

X_3 = Average processing time $1/\mu$

X_4 = Deviation of daily processing rate μ from average processing rate μ

The fact that the coefficient of X_4, the variance in processing rate, is negative means it has a positive effect on lead time. This contradicts with queuing theory, most likely because of the four reasons mentioned before. The result of the regression analysis is an indication that the theoretical explanation of increasing lead times is not suitable for the practical situation at Company X.

3.4.2 Explaining waiting times qualitatively

Because a completely satisfying quantitative explanation for the (increase) in lead time is not found, this section is meant to explain the variance which has not been explained in a qualitative way. Therefore two fictitious products A and B (Figure 3.9) are scheduled to illustrate the occurrence of waiting times and help to build an understanding of the problem. Product A is a “complex” product, with three parts, all of them requiring CNC laser cutting, two of them requiring drilling and one of them requiring bending. Afterwards the three parts are welded together. Product B is a “simple” product. A single part, requiring CNC laser cutting and bending.

The finished products are both required to be delivered at time t = 0. The structures in Figure 3.10 show how the offset lead time in MRP schedules the various underlying processing steps according to the required start date of their parent. At time t = -1, two different production orders

are to be completed by the cutting department. Suppose the capacity is not enough to complete both orders and product A has an assumed priority over product B. In this example the waiting time for the cutting department increases (presumably with a day if capacity is sufficient on the next day) and the total lead time of product B increases as well. Unless one of the other departments has excess capacity that could be interchanged to meet the required capacity. In that case, waiting time would be predicted for the cutting department, yet does not occur. (In reality the limiting factor for the cutting department is machine capacity, therefore flexible capacity would not apply.)

Imagine this process to entail a multitude of finished products with varying delivery dates and component structures many more levels deep. Where lack of management, planning and control leads to:

- Maximum capacity levels per department which are not strictly guarded;
- Flexible capacity interchanges which are not formally controlled;
- Order priority which is not based on actual customer importance, frequently changed and sometimes not adhered to in production.

In such a situation, the frequent occurrence of waiting time is understandable, but difficult to predict because there is no explanatory variable accounting for the effect of lack of management, planning and control.

Based on the quantitative analysis the conclusion could be drawn that Kingman's formula does not apply for Company X, although it is known to be quite accurate in predictions. Based on qualitative (intuitive) analysis, on the other hand, it could be argued that the fact that there is lack of control any relationship that does in fact exist is simply obscured. The question that arises: can we control the situation or do we need to accept the situation as it is and look for other approaches?

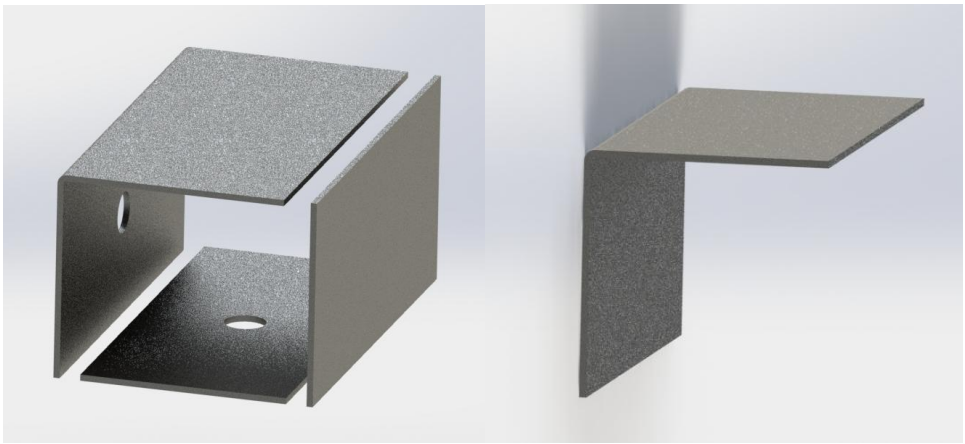


Figure 3.9: Example product A (left) and B (right)

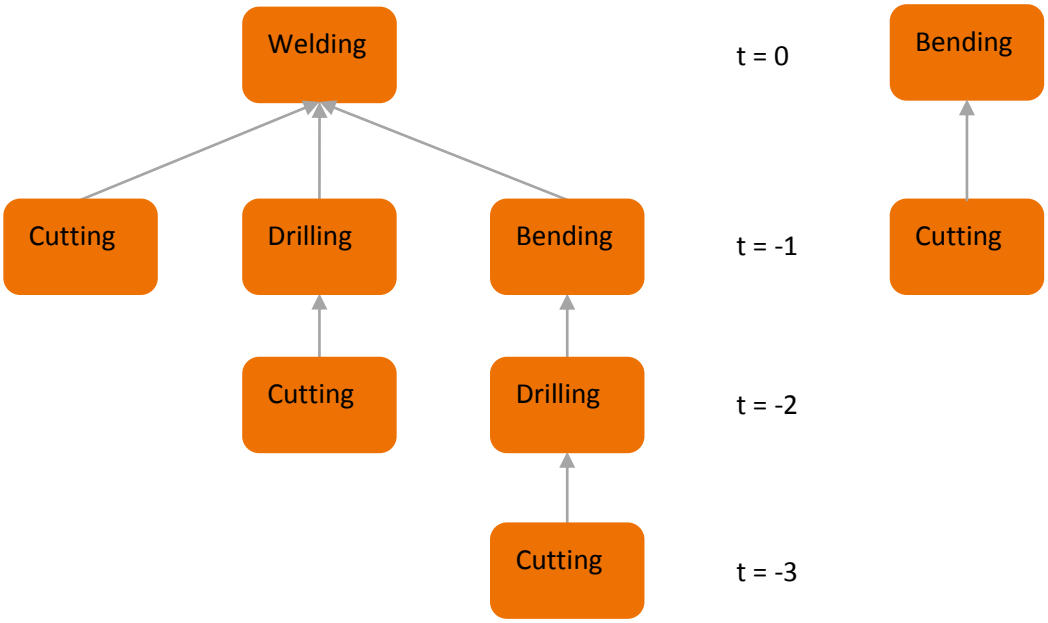


Figure 3.10: Example of two product structures and schedules to explain occurrence of waiting time

3.5 Lead time and waiting time controlled

Waiting and lead time can theoretically be explained by arrival intensity (λ), processing rate (μ) and the variation related to them. If this is the case, we can subsequently control waiting and lead time as well by manipulating these variables. Limiting the arrival rate and increasing the output, or a combination of both. The assumption is that these variables λ and μ are known and constant. The time between arrivals might be subject to change, but from day to day the same number of orders are required to be processed based on some known or predictable demand. Same goes for processing time. Between the different production orders to be processed, there might be differences. Some require more processing time than others. From day to day though, the same constant product mix has to be processed. This is where the practical problem of Company X occurs.

For Company X, the arrival rate and the product mix (processing time) changes from day to day. This means that applying Kingman's formula predicts a changing lead time from day to day, where we want to achieve a continuous flow. We could either try to balance λ and μ on a daily basis or take a longer period in consideration and hope for an even arrival rate and product mix. Considering a longer period is a problem for two reasons:

1. Customers place their orders on a short term, as such demand cannot be observed very far in the future. Even if arrival rate and processing time would be more consistent over time, we would never know for sure;
2. The same short delivery times required by the market force us to be able to react quickly. We must balance on a daily basis.

What we are left with is balancing from day to day, by adjusting arrival and output rate to find a daily balance. Here we encounter new difficulties.

3.5.1 Arrival rate

We should make a distinction between companywide arrival rate (sales orders) and department specific arrival rate (production orders). The actual arrival intensity of sales orders and its variance is a parameter which is difficult to influence. Because Company X is an OEM supplier, its

position to negotiate longer lead times in order to spread out the arrival of orders is not strong. This does not mean that no effort should be made to try to spread out the customer deliveries and associated workload. The potentially limited effectiveness of the effort spent on influencing this part of the process should be recognized though.

As far as the production orders are concerned, there is some leeway in the arrival rate. When arrival rate exceeds output rate at a departmental level, we could reschedule production orders in order to control arrival rate and lead time. The decision to adjust arrival rate by rescheduling production orders, forces us to reconsider the production schedule for every production order associated with the particular final assembly. First of all to avoid accelerating the lead time for one production department and subsequently lose this advantage if the synchronized availability for final assembly is no longer true. Secondly, we must also consider the situation where the problem is moved to a different department. Lastly, if the decision involves pushing a production order back there is always a possibility of delaying customer delivery, which should be carefully considered.

The decision process to choose which production order is best suitable for rescheduling requires information and expertise related to customer importance not readily available in ERP. Furthermore it depends on and changes with the dynamic demand situation. The difficult and changing parameters make the process unsuitable for generalization and automation in one tool.

3.5.2 Processing rate

On the other hand we could try to improve the processing rate when needed by applying extra capacity. Increasing the output temporarily by increasing capacity is certainly a possibility on departmental level by applying the flexible capacity when possible. This is what happens in practice at the moment, where decisions regarding interchanging capacity are made on the work floor.

If we want to centralize this decision process we have a number of difficulties. First of all we would need to identify when interchanging the capacity is helpful. Secondly, we would need a personnel matrix listing the skill set per employee, which is not available. Thirdly, we would need a detailed list of employees scheduled to work that particular day and figure out which employee can be assigned to resolve a shortage. The situation might change from day to day and scheduling this on a long planning horizon would be difficult. Furthermore the varying parameter settings or absence of input make central, automated control extremely difficult.

3.6 Theoretical landscape

We now have an analysis of the current situation and the results of the literature review in chapter 2. Combining both of these, we construct the landscape as shown in Figure 3.11. In this landscape, we have a horizontal axis representing variability in demand, products and processes. This ranges from deterministic to stochastic. On the vertical axis we put control, from pure push to pure pull and hybrid in the middle.

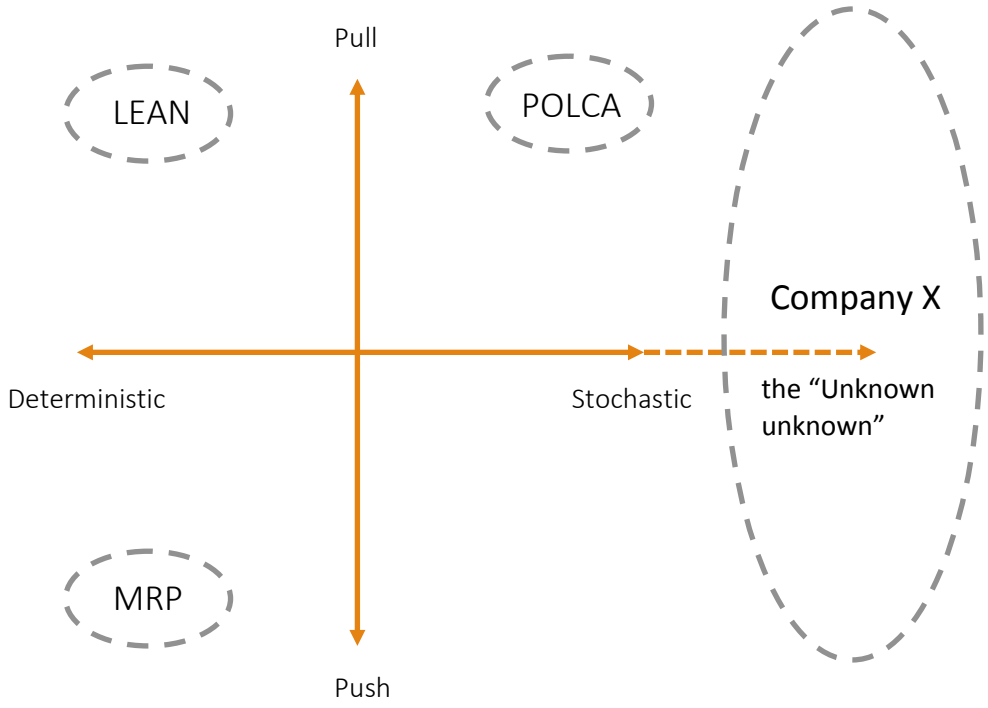


Figure 3.11: Position of the planning approaches from the literature review and the position of Company X

In the literature we identified both Lean and MRP as deterministic approaches. Not suitable for an organization with stochastic processes. We concluded that POLCA is better suited for organizations with uncertainty. It combines a pull approach[2] with product-anonymous capacity availability checks to promote an efficient order flow. POLCA still assumes an organization with its own products, implying stochastic processes within certain boundaries. This is where we identify an interesting gap between literature and practice. Company X operates somewhere in the environment of unknown unknowns. Not only is demand unknown, neither is the product which will be manufactured in the future. It requires a degree of flexibility which we have not found in existing literature. This forms the starting point for a potentially new, complementary approach to existing ones. An approach suitable for unknown unknown environments.

3.7 Summary

Company X is an ETO/MTO supplier for OEMs. Rather than manufacturing its own products. It manufactures customer specific products on order (demand driven). These products are manufactured in a job shop structured facility with five main departments:

1. Cutting (sheet metal center)
2. Bending (pressing, punching, drilling)
3. Welding
4. Painting (powder coating)
5. Assembling

Two types of production orders can be distinguished: initial orders and repeat orders. For an initial order, the product is engineered. This means a BOM and routing with processing steps are made and documented in ERP, along with calculated processing times. Subsequently, orders follow

[2] POLCA can be combined into a hybrid approach by applying a higher level push strategy, i.e., MRP. This is not the original intention though, therefore it is classified as a pull approach.

the routing along all required departments where the process steps are executed. For repeat orders the engineering phase has already been performed. In this case the production orders are released and follow the routing like the initial orders.

The average sales order lead time is X workdays at the moment and Company X produces about X sales orders per week. The average internal lead time is 13.8 workdays, with a standard deviation of 6.7. Of this total internal lead time, 57% is waiting time. Decreasing the waiting time is the best possibility to decrease the lead time, as this research is focused on planning and control rather than manufacturing technology (reducing processing time). The entire actual production time is 7.5 hours on average. Each processing step is scheduled on a separate day however, because it is unrealistic to assume the next department is idle and able to start processing right away. In MRP an explicit off-set lead time of one day is chosen. This forms the basis for another important opportunity to improve lead time. Perhaps several processing steps can be conveniently clustered, thereby eliminating several off-set lead times. This is however, neither a responsibility of the planning function, nor does it help controlling the MRP schedule.

Analysis of the lead time using queuing theory proves that we cannot explain or predict lead time for Company X. It seems that arrival and processing rate are subject to change in such a way that the underlying steady-state assumption is not met. Because we cannot give a satisfying quantitative explanation of lead time, we try to qualitatively explain what happens. We observe three phenomena. Capacity can be shifted around to solve shortages, maximum capacity levels are exceeded and order priority is frequently changed or not followed. Waiting time is clearly to be expected, but predicting when and where is difficult. We could try to create a steady-state by controlling input and output, and be able to predict and control lead time. Input and output must be known and constant to a degree. From day-to-day, roughly the same amount of orders to be processed and roughly the same product-mix requiring the same processing time. Because input and processing time change on a daily basis, we would need to rebalance on a daily basis by adding capacity or rescheduling. This is impossible to capture in an automated system because it depends on difficult trade-offs. Another option would be to consider a longer period and hope for an even order arrival and processing rate. Most customers place their orders on a short term, therefore we cannot observe demand on a long term, nor its processing requirements.

We construct a landscape in which we place the existing planning approaches from our literature review. In this same landscape we place Company X and we identify a gap between theory and practice. This gap is the starting point for the next chapter in which we look for a new approach for production control.

4. Workload and lead time control

This chapter answers the third research question "*how can Company X control and reduce the internal lead time?"* In review of the last three chapters, what is the key problem standing in the way for Company X to take the next step in reducing the lead time? Data analysis in chapter 3 shows that we cannot predict the lead time at Company X. The basic fact is, in order to control the lead time we should be looking for a balance between arriving production orders and the capacity to process them.

The theoretical explanation of lead time does not apply because the situation is continuously subject to change. Today we would be trying to balance the situation by adjusting the input λ and output μ, while tomorrow the situation is completely different and the process starts all over again. If this would be a process that could be automated and a computer could do the work quickly that would be acceptable. However, what we see in practice, is that the possibilities and considerations are too complex for automated solutions or at the very least would take a considerable amount of time and resources to implement.

What are the implications of this situation? First of all we must realize the simple fact that the arrival rate of production orders cannot exceed the processing rate or the factory will become blocked with WIP. Beyond this we must accept that the operational variability is not suited for a very rigid (MRP) schedule. Because the day to day situation is variable and unpredictable, a lot of effort spent on detailed scheduling using MRP is wasted. These schedules are frequently infeasible and require rescheduling. The situation is not suitable for a material control system based on a push principle, with a high level of variation.

From literature we know that vertically integrated push-pull systems were quite rare in actual application. However, a *vertically integrated hybrid system* (VIHS) is developed specifically for MTO and ETO organizations (see section 2.4). In the following sections we describe how it can be applied in some form, to help control the lead time for Company X.

4.1 Push-pull hybrid system

Suppose we accept that we cannot completely control the situation in order to control the internal lead time beforehand. What do we need in order to keep the lead time from dramatically increasing?

5. A method to roughly assess required versus available capacity;
6. A method to schedule the production orders;
7. A flexible material control system to achieve an efficient throughput;
8. A mechanism to guard the scheduled production order lead time between departments.

4.1.1 Push control

Company X's ERP system, Company Z, has a functionality which shows us what the workload is per department. This can be used to assess required production capacity versus available production capacity based on (desirable) sales order delivery date. When this is used properly, we are able to avoid the factory from being overloaded. Because there is inherent uncertainty in some parts of the input information, we call it workload balancing, high-level planning and scheduling. We limit the total capacity to be used to 80%, a higher utilization has a negative influence on lead time (Riezebos, Polca scanningtool).

The same ERP system can be used to schedule all production orders backwards in time, using an offset lead time. These first two planning methods are based on MRP's push principle. The question that follows: how can we make sure that this workload moves through the factory at a continuous rate? Or even with a reduced lead time? This is where we switch our point of view to pull.

4.1.2 Pull control

When a backwards MRP schedule is made, it is based on a predetermined off-set lead time. As long as we keep to this schedule, all production orders should move through the factory in synchronicity towards final assembly. However, as we have seen, the lead time is stochastic and depends on utilization. Therefore it would be unwise to release all production orders.

We schedule a workload per department which we know is (roughly) feasible within the particular off-set lead time. We can treat a department as a black box where we are not interested in the sequence in which the processing takes place. We only require its timely completion. Our objective now becomes guaranteeing a controlled internal lead time between departments. The challenge in this situation is to avoid insufficient synchronization between departments. This would result in waiting time between the completion of a process at one department and start at the following department (Riezebos, 2010). The approach we propose is derived from the POLCA method as developed by Rajan Suri in 1998 (Krishnamurthy & Suri, 2009) (see section 2.4).

4.2 POLCA adaptation

The POLCA material control system describes what we are looking for. A concept of multiple, routing-specific capacity loops which control the WIP in an operational setting, combined with an ERP system limiting WIP in a tactical setting. There are two challenges that remain if we want to model the system of Company X as a series of paired-cell overlapping loops:

- Complex routings;
- Calculating the number of authorization cards.

4.2.1 Complex routing

Complex routings are a result of processes with diverging and converging manufacturing steps. An example of a diverging routing is shown in Figure 4.12. In the set-up, cell *A* can produce anything as long as it has authorization of cell *B*. What we want to avoid is the situation where cell *B* only receives orders for either cell *C* or *D*. This could cause cell *C* or *D* to become idle if cell *B* has no orders to forward. Our proposition is to monitor the number of production orders in queue at cell *B*, when it signals capacity availability. Suppose cell *B* has just one production order left to forward to cell *C*. In a simple solution, we require at least one production order for each routing. In this example, the capacity signal from *B* to *A* can be accompanied by a signal that an *A-B-C* product is preferred, ensuring that the last order for cell *C* is replenished at cell *B*.

In a more elaborate control system, we can monitor the mix of production orders at cell *B*. The newly processed order at cell *A* should be the one which restores the balance in ratio *C:D* in queue *B*. The exact ratio should be based on the ratio of production orders cell *B* to *C* and cell *B* to *D*. Suppose that 50% of orders follows routing *A-B-C* and 50% follows *A-B-D*. We would release an order at cell *A* which restores this 50-50 mix. The advantage of this system, we guarantee the possibility of a sensible choice to process the right order at station *B*, because we guarantee a diverse mix to choose from. However, this requires additional data to determine the product mix. The data is not readily available and can also change over time. For the time being, the less complex method should suffice.

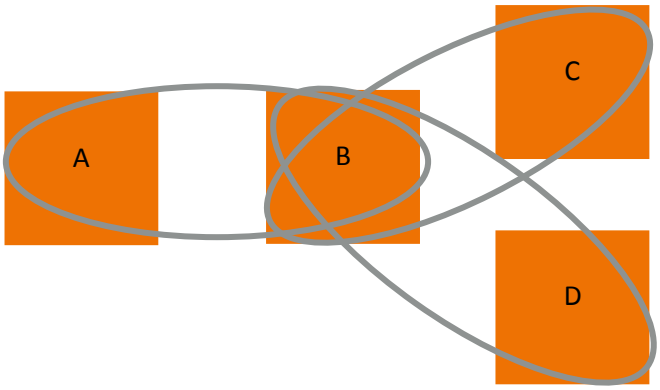


Figure 4.12: Example of diverging routings

Converging (assembly) steps are only mentioned casually by VanDaele *et al* (2008), who advice multiple intermediate authorizations. Riezebos (2010) argues that a converging step only requires all components to be available in the cell before processing can start. Based on the example in Figure 4.13, a job to be processed in cell *C* needs a card *AC*, *BC* and *CD*. In our opinion, this is no more than a material availability check. It does not help us to synchronize the moment at which components (*a)* and (*b)* are available. When cell *C* signals capacity availability, we still need a mechanism which triggers production at cell *A* and *B* of components (*a)* and (*b)* respectively for assembly in *C*.

Instead of complex interdepartmental communication and coordination to start production for the same assembly, we propose that this is where we profit from the strong point of MRP. The earliest start date of assembly in cell *C* dictates the earliest start date for components (*a)* and (*b)* based on their respective offset lead times (Figure 4.14). When offset lead time in ERP and the actual lead time are equal, synchronized availability is guaranteed. When the off-set lead time is not feasible, perhaps because the workload was higher than allowed, some production orders will not be finished on time. As a result, the MRP synchronization of simultaneous component availability is lost.

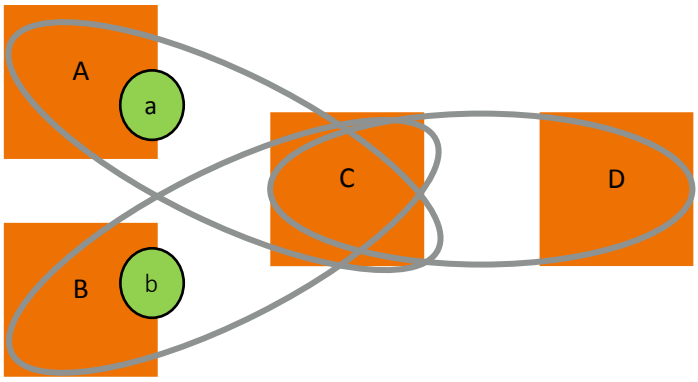


Figure 4.13: Example of converging routing

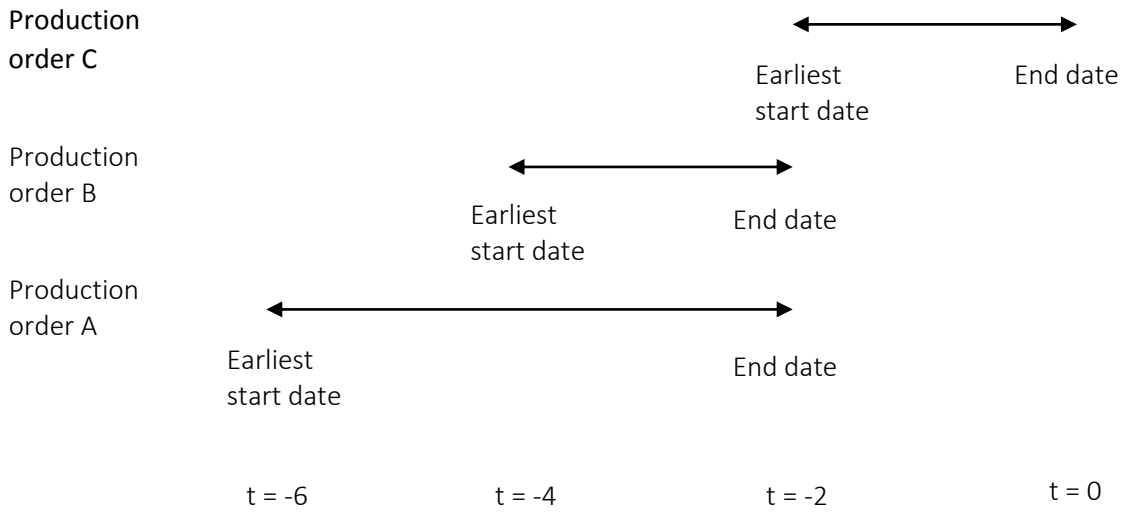


Figure 4.14: Example of MRP scheduled earliest start date authorizations

4.2.2 Real-time utilization

POLCA uses cards to signal capacity availability and trigger production upstream. The number of cards is calculated using a formula based on Little's Law. Several authors have attempted to further develop the method by improving the formula for the number of capacity cards (see Riezebos, 2010; Vandaele *et al*, 2008). Unfortunately, these improvements all still use Little's Law. All methods we have found in the literature calculating the number of production orders allowed in the production cell use averages and expectations. The combined average lead time for a loop of production cells and the expected number of production orders flowing between cells. This is more suitable for an MTO organization which (presumably) manufactures its own products, but in high-variety, low-volume setting. Our analysis in chapter 3 has proven that we cannot estimate the average lead time and the problem for Company X remains that the arrival rate can change from day to day. The expected lead time would be changing from day to day as well, requiring continuous updating of the number of cards in circulation.

The production orders in a production cell consist of the ones in queue plus the one(s) currently in progress. The combined expected processing time relative to the actual capacity of the department forms the real-time utilization. As a production order is finished, it leaves the system and vacates its processing time in the queue. This is a signal to a preceding production cell, to trigger production and avoid lack of WIP to continue to flow of production orders.

To signal available capacity without using cards, we propose to use an electronic authorization instead. Company X has television screens throughout the factory for all production departments. On the screens the available orders to start are currently shown. We propose a modified visualization of the production orders based on the two authorization checks. First of all, the MRP authorization. The computer system should sequence production orders based on increasing earliest release date. Second of all, the system reviews the queue of its successors to look for the utilization trigger. In essence, the *polca* card in an available capacity check, is replaced by a real-time electronic signal. The computer program currently running the system of television screens can be adapted for this purpose. It is believed that this will be a relatively simple adjustment[3].

[3] Based on the opinion of the system administrator of the television screens.

4.3 Approach

The approach we propose for Company X, is the adaptation of the POLCA method as we previously discussed. We distinguish two planning levels. Firstly, the tactical MRP schedule and secondly the operational level of overlapping capacity loops.

4.3.1 Tactical (MRP) planning

The MRP planning level is meant as a tactical approach. The intention is to consider a mid-term planning horizon, i.e., the upcoming three or four weeks. The decisions we take involve:

- Scheduling production orders based on sales order delivery dates;
- Monitoring the overall workload for capacity groups, anticipate and take action;
- Schedule extra capacity;
- Reschedule production orders;
- Order intake and issuing delivery dates.

4.3.2 Operational planning

The approach of overlapping capacity loops is an operational approach. At this level we consider the day-to-day planning. The capacity loops form a trigger mechanism with a double functionality. First and foremost, the trigger mechanism determines which production order is the right one to be processed. When a capacity group is ready to process a new production order, we apply a double authorization release check. Only if a production order meets both requirements is it eligible for processing. The authorization process is an automated checklist:

1. High level MRP release date is reached;
2. Utilization of succeeding capacity group is below critical level.

Second of all, if the mechanism cannot trigger a production order to be processed, this is a trigger to apply flexible capacity. The decision to reassign capacity depends on who is exactly working on a given day and what their specific skill set is compared to where extra capacity is useful. Actual reassigning of the employees is a process unsuitable for an automated system. We intend to let this process be controlled by the foremen, where they check the skill set of the employees in their capacity group. Subsequently, one of three decisions is taken:

1. Direct move to bottleneck. Reassign to different downstream capacity group with capacity problem;
2. Indirect move to bottleneck. Reassign to different non-bottleneck capacity group, freeing up capacity there to reassign to a bottleneck capacity group.
3. Work ahead of schedule.

N.B. Reassigning , either direct or indirect (Figure 4.15), is always preferable. Even if there are no other capacity groups behind on schedule, it is more preferable to speed up the process downstream than creating more intermediate WIP by working ahead of schedule. The only decisions we take in advance at this level are related to the utilization of the capacity group. The objective is to ensure a smooth flow of production orders. We determine how much required processing time is allowed in queue, compared to the total available capacity on a given day.

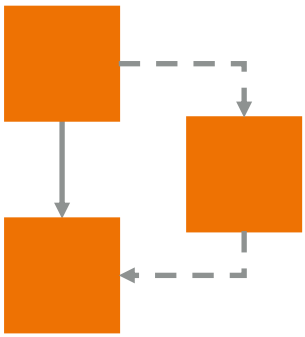


Figure 4.15: Direct flexible capacity move (solid line) and indirect flexible capacity move (dotted line)

The approach is schematically shown in Figure 4.16. The dotted line represents the division between push and pull. Above the dotted line we see a high level backwards MRP schedule using an (arbitrary) off-set lead time of two days per process step. Beneath the dotted line we have a pull approach where available capacity determines release. On the dotted line we see where the two approaches meet in a double authorization check. For every production order in queue, the red arrows represent an earliest release check and capacity check. If and only if both authorizations check out, does the green arrow occur, signaling the release of the particular production order. For example, suppose cell *B* just finished processing an order and is searching for the next one. If the next order in queue has cell *C* as its successor, it is checked for authorization based on (1) the earliest release date according to MRP and (2) available capacity at station *C*. Only if both checks are cleared, does (3) the authorization to release the production order occur.

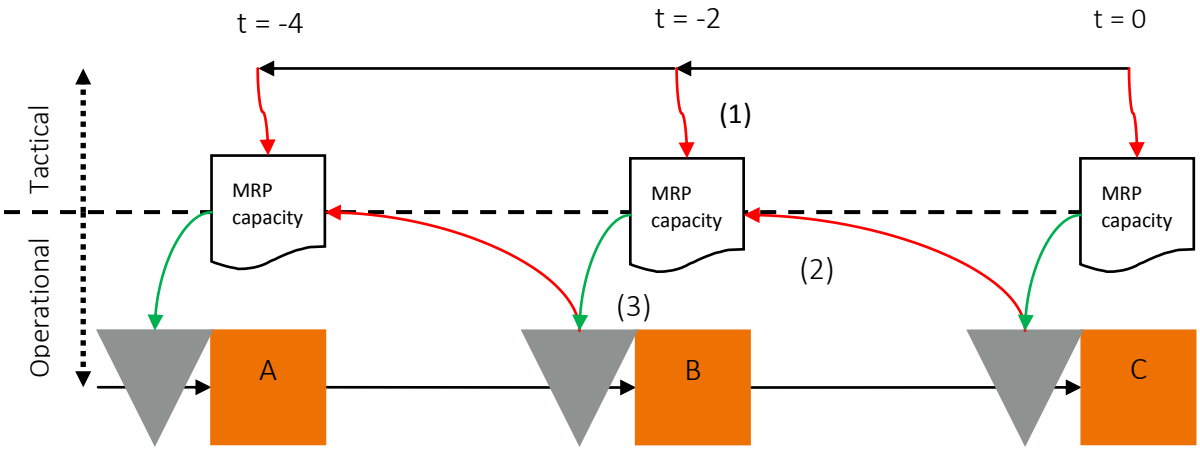


Figure 4.16: Schematic representation of the hybrid production order release mechanism

4.4 Electronic capacity loops

4.4.1 Theoretical capacity loops

Every capacity group forms a number of capacity loops, one loop with each immediate succeeding capacity group. The production orders can follow any number of combinations of these capacity groups, i.e., follow a routing. The product specific routings determine the capacity loops. In general, we see that some process steps precede others, i.e., bending is always before welding. However, some products follow an extraordinary path with a “loop back” before it moves on. For instance bending, then welding and back to handcraft bending to punch some hole (compare the routings in Figure 4.17). Because Company X is an ETO/MTO manufacturer, the routings and loops

also differ in time. Changes form no problem as we use electronic loops, e.g., if no production orders follow a partial routing from bending to stainless steel welding, this capacity loop is temporarily nonexistent.

4.4.2 Practical capacity loops

In practice, every capacity loop is an authorization process between two buffers, the first we will call the *requesting* buffer, the second the *receiving* buffer. The requesting buffer requests if the receiving buffer has sufficient capacity to receive extra workload. The nature of the manufacturing process at Company X leads to multiple overlapping loops, which cause two challenges to deal with.

First of all, multiple *requesting* buffers can ask a *receiving* buffer for an authorization, an *N:1* ratio (left side of Figure 4.18). When this occurs simultaneously, we could be faced with multiple releases, causing the workload of the receiving buffer to increase more than is allowed. Second of all, one *requesting* buffer can have multiple *receiving* buffers with available capacity to choose from, an *1:N* ratio (right side of Figure 4.18). The question is which production order to choose.

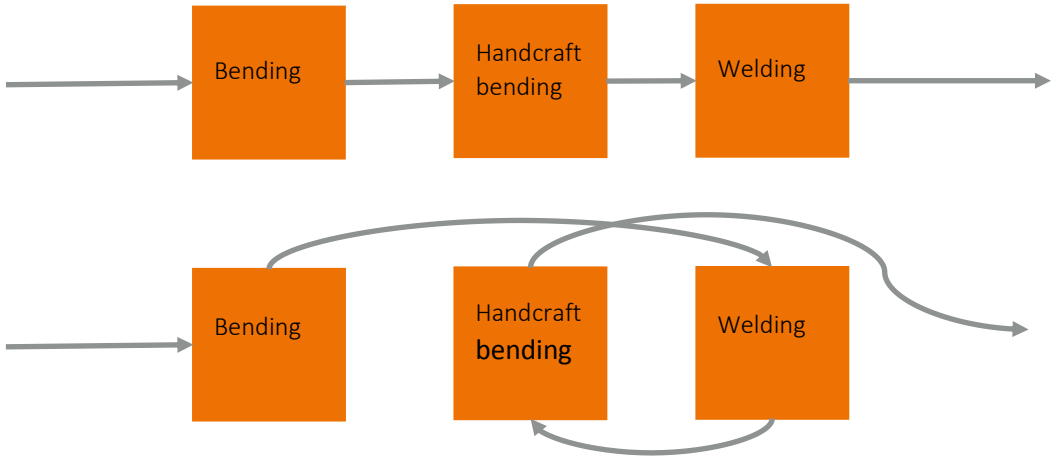


Figure 4.17: Example of a straight forward routing and a routing with a "loop back"

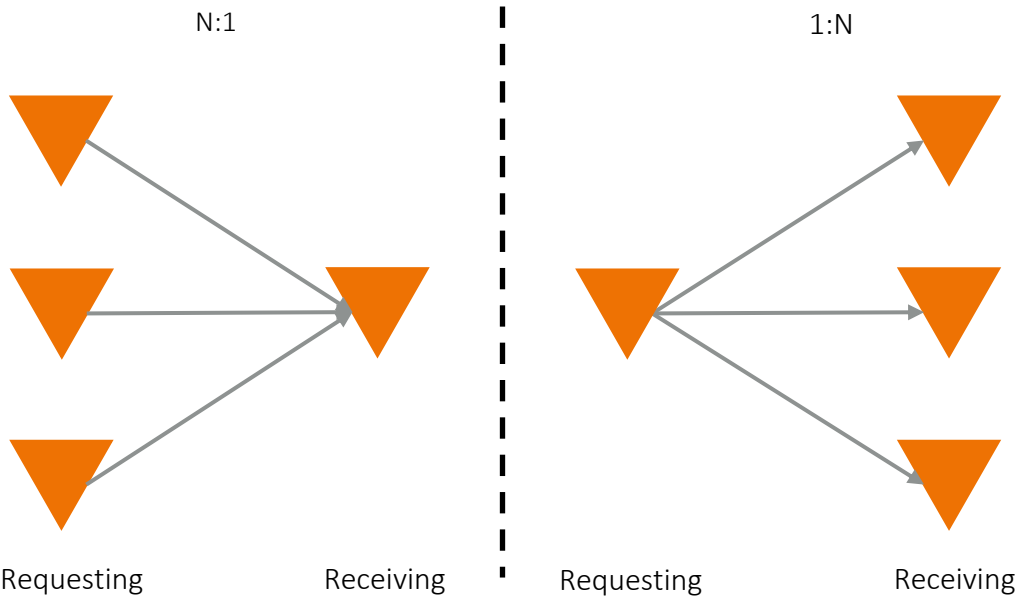


Figure 4.18: Schematic representation of the N:1 and 1:N ratio between requesting and receiving buffers

We intend to solve the challenges occurring in the *N:1* scenario, by incorporating *reserved* time in the *receiving* buffer. This reserved time is the processing time of the production order at the receiving side, while it is in progress at the requesting side. As soon as the *requesting* buffer receives the authorization to start processing, the process time of that particular order at the *receiving* (succeeding) station is added to the allocated capacity.

N.B. The processing time at the requesting side, i.e., the time it takes before the new order arrives at the receiving buffer, is *not* yet considered in this matter.

Difficulties in case of multiple available capacity authorizations, i.e., a *1:N* scenario, will be handled with the use of a priority rule. These priority rules are discussed in more detail in section 4.8.

4.4.3 Long lead time items

We did not yet take into account the processing time at the *requesting* side of the capacity loop. When the processing time is long, we run the risk of the *receiving* side running out of work. The replenishment time of the receiving buffer takes longer than its depletion time. To avoid this we should consider the processing time on the *requesting* side for future available capacity checks on the *receiving* side. The *surplus* time is the difference between the processing time at the *requesting* site, minus the time required to finish the workload at the *receiving* site. This is potential down time. This difference is added to the capacity left to allocate on the *receiving* site, for other replenishment orders.

We clarify this with an example; suppose we have two capacity groups, *A (requesting)* and *B (receiving)*, both with a maximum capacity of 8 hours. Capacity group *B* has a number of orders in queue, summing up to 4.5 hours in processing time. Furthermore, it is currently working on an order with a processing time of 1 hour. In total, 5.5 of the 8 hours of maximum capacity are allocated and 2.5 is free to allocate. Capacity group *B* signals available capacity to group *A*, which happens to have a suitable order. This particular order requires 8 hours right now at group *A* and 1 hour afterwards at group *B*.

Capacity group B

- In buffer: 4.5 hours of work
- In progress: 1 hour of work
- Maximum capacity: 8 hours

Capacity group A

- Processing time at B: 1 hours
- Processing time at A: 8 hours

Suppose we do not account for the surplus time. 5.5 hours later, all work at group *B* is finished. Group *A* still requires 2.5 hours to finish and send the new order to group *B*, leaving group *B* to become idle. To avoid this, we allow 2.5 hours of capacity for group *B* still to be allocated to other orders to be send its way.

4.5 Capacity levels and authorization

When the capacity groups are defined, we need to assign a capacity to each one, define when there is enough capacity available to authorize a release and which production order is best suited for release.

4.5.1 Capacity authorization

Taking into account the considerations in sections 4.2.2 and 4.2.3; when can a production order be released for processing? A production order is authorized for release if the current workload plus reserved capacity for orders on its way minus any surplus time, is smaller than the allowable workload. Expressed in a formula:

$$Capacity\ to\ allocate = Max\ capacity - workload\ receiver - reserved + surplus\ time$$

Where workload at the receiver is:

$$Workload = \left(\sum in\ buffer + \sum WIP\right)$$

The surplus term is only required if the order processing time is larger than the workload. Therefore the surplus term is dropped if it becomes negative:

$$Surplus\ time = Max(processing\ time\ requester - current\ workload\ receiver; 0)$$

N.B. Because we use the existing television screens (electronically linked to one another with computers) we feel it is both possible and useful to create a dynamic, real-time system as much as possible. We define the sum of WIP as the total remaining capacity required to finish the production orders in progress. When an order is half finished, half of its production time is free for authorization of a new order. This is a distinct deviation from the POLCA theory, where the capacity becomes available again once processing is completed and the polca-card is returned. This will enable the system to react a lot quicker.

We demonstrate and clarify the concept with an example of a hypothetical decision moment based on Figure 4.19 and Figure 4.20. All capacity groups have a maximum capacity of 8 hours. Capacity group *A* finished an order and turns to the television screen to find the next order to start. The orders with the earliest release date are meant for capacity group *B*. Unfortunately, the maximum workload is reached and we move down on the list. *P005* is the first order for which the next capacity group has available capacity. This will be the next order to be processed at group *A*.

At the same time, capacity group *A* (Figure 4.20) is also a *receiver*. It has a current workload of 5.5 hours (50+35+65+80+40+0.67[4]*90), leaving 2.5 hours left for any *requester* to send its way. Suppose a *requester* has an order requiring (*reserving*) exactly 2.5 hours at *receiving* group *A* and requiring 7 hours of processing time at the *requesting* side. The capacity left to allocate to group *A* becomes 1.5 hours (2.5-2.5+(7-5.5)). We have accounted for the *surplus* processing time at the *requesting* capacity group.

[4] Real-time utilization is applied. 33% of required processing time is completed, leaving 60 minutes.

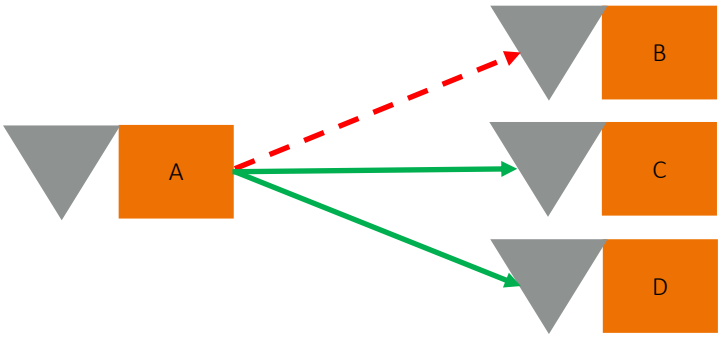
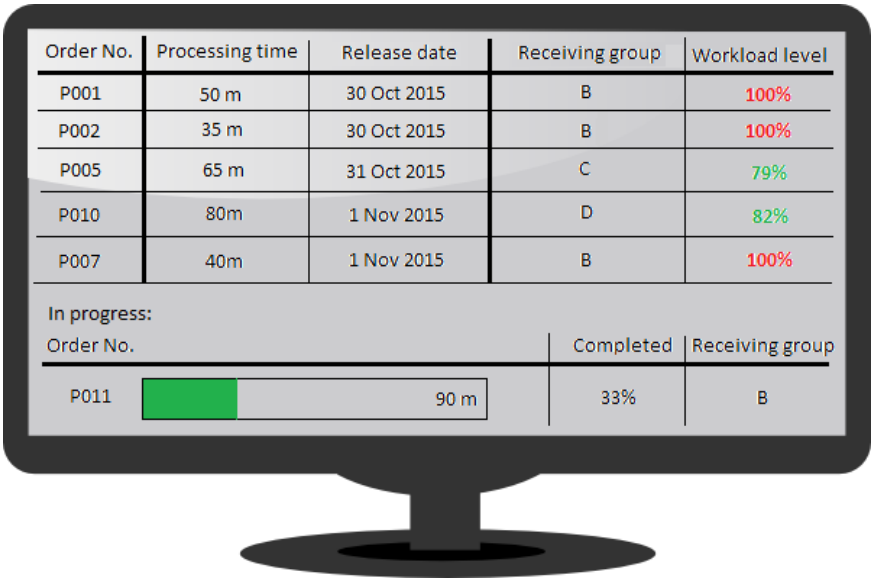


Figure 4.19: Example of a decision moment for the capacity authorization approach



Order No.	Processing time	Release date	Receiving group	Workload level
P001	50 m	30 Oct 2015	B	100%
P002	35 m	30 Oct 2015	B	100%
P005	65 m	31 Oct 2015	C	79%
P010	80m	1 Nov 2015	D	82%
P007	40m	1 Nov 2015	B	100%

In progress:

Order No.		Completed	Receiving group
P011	90 m	33%	B

Figure 4.20: Example of what the order release mechanism might look like on the screen at capacity group A

4.5.2 Priority rules

In section 4.6.2 we already mentioned the use of priorities rules in the context of the 1:N relationship between *requesting* and *receiving* buffers. When multiple production orders in a queue are eligible for release based on release date and capacity, which one should be next? We are talking about additional priority rules, other than time and capacity. One of the possibilities is prioritizing based on the utilization of the *receiving* buffers. The next production order to be processed should be the one which replenishes the *receiving* buffer with the lowest workload compared to its total capacity. In the example on the screen in Figure 4.20, order *P005* would have priority over *P010*, even if the release dates would be the same. Another possibility for an additional rule, is to give priority to products for converging process steps, i.e., give priority to the last part required to start an assembly or complete a shipment.

4.5.3 Utilization

On a higher, tactical level we should limit the amount of work we allow in a capacity group. What we are talking about is a maximum utilization parameter to adhere to. First of all, a high utility leads to higher waiting times. We know this from queuing theory. Second of all, there is some inherent uncertainty in the processing times, especially for initial products. If we allow for some margin of error, this will prevent the capacity groups from being overloaded in case of erroneous processing time estimates. How much work we actually schedule compared to the total capacity, is an arbitrary limitation. In the quantification in chapter 5, we experiment with different levels.

4.5.4 Capacity levels

The amount of capacity we define for each group is a combination of two things. First of all, it is based on the actual number of people employed and their skills. Second of all, the capacity we think is required based on historic data. It is an important strategic decision making process, because we can already identify a mismatch between required and available capacity. By applying well-designed capacity levels we can identify potential bottlenecks. Furthermore, we avoid the opposite as well. The unnecessary complexity of reassigning capacity to work outside the own capacity group when there is overcapacity.

4.6 Expected benefits

We expect this approach to have a number of positive effects, some related to the pull perspective of the approach, others related to the push strategy. The operational pull approach limits the WIP, ensuring a controlled lead time. Even if our MRP schedule theoretically overloads the manufacturing system, the capacity authorization will prevent this overload from flooding the work floor. Moreover, the pull function also offers a trigger function for flexible capacity. A lack of production orders authorized to start, is a signal for capacity problems downstream and help can be offered. The push approach on the other hand, helps us with a more clear prioritization. Production orders are not started before the earliest start date, which means that no unnecessary effort is spend working ahead of schedule.

Furthermore, the pull approach to control the amount of work on the work floor is clear and unambiguous. It fits well with the objective to find a transparent and understandable planning method (section 1.3). First introduction with the concept through exploratory meetings have resulted in a generally positive attitude towards the idea both at the shop floor level as well as on a management level. If introduced and implemented correctly, the approach can therefore also count on the broad support Company X is looking for.

One more benefit of the approach is related to implementation. The method can be implemented as a logical extension of the currently applied network of television screens. Because it is an extension rather than a radically different approach, the probability for successful implementation and acceptation is higher. As Hopp and Spearman (2000) write: “not every improvement needs to be presented as a new way of life ” (p. 373). After the limited success of the application of Lean manufacturing at Company X, we believe introducing yet another initiative as if it were a game changer could induce skepticism.

The rest of this chapter is a specification of some of the details and decisions to be taken before the approach can be implemented:

- Identification of production cells (the capacity groups);
- Capacity levels per capacity group;
- The utilization limit (total to schedule);
- Priority rules.

4.7 Capacity groups

We define a capacity group to be a (collection of) process step(s) and related production resources, for which the available capacity may be summed. For each of the process steps within the capacity group, at least one employee must be present to perform the task at all times. This is to ensure that a production order can always be processed within the off-set lead time. Furthermore we prefer a capacity group to require at least one FTE worth of production orders. This is to reduce the number of groups and complexity. The capacity between capacity groups is non-overlapping, i.e.,

the capacity of an employee with the skills to perform multiple (capacity group transcending) tasks is assigned to only one capacity group. This is to avoid his or her capacity potentially being allocated more than once. The capacity groups we have defined are based on experiences shared by the foreman, planning department, management and on historical workload:

4.7.1 The capacity groups

- Bending programming
- Bending
- Robot bending
- Handcraft bending
- Welding
- Stainless steel welding
- Programming welding robot
- Welding robot
- Spot welding and stud welding;
- Painting
- Assembling

N.B. The cutting department is separately controlled. The machinery at the cutting department is specifically designed for production of larger batches. Machine and material efficiency are most important. It is the only step where parts are produced in a batch (nest) and production is not strictly demand driven, some of the parts are not needed until later. We propose a CODP-based push-pull boundary (see section 2.4.1) between the cutting department and the rest of the factory. The tangible boundary being the warehouse for components cut from sheet metal, awaiting further processing. We offer the cutting department two days' worth of orders to be finished in time. Thereby we enable the department to schedule its work as efficiently as possible. We only require all components to be cut and available at the predetermined day, on the next day, the rest of the capacity groups can start pulling these components from inventory based on what the capacity loops indicate. The actual components to cut is based on the customer orders to be manufactured in the upcoming unit of time. We suggest to work with a time bucket of two days. We want to limit the amount of intermediate components in inventory, yet also have a large enough workload to have room for efficient production based on sequencing of jobs on the machine.

4.7.2 Explanation bending groups

Some of the decisions are simply based on product characteristics, a product either is suitable for automated bending using a robot or not. The programming part of the bending process is defined as a separate group, because the task is performed by different, specialized programmers. Their time is limited, therefore it makes sense to control what they program, based on the machine's workload.

4.7.3 Explanation welding groups

Another product characteristic, a product is either made of stainless steel or not. These materials cannot come into contact with one another and are therefore separated. The decision is also based on personnel and skills. For instance, not every welder is capable of performing spot welding and stud welding. If we were to incorporate those two in a capacity group with normal (stainless steel) welding, it could lead to a situation where no one is available to process a spot welding order, because the specialists are working on normal welding orders.

Spot welding and stud welding are incorporated into one capacity group together, because of the limited workload. Based on historical data between December 2014 and June 2015, spot welding orders accumulate to a workload of at least eight hours a day in 45% of the days. For stud welding this is only 4%, therefore it is decided to combine these processes.

Robot welding is defined separately, because the work cannot be simply interchanged between it and conventional welding by hand. The product needs to be programmed first, which takes a while, making it economically infeasible. Vice versa, if the program is present, we cannot simply reassign the product to conventional welding. Conventional welding is slower, therefore the manual labor would be more costly. The programming work for the robot is also defined as a separate group. It can be performed offline, meaning that the robot can be working on something else in the meantime. But the program needs to be written beforehand none the less and capacity for it must be reserved.

4.7.4 Explanation painting group

Painting as a group has a couple of distinct processes, but these can all be performed by the people working at the painting department. As such, the capacity can be treated as a whole.

4.7.5 Explanation assembling group

The capacity for assembling, i.e., the employees, are divided over the assembling orders at hand, based on their experience with the product. As such, the ones most experienced with a product are tried to be assigned to its assembly. If the workload requires it, less experienced employees can always perform an assembly, there is no risk of a job having to wait for an employee with the right skill set. Furthermore, there are no distinctive processes for a logical separation into different capacity groups.

4.8 Summary

Based on the conclusion that controlling and scheduling the production process is impossible beforehand, we are looking for an approach to cope with the inherent uncertainty Company X faces. Bottom line, we want to roughly assess required versus available capacity and schedule production orders accordingly (production planning). Once this has been done, we require some method to ensure that the scheduled production orders are finished on time, meanwhile achieving an efficient order flow (production control). A combination of a push and pull strategy, or hybrid approach.

Company X's ERP system is suitable for the push strategy, it has the functionality to automatically schedule production orders (MRP) and show us scheduled workloads to assess remaining capacity. When a backwards MRP schedule is made, it is based on a predetermined offset lead time. As long as we keep to this schedule, all production orders should move through the factory in synchronicity towards final assembly. We still require a control mechanism for internal lead time between departments, avoiding excessive waiting time. The POLCA material control system describes what we are looking for. A concept of multiple, routing-specific capacity loops which control the WIP in an operational setting, combined with an ERP system limiting WIP in a tactical setting.

The main challenge we identified is suitable method to signal available capacity. POLCA uses routing-specific authorization cards. The number of cards is difficult to determine for Company X, because the product mix changes frequently. Keeping in mind that the card is nothing more a way to signal free capacity, we propose to use the existing network of television screens for electronic authorization instead. The screens can be modified to visualize the production order priority based on MRP earliest release and capacity availability.

We distinguish two planning levels. Firstly, the tactical MRP schedule and secondly the operational level of overlapping capacity loops. The MRP planning level is meant as a tactical approach, scheduling production orders based on sales order delivery dates and monitor the overall workload. If capacity is insufficient we take action accordingly. At a tactical level we also manage order intake and issue delivery dates. The approach of capacity loops is an operational approach, considering the day-to-day planning. The capacity loops form a trigger mechanism with a double functionality, prioritize on MRP release date and assess if the succeeding capacity group has available capacity. If the operational mechanism cannot trigger a production order, this is clear signal to apply flexible capacity.

Every capacity group forms a number of capacity loops, one with each if its immediate successors. Because Company X is a ETO/MTO manufacturer, the routings and loops differ over time. The use of electronic capacity loops on television screens helps us to cope with change, if no production order follows a partial routing through the factory, the associated capacity loop is temporarily nonexistent. We also try to take advantage from the screens as much as possible by creating a dynamic, real-time system, capable of quick response. If a production order is partially completed, the finished time is signaled as free capacity. We also incorporate reserved processing time in a capacity group. If a production order is being processed by a predecessor, we reserve its processing time in the next station to avoid unwanted simultaneous releases. Finally, if a replenishment order takes more time to be finished at the current station than it reserves at the next, this additional time is marked as a surplus, capacity still available to assign. This is to avoid a buffer from depletion because the replenishment time is higher. Every capacity group signals available capacity if the current workload plus any surplus time is below its maximum capacity level.

We expect this approach to have a number of positive effects, some related to the pull perspective of the approach, others related to the push strategy. The operational pull approach limits the WIP, ensuring a controlled lead time. Even if our MRP schedule theoretically overloads the manufacturing system, the capacity authorization will prevent this overload from flooding the work floor. The push approach on the other hand, helps us with a more clear prioritization. Production orders are not started before the earliest start date, which means that no unnecessary effort is spent working ahead of schedule. Furthermore, the pull approach is clear and unambiguous. It fits well with the objective to find a transparent and understandable planning method. Introduction of the concept has resulted in a generally positive attitude both at the shop floor level as well as on a management level. The method can be implemented as a logical extension of the currently applied network of television screens. Because it is an extension rather than a radically different approach, the probability for successful implementation and acceptation is higher.

5. Simulation study

In this chapter we use a simulation study to quantify the new approach from chapter 4. In section 5.1 we explain why we use a simulation model. In section 5.2 we introduce the model and input. Section 5.3 describes the concepts of interests, these are tested and the results are shown in section 5.4 and 5.5. Finally, in section 5.6 we draw conclusion based on the results and mention some of the limitations of this simulation study.

5.1 Simulation

In chapter 3 we conclude that we are not able to predict lead time under the conditions at Company X. Before we implement the approach from chapter 4, we do like to have an indication of the performance of the new order release mechanism in comparison to the current release. Therefore we apply a simulation study. It enables us to take the variability[5] of Company X into account, more than an analytical approach can. The objective of this simulation study is to deliver a proof of concept. We want to research the effect of the experimental release mechanism in a congested system, i.e., with bottleneck. On the one hand the situation of order release as it is now, on the other hand the new order release based on the capacity loops. With the use of a simplified set-up of the factory we compare order release strategies. We use as much real, historic data from chapter 3 as possible to attain a degree of validity. The objective however, is not to exactly simulate the current situation at Company X and compare this to a situation where the new approach is applied. We do not attempt to exactly recreate all routings, product mix, arrival rates or total required and available capacity. After all, there is no such thing as a typical, representative product-mix.

5.1.1 Nature of simulation

A distinction can be made between terminating and non-terminating simulation studies. A terminating simulation run is ended by some natural event, i.e., the end of a work day. In a non-terminating simulation no such event occurs. A non-terminating simulation is associated with testing adjustments to existing systems and comparing the results to the status quo. The long-run performance is of interest. The simulation of a system can be terminating or non-terminating depending on the objectives we are interested in (Law, 2007).

In this simulation study we encounter natural events, the end of a shift on a given work day, leading to a termination of the simulation. However, we are primarily interested in the long-run performance of the order release mechanism. Therefore in section 5.3 we determine what “in the long-run” means for our model, by finding a warm-up period.

5.2 Model

The input of the simulation study is discussed in detail in the following sections:

- Factory model, routings and production orders;
- Processing time and uncertainty;
- Arrival rate;
- Capacity;

[5] We can take more of the variability into account, but unfortunately not all of it. For instance, the flexible interchanges of capacity between groups is not modeled.

5.2.1 *Factory model, routings and production orders*

The manufacturing facility we model is a simplified version of the production facility at Company X. Figure 5.21 shows the six capacity groups used in the simulation study. A capacity group is a collection of similar processes, requiring similar resources. The production orders can follow a number of routings along these capacity groups:

- Bending;
- Robot bending;
- Welding;
- Stainless steel welding;
- Painting;
- Assembling.

These routings are shown in Table 5.5. For example, a production order might require processing at the bending department, stainless steel welding department, painting department and finally require an assembly step (Figure 5.21). The production orders are modeled as one component to be processed at a sequence of processing stations. We model 12 different routings (Table 5.5), simulating 12 different production orders. These are typical routings for Company X, however not exhaustive. The production orders are generated randomly with an equal probability of occurrence.

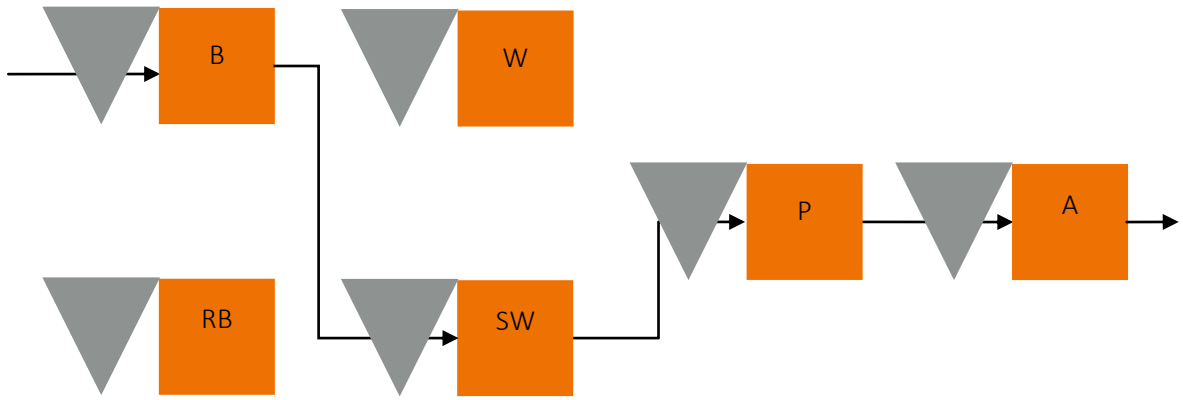


Figure 5.21: Simple set-up for simulation study with example routing (B=Bending, RB= Robot bending, W=Welding, SW=Stainless steel welding, P=Painting, A=Assembling)

Table 5.5: Simulated routings with equal probability of occurrence

Component	Routing	Occurrence
1	RB-SW-P-A	8.33%
2	RB-W-P-A	8.33%
3	B-SW-P-A	8.33%
4	B-W-P-A	8.33%
5	RB-SW-A	8.33%
6	RB-W-A	8.33%
7	RB-P-A	8.33%
8	B-SW-A	8.33%
9	B-W-A	8.33%
10	B-P-A	8.33%
11	RB-A	8.33%
12	B-A	8.33%

5.2.2 *Processing time and uncertainty*

For every production order the processing time at every process step in its routing is randomly generated. This processing time is generated using a normal distribution with a mean and standard deviation that corresponds with the historical data from chapter 3 (see Table 5.6). A lower bound is set to avoid a negative processing time (which is theoretically possible) and an upper bound is set at 16 hours, or two full shifts. An upper bound of 16 hours is sufficient, as the processing times rarely exceed this. This processing time is the planned (estimated) processing time. For Company X this would be the VoCa time (Dutch: *voorcalculatie*).

We also generate a processing time margin of error (standard normally distributed) between -25% and +25%. This margin is either added or subtracted from the planned processing time to get a real processing time. This would be the NaCa time (Dutch: *nacalculatie*). For example, the generated planned processing time might be 60 minutes, with a margin of error of +14%. The actual processing time would therefore be 68.4 minutes. According to Table 3.4 (section 3.3) the actual processing time is between 77 and 105% of planned processing times. Recent order intake is predominantly initial work, deteriorating the actual/planned processing time ratio. For this reason we have chosen a little wider range.

Table 5.6: Planned processing time (hours) and standard deviation as measured in actual production

Component	μ	σ	Lower bound	Upper bound	Processing time > upper bound
Bending	1.1	1.6	0.08	16	0%
Robot bending	3.5	3.4	0.08	16	1%
Welding	3.1	5.6	0.08	16	3%
Stainless steel welding	2.9	5.8	0.08	16	2%
Painting	0.4	0.8	0.08	16	0%
Assembling	2.6	7.2	0.08	16	4%

5.2.3 *Arrival rate*

A production order is generated according to a normal distribution[6], with a mean of 46 minutes and a standard deviation of 15 minutes. The lower bound between order arrivals is a minute, the upper bound is 90 minutes. The arriving order will be one of the 12 types as discussed in section 5.2.1.

5.2.4 *Capacity*

The capacity groups have the following capacity on a daily basis. This capacity is doubled in case of a two day off-set lead time. These are not representative for Company X.

- Bending, 16 hours
- Bending robot, 24 hours
- Welding, 24 hours
- Stainless steel welding, 24 hours
- Painting, 8 hours
- Assembling, 80 hours

[6] A normal distribution is chosen for simulation purposes, the arrivals at Company X do not necessarily follow this distribution. The mean value of 46 minutes is chosen to generate enough orders to create WIP in the buffers, without overloading the system.

For simulation purposes, the previously discussed input creates a mix of production orders flowing through the factory with varying arrival rates and estimated processing time errors. It should be noted that this product mix is not representative for Company X. To ensure the comparison is fair, all input variables are fixed. The only difference is the changing release strategy. We are also interested in the effect of the new approach in combination with utilization. In other words, we study the resulting performance when workload increases. To do this, we change the input variable utilization between experiments. As such, we can compare the new approach under different levels of allowable workload. This input information is incorporated in a model, build and tested using Tecnomatix Plant Simulation 11 (Figure 5.22).

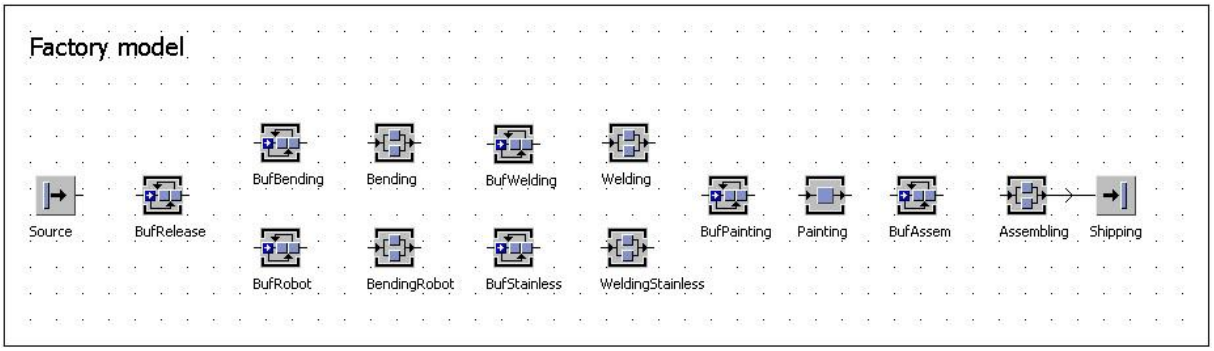


Figure 5.22: Model of manufacturing process at Company X using Tecnomatix Plant Simulation 11

5.2.5 Warm-up period and run length

When we are interested in performance over time, we are looking for a steady-state parameter. The value of this parameter over time takes on a steady value, i.e., the performance over time is about the same. This performance can be compared to other steady performances of the same system under different experimental designs.

The main interest of our study is the production lead time, therefore it makes sense to use it as steady-state parameter. It is a performance indicator which is strongly influenced by the effect of a half-empty production facility, because we just started generating production orders. Therefore we determine the so-called "warm-up" period and exclude the data from our analysis. Figure 5.23 shows us the long-run lead time for the current situation as a baseline. After 140 days, the moving average[7] is a steady-state performance with a mean around 80 hours per production order. The simulation is run for a total of 1020 days (4 years of 255 workings days), the data obtained up until the 140th day is not taken into account in the analysis of the current and the experimental release method.

[7] We use a moving average with a window W=20, see (Law, 2007) p. 508.

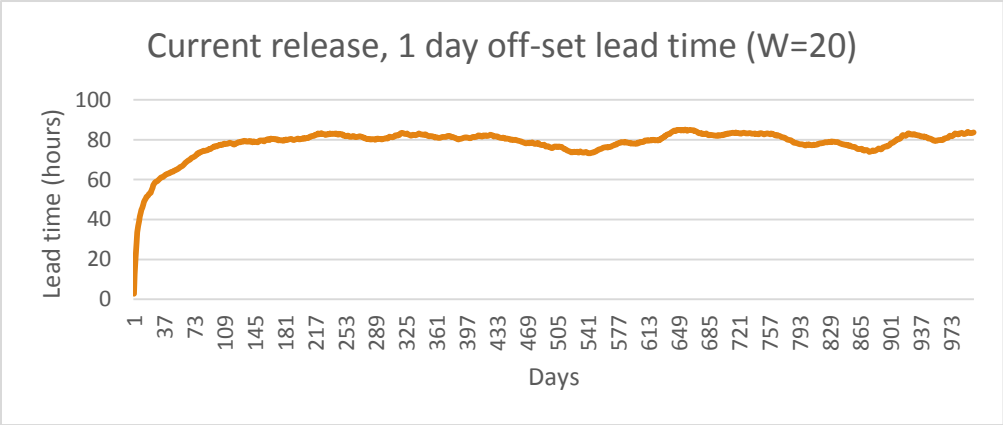


Figure 5.23: Graph of moving average daily lead time (W=20)

5.3 Concepts of interest

5.3.1 Current release strategy

For the current situation, the release of production orders is modeled as a release mechanism with just the earliest release date as priority rule. Whenever the planned start date is reached, the order is eligible for release (for the sake of simplicity, we assume that the materials are available). If no orders with a higher priority are present, an order which has not yet reached its earliest release date will be processed. Production orders are stored in a *first come first serve* (FCFS) buffer with a capacity assumed to be infinite. Each capacity group has such a buffer. After processing is completed, the production order moves on to the next buffer in the routing. FCFS is not entirely representative for the current situation, where orders are sometimes given priority. Again, we are interested in a comparison between pure push (MRP) release strategy and a hybrid approach for the entire workload without resorting to rush orders for one particular customer. The exact sequence of events is shown in the flow chart in Appendix A.

5.3.2 New release strategy

In the new release mechanism every capacity group still has a buffer with infinite capacity, where production orders are stored. We model the release of production orders to have the double authorization check (DA method), the highest priority based on earliest release date and the next station in the capacity loop must have available capacity. The exact sequence of events is shown in Appendix B.

To determine if there is available capacity, we compare the workload ratio of a capacity group to a predetermined utilization parameter. The workload ratio of a capacity group is calculated using the formula from section 4.7 and dividing it by the total available capacity. Subsequently this number is compared to the allowable workload on the right-hand side:

$$\frac{Total\ capacity - workload + surplus\ time^{8}}{Total\ capacity} < utilization\ parameter$$

5.3.3 Workload level

We limit the allowable workload level because of the inherent uncertainty in the planned, estimated processing time. In a practical implementation we would limit the allowable workload for

[8] Surplus time is used when the processing time at the sending capacity group is longer than the workload at the receiving capacity group. This could cause a capacity group to become idle because it must wait on replenishment orders. See section 4.5.

a capacity group on a tactical level. If we allow 100% of the total capacity to be allocated to order intake, any delays due to unforeseen errors in processing time potentially effect a sales order. If we limit the order intake, yet allow 100% of the workload on the shop floor itself, we expect to achieve an output as high as possible without risking sales order due dates.

Our simulation model is a theoretical model at an operational level, i.e., we cannot simulate the allowable workload at our preferred level. Still, we are interested to see what the effect is and limit the workload with a utilization parameter. The actual workload entering the work floor is the same (e.g., 90% of 100h = 90h = 100% of 90h).

5.4 Allowable workload

5.4.1 Allowable workload experiment

First we simulate the allowable workload under different levels of variability in processing time. Deterministic processing time, planned and actual processing time are equal (no margin of error). Low variability, actual processing time is between 75% and 125% of planned processing time. High variability, actual processing time is between 50% and 150% of planned processing time. The range of variability is the same for each process step. We expect that a situation with a higher degree of variability benefits from a lower allowable workload. We perform the experiment for both a one day and two day off-set lead time. A one day off-set lead time is used because it is the current situation. We apply a two day off-set lead time because we are interested in finding out if a little more slack in the schedule might be beneficial. Two days is the maximum however, a three day off-set lead time is not acceptable. Applying three days would mean scheduling each process step within a three-day time frame, an increase of 300% to schedule sales orders compared to the current situation. We cannot justify this to management and sales. We simulate an allowable workload between 80 and 100%. No less than 80% because of underutilization, no more than 100% for obvious reasons.

We compare the results on three different KPIs. We are not only interested in better average performance based on lead time. Perhaps more important is to find out if the degree of variability influences the ability to control the MRP schedule, i.e., enable a consistent flow and avoid (intermediate) missed due dates.

1. The average order lead time and its standard deviation;
   The average production order lead time is measured as the time lapse between order entry into the factory until completion at the assembling department. Order entry in this instance refers to the moment at which the production order leaves the release buffer.
2. The average completion time per capacity group and its standard deviation (MRP control);
   The time lapse between order entry in the buffer of a capacity groups and moment of finish of the processing.
3. The average order throughput and its standard deviation;
   The number of production orders entering the production facility, i.e., exit from release buffer, and the number of production orders leaving the production facility via the shipping department[9].

5.4.2 Lead time

The lead time results of the simulation are listed in Table 5.7. If allowable workload remains unchanged, but the level of variability in processing time increases, we observe what is to be

[9] The shipping department is simulated as a sink in the simulation model. It does not do anything except remove the production order from the model.

expected. The average lead time and its variability increases in every scenario, one or two days off-set lead time with 80, 90 or 100% workload.

Table 5.7: Average lead time (hours) for the new order release mechanism under different levels of processing time uncertainty

Allowable workload	Deterministic		Low variability		High variability	
	μ	σ	μ	σ	μ	σ
100% (1 day)	62.3	4.2	62.4	4.1	65.0	4.6
90% (1 day)	65.2	4.2	65.0	4.7	68.9	5.0
80% (1 day)	69.1	4.7	70.7	5.1	72.8	5.3
100% (2 days)	78.3[10]	13.8	73.5	4.5	75.0	4.7
90% (2 days)	70.4	4.4	70.6	4.2	71.8	4.5
80% (2 days)	67.8	4.1	68.1	4.4	69.0	4.4

If we compare the average lead time with a changing allowable workload, while the variance in processing time remains the same, we see an interesting result. When we apply an off-set lead time of a day, we observe that a decrease in allowable workload causes the lead time to increase. We observe this result for a deterministic processing time, low variability and high variability. When we make the same comparison, but apply an off-set lead time of two days, we observe exactly the opposite. As we reduce the amount of work we allow, the lead time and variance decreases.

5.4.3 Completion time – MRP control

The ability of the new order release mechanism to control the MRP schedule with a one day and two day off-set lead time is shown in Table 5.8 and Table 5.9 respectively. The ability to control the schedule is measured as the probability of a timely completion of production orders at the capacity groups. We use the normal cumulative probability density function with the average and standard deviation of completion time.

Table 5.8: Probability of order completion within 1 day off-set lead time under different levels of variability in processing time (probabilities based on normal distribution)

Workload and variability	Bending	Robot	Welding	Stainless	Painting	Assembling
100% - no variability	90%[11]	3%	100%	100%	100%	100%
90% - no variability	41%	5%	100%	100%	100%	100%
80% - no variability	6%	7%	100%	100%	100%	100%
100% - low variability	88%	3%	100%	100%	100%	100%
90% - low variability	42%	5%	100%	100%	100%	100%
80% - low variability	3%	7%	100%	100%	100%	100%
100% - high variability	63%	3%	100%	100%	100%	100%
90% - high variability	14%	5%	100%	100%	100%	100%
80% - high variability	2%	7%	100%	100%	100%	100%

From Table 5.8 (one day off-set lead time) we conclude that the bending robot capacity group is the bottleneck. Limiting the allowable workload is beneficial according to the simulated results. The queue of a bottleneck is full, therefore choice is not an issue. In this case we benefit more from limiting the workload thereby decreasing the variability. This is in line with the conclusion

[10] The scenario forms the exception on the conclusion that a higher variability leads to a higher production lead time.

[11] $P(x < \bar{X})$ where *x* is 24 hours (one off-set lead time) and $\bar{X} \sim N(18.3, 4.4)$

we draw based on lead time. We can also see that the degree of variability does not make a difference for the bottleneck anymore, when the off-set lead time is a day. It is 3, 5, 7% probability in each instance. For a non-bottleneck capacity group we do observe what is to be expected; as the variability increases, the ability the control the off-set lead time decreases. For instance, if we look at bending at 100% allowable workload with no variability, low variability and high variability. We observe 90, 88, 63% probability for on time completion.

Table 5.9: Probability of order completion within 2 day off-set lead time under different levels of variability in processing time (probabilities based on normal distribution)

Workload and variability	Bending	Robot	Welding	Stainless	Painting	Assembling
100% - no variability	97%	28%	100%	100%	100%	100%
90% - no variability	100%	56%	100%	100%	100%	100%
80% - no variability	100%	87%	100%	100%	100%	100%
100% - low variability	100%	34%	100%	100%	100%	100%
90% - low variability	100%	60%	100%	100%	100%	100%
80% - low variability	100%	87%	100%	100%	100%	100%
100% - high variability	100%	20%	100%	100%	100%	100%
90% - high variability	100%	53%	100%	100%	100%	100%
80% - high variability	100%	82%	100%	100%	100%	100%

In Table 5.9 we see that the off-set lead time of two days offers enough leeway to avoid the bending department from becoming a bottleneck. For the bottleneck capacity group (robot) we see that for every level of variability, limiting the workload improves the performance regarding MRP control. If we compare the levels of variability to each other, we do not see a clear pattern. The performance seems to benefit from more from lower variability compared to no variability, which is not logical. Then performance decreases again if we compare low and high variability, which does stroke with our expectation.

5.4.4 Throughput

Table 5.10 and Table 5.11 show the throughput performance in case of a one day and two day off-set lead time respectively. The average and standard deviation of the daily number of orders entering and exiting the manufacturing facility are given. We can clearly see in Table 5.11 that an off-set lead time of two days gives enough leeway such that throughput performance is no longer influenced by variability. The results in Table 5.10 tell us that different degrees of variability do have a slight influence when off-set lead time is a day. Though only slightly, under high variability the throughput performance decreases. Furthermore, under decreasing allowable workload, the throughput performance is not affected in the case of a two day off-set lead time. This means we can decrease it to improve MRP control without compromising on throughput performance. This is not the case when we apply a one day off-set lead time, where the throughput decreases when we decrease the allowable workload. The differences might look small, but on a yearly basis the difference between highest (7.4) and lowest (6.2) is 306 sales orders, or 19%.

Table 5.10: Daily number of orders entering and exiting the simulated factory using a one day off-set lead time

Workload and variability	In		Out	
	μ	σ	μ	σ
100% - no variability	7.0	0.6	7.0	0.6
90% - no variability	6.7	0.7	6.7	0.6
80% - no variability	6.5	0.7	6.5	0.6
100% - low variability	7.0	0.6	7.0	0.6
90% - low variability	6.8	0.7	6.8	0.6
80% - low variability	6.4	0.7	6.4	0.6
100% - high variability	6.9	0.6	6.9	0.6
90% - high variability	6.6	0.7	6.6	0.6
80% - high variability	6.2	0.8	6.2	0.6

Table 5.11: Daily number of orders entering and exiting the simulated factory using a two day off-set lead time

Workload and variability	In		Out	
	μ	σ	μ	σ
100% - no variability	7.2	0.6	7.2	0.6
90% - no variability	7.3	0.6	7.3	0.6
80% - no variability	7.3	0.6	7.3	0.6
100% - low variability	7.4	0.6	7.3	0.6
90% - low variability	7.3	0.6	7.3	0.7
80% - low variability	7.3	0.6	7.3	0.6
100% - high variability	7.3	0.6	7.3	0.6
90% - high variability	7.3	0.6	7.3	0.6
80% - high variability	7.3	0.6	7.3	0.6

5.4.5 Conclusion

Based on the result of the simulated example setting, we conclude that an off-set lead time of two days performs better than a one day off-set lead time under different levels of variability. Average lead time performance is a little less favorable, but the ability to control the MRP schedule significantly improves. The throughput of the simulated facility is higher and more stable when using a two day off-set lead time. In combination with an allowable workload of 80%, the performance for a two day off-set lead time is consistently best. If we apply a one day off-set lead time we observe the opposite, 100% workload performs better.

An explanation for these contradicting results could be related to the actual number of orders that make up the WIP we allow. This number of orders determines the number of choices we have for our next production order to process. Suppose we apply an off-set lead time of one day and allow just 80% of total capacity to enter a capacity group, we restrict the amount of orders in queue. Maybe so severally, the new approach is left without enough production orders to make a sensible choice. The average number of daily production orders to choose from in Table 5.12, corroborates that the number of orders to choose from at the robot decreases. Table 5.12 offers an alternative explanation as well. A one day off-set lead time and restricted workload seem to cause a second bottleneck to emerge. The bending capacity group. Every production order must either pass bending or the bending robot, if both become a bottleneck because of restricted allowable workload, lead time will increase.

Table 5.12[12]: Average number of production orders in queue at different levels of workload

Workload	Bending	Robot	Welding	Stainless
100% (1 day)	2	3	0	0
90% (1 day)	4	3	0	0
80% (1 day)	5	2	0	0
100% (2 days)	1	8	1	1
90% (2 days)	0	7	1	1
80% (2 days)	1	6	1	1

With the results from the simulation on allowable workload, we can construct a rule of thumb for Company X. We want to gain as much advantage from our new approach, based on the trade-off between having choice of production orders in queue and uncertainty in processing time. In general, we can say that the number of orders allowed in a capacity group during the off-set lead time, is related to the ratio between available capacity and average processing time:

$$\frac{Utilization\ parameter * Total\ capacity}{Average\ processing\ time} = No.\ orders\ in\ capacity\ group$$

We can deduce that a large average processing time, compared to the available capacity results in a relatively low number of orders in the capacity group. In such a case, we could benefit from using all available capacity. We allow all capacity to be used, thereby increasing the possibility of choice in queue. Ultimately trying to leverage the benefits of the new approach. On the other hand, if the average processing time is small, we will end up with multiple orders to choose from anyway. At which moment, we would be better of decreasing the number of orders, because with every additional order in the capacity group, the processing time uncertainty increases.

We demonstrate with an example. Suppose we have two capacity groups, A and B, both have a total capacity of 24 hours. Our average processing time in capacity group A is 6 hours, for capacity group B it is 1 hour. In capacity group A, we would benefit from allowing all 24 hours to be available, thereby ensuring we can choose from $\frac{100\% * 24}{6} = 4$ production orders. In capacity group B though, we could easily restrict available capacity to 80%. We still have $\frac{80\% * 24}{1} \approx 19$ production orders to pick from, but if any of them takes longer than anticipated we are not immediately in trouble.

[12] Painting and assembling are not included, because there is no choice involved. The following process step is deterministic.

5.5 Current vs. new approach

5.5.1 Experimental scenarios

In this experiment we compare the current against the new release strategy in three different simulation scenarios (Table 5.14). One day off-set lead time is the current practice and is therefore the baseline choice of parameter setting. For the new approach we compare one and two days of off-set lead time. No more than two days, see section 5.4.1. In this simulation, we will use the most favorable workload level for each scenario, based on the KPIs. According to our findings in the last section, this means 100% for scenario 2 and 80% for scenario 3. It does not apply to the current release of course, as it does not take workload into account. We apply a low variability (see section 5.2.2).

Table 5.13: Experimental design scenarios to study using the simulation model

	Simulated release model	Off-set lead time	Workload level
Scenario 1	MRP earliest release (baseline)	1 day	N.A.
Scenario 2	Double authorization (DA)	1 day	100%
Scenario 3	Double authorization	2 day	80%

To find out if the new approach outperforms the current method of order release, we use the same KPIs as for the workload:

- The average order lead time and its standard deviation;
- The average completion time per capacity group and its standard deviation (MRP control);
- The average order throughput and its standard deviation;

5.5.2 Lead time

The average lead time and standard deviation for the current order release method (the baseline) and the new double authorization release are given in Table 5.14. Based on the averages, we can conclude that the experimental release outperforms the baseline in both scenarios. Furthermore, based on the standard deviations, we can also draw the conclusion that our new order release reduces the variability in order lead time as well.

Table 5.14: Lead time (in hours) of simulated scenarios

	Average lead time (hours)	Standard deviation
Scenario 1	80.2	7.0
Scenario 2	62.4	4.7
Scenario 3	68.1	4.4

5.5.3 Completion time – MRP control

The probability of an order being finished within the off-set lead time is shown in Table 5.15 (probabilities are calculated as in Table 5.8). The results show that the new approaches improve the performance in this regard. Scenario 2 might have the best average lead time, scenario 3 out performs it when it comes to controlling the MRP schedule. For both scenario 1 and 2, the welding and stainless steel welding capacity group have a high probability of missed MRP due dates. The MRP schedule is therefore infeasible with a high probability, requiring corrective actions.

Table 5.15: Probability of completing a production order within the MRP off-set lead time

	Bending	Robot	Welding	Stainless	Painting	Assembling
Scenario 1	100%	0%	67%	68%	100%	100%
Scenario 2	88%	3%	100%	100%	100%	100%
Scenario 3	100%	87%	100%	100%	100%	100%

5.5.4 Throughput

The average order output in scenario 2 is a little lower (Table 5.16). Most likely because the number of orders entering the factory is also lower (Table 5.16). A lower number of orders entering can be explained by the more restrictive one day off-set lead time. The order output of scenario 3 is almost equal to the current. We expect the difference to be related to the simulation model's lack of immediate (real-time utilization) response[13]. Again, the difference between the best output performance and the worst is only 0.5. On a yearly basis this is a difference between 1785 or 1913 sales orders, or 7%.

Table 5.16: Average daily order intake and output

	Average intake	Standard deviation	Average output	Standard deviation
Scenario 1	7.5	0.3	7.5	0.6
Scenario 2	7.0	0.6	7.0	0.6
Scenario 3	7.3	0.6	7.3	0.6

If the number of arriving orders exceeds the capacity, in the current approach these production orders enter the capacity groups and cause congestion. In our new order release simulation the congestion occurs at the release buffer. We exactly observe what is to be expected. Our new release mechanism does not increase the output (it does not create more production orders or capacity to process them). The new approach gives the system some self-regulation. If the incoming production orders temporarily exceed the outgoing rate, the number of orders in the release buffer increases. As soon as the arrival intensity decreases, the number of orders in the release buffer will stabilize. Furthermore, if the outgoing rate exceeds the incoming rate in the following period, we observe the pull function. The system will pull and the number of orders in the release buffer will decrease. We observe this effect in Table 5.16, as the larger range of order intake. If we observe that the increasing number of orders in the release buffer is no longer controllable, we are forced to take a tactical decision. In order to solve the congestion, we should either decrease input or increase output. The current release approach ignores this problem and it deteriorates our operational performance. The self-regulating effect can be strengthened by applying flexible capacity. This and the tactical decision to be taken are discussed in the next chapter.

5.5.5 Return on investment

There is of course also a financial aspect involved. If we limit the allowable workload on a tactical level, we allow a maximum allocation of resources for order intake. By allocating all capacity, we commit ourselves to more customer due dates. Because of the inherent uncertainty, the probability and related cost of extra capacity should not be forgotten. Neither should the loss of goodwill if customer due dates are not met. Therefore we recommend to make the trade-off in favor of lead time control (guaranteeing the MRP off-set lead time), rather than return on investment.

[13] Simulation runtime considerably increases if we are to program a workload update every 5 minutes, like the system behind the current television screens.

Important to keep in mind; we plan at 80%, we work for 100%. This means the capacity reserved for uncertainty will not go to waste if it is not needed. If the workload is completed without extra capacity, remaining capacity can be used for several purposes. Accepting short term customer orders, catching up with backlog or working ahead of schedule to avoid backlog when capacity is more critical.

5.6 Conclusions and limitations

First and foremost, the proof of concept is found. Based on the simulation results we conclude that the new approach attains better lead time and MRP control (timely completion of orders). The implication for Company X is that this approach is an appropriate operational planning tool. The simulation results confirm that objectives 3 and 4 from section 4.1 can be accomplished by implementing the double authorization check. We have a controlled MRP lead time, while the production orders flow efficiently through the facility. Numerically speaking, we observe a decrease of 15% in average lead time, based on this system's average of 3⅛ production steps in a routing. As there is no such thing as a typical routing for Company X, we cannot express the improvement in a number with certainty. This was however not the intention of this simulation study.

The choice of allowable workload is a trade-off between enough orders in queue to give our new approach possibilities to choose from and limiting the number of orders to reduce the risk of variability in processing time. Based on the result of the simulated example setting, we conclude that the combination of the simulated processing time and an off-set lead time of one day already restricts the option of choice. If we limit the workload even more, there remains no choice to start the best suitable order. As a result the lead time increases. An off-set lead time of two days allows more orders in queue to choose from. An allowable workload limited to 80% does show to be beneficial, because we limit the possibility of a problem when actual exceeds planned processing time. The actual choice on a tactical level remains an arbitrary one. We suggest to plan against 80%, work for 100%. Any remaining capacity will not be wasted. It will be put to good use, for example to accept short term customer orders.

5.7 Summary

As a proof of concept for the new order release method, we use a simulation study. In the simulation model we use a virtual factory. We have modeled six capacity groups; bending, robot bending, welding, stainless steel welding, painting and assembling. Production orders are randomly generated and follow 12 different routings along the capacity groups. Each production order has a randomly generated processing time for each of its process steps, based on historic data. We have also included a planned processing time margin of error between -25% and +25%. The interarrival time between orders is 46 minutes on averages with a standard deviation of 15 minutes. These input parameters create a varied production order and work load mixture to put our new release approach to the test. We measure the performance on three different criteria:

- The average lead time and its variance;
- The timely completion at a capacity group, i.e., the ability to control the MRP schedule;
- The average throughput and its variance.

Firstly, we study the effect of a limited allowable workload under different levels of variability for the new order release approach (double authorization). This is to account for the inherent uncertainty in the estimated processing times. The results show the choice is a trade-off. On the one hand, allowing enough work for the release approach to have the option of choosing a suitable next order to release. On the other hand, limiting the workload to decrease the probability of requiring more processing time than planned.

Secondly, we simulate and compare the current order release against the new order release. In the current situation, production orders in queue are ordered based on earliest release date (MRP) only. The new release method has a double authorization based on earliest release and on available capacity in the next capacity group. We apply a one day off-set lead time, as is the current case in MRP, and a two day off-set lead time. More than two days is not acceptable, as in increases the scheduled due date by too much.

With the simulated results we have a proof of concept. The new order release method outperforms the current one. We conclude that the application of an off-set lead time of two days in combination with an allowable workload of 80% of the total capacity creates the most preferable performance under the simulated circumstances. It shows a promising 15% reduction in lead time (although not the highest reduction). The variability in lead time is also reduced. It performs as good as the current situation based on the order throughput. It scores the best on its ability to keep up with the MRP schedule, making it very suitable as a method to control the operational order flow. We cannot with certainty translate the relative reduction in lead time to Company X, due to the fact that the simulated scenarios are not strictly representative.

The intention is to plan against 80% capacity, work for 100%. This is only to guard against unforeseen circumstances. If the workload is finished within the planned time, the unused capacity can be used for short term customer orders or backlog. We do not expect a negative impact on financial performance.

6. Planning structure

This chapter answers the fourth research question *"How could Company X structure its planning function?"* In chapter 4 we have introduced an approach to control the production flow in the manufacturing facility at Company X. It is based on overlapping loops of capacity groups and order release authorization between them. In chapter 5 the new approach is compared to the current approach using a simulation model and it demonstrates the benefit. The approach is not a stand-alone solution to the challenges at Company X. We already discussed how the approach is to be combined with a higher-level tactical MRP schedule. For the solution to be successful, it requires careful embedding within the organization. In this chapter we expand on the subject of embedding by introducing a hierarchical framework for planning as a function as introduced in the theoretical framework in section 2.5. We try to make the function of planning more transparent and understandable. The section 6.1 defines what the objective of planning at Company X is and how the entire function can be divided into more manageable pieces, or subproblems. In sections 6.2 through 6.4 these subpoblems are discussed. Subsequently, we can identify where the new approach fits within the planning function and how it can be incorporated to increase to chance of success. Last but not least, the provided transparency can help understanding at which level a particular problem should be solved, the implications for the other levels are and how to translate a solution to lower levels. Point all noses in a common direction as it were.

6.1 Objective of planning

The objective of a well-structured planning function is matching supply and demand. Supply being the available capacity, materials etc., and demand being the customer needs. The added value is all about supporting the manufacturing process to deliver the right product, the right quantity at the right time (at the best possible cost). A planning function forms the linchpin in a supply chain where the needs of the customer are translated into a manageable (cost) effective sequence of processes to meet that need (Figure 6.24). This is of course from a planning point-of-view, the picture probably looks different from the perspective of another department and we certainly do not mean to say that planning is more important than other functions. This research however, is from a planning perspective and therefore planning as a function is the center of attention.

To divide the large, complex planning problem into smaller subproblems, the hierarchical planning framework is used as introduced in section 2.5. The planning function at Company X is predominantly of operational nature at the moment, the framework also serves as an aid to take planning as a function to a higher, more efficient level. The hierarchical framework for Company X is displayed In appendix C, consisting of a strategic, tactical and operational level.

The following sections will describe these levels within the context of Company X. We would usually start at the top, at a strategic level. After all, the decisions taken on a higher level are the boundaries for the levels beneath. From a practical point-of-view however, it is easier to start bottom-up, i.e., the operational level. The objectives, decisions and considerations become more difficult when we go up the levels. Besides, if we want to embed the new approach in the organization it makes more sense to start from its own planning level.

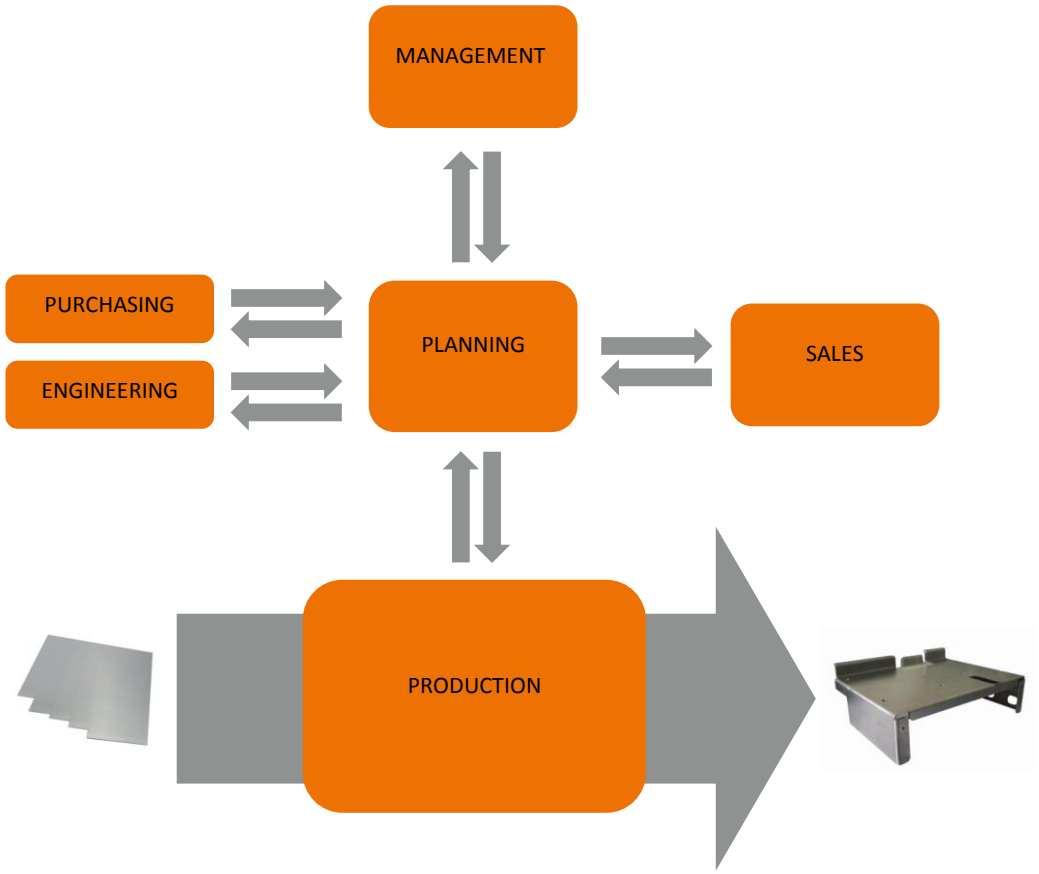


Figure 6.24: Position of planning in the bigger picture

6.2 Operational level

6.2.1 Objective

We start bottom-up in Appendix C, at the operational planning level where we apply the new approach of paired capacity groups with available capacity checks. Our objective is to complete the required workload as soon as possible, with the available capacity. In other words, we are controlling the order flow pattern. The workload to be finished, is based on the predetermined order pool. A capacitated Master Production Schedule (MPS). For the planning function, the operational level is about monitoring and adjusting. We monitor the progress of production orders, signal any delays and take action accordingly. The pressure for control at this level should be minimal, we apply the capacity loops especially because it redirects the order flow away from the problem areas.

6.2.2 Responsibility

The responsibility at this level first of all lies with the foremen and it is their duty to monitor the process. The sequence in which the production orders are finished within the time frame is left to their expertise and judgment. Gains in throughput time because of shorter changeover times are of course greatly appreciated. The bottom line is guaranteeing the workload is finished within the time frame. Secondly, the (operational) planner also has the responsibility to monitor the progress and signal any problems.

6.2.3 Decision process and practical approach

The problem owners of this planning level are the (Operational) planner and the foremen on the shop floor. See Appendix D for an organization chart. Together they are responsible for completing the production orders. The decisions to be taken revolve around solving any bottlenecks.

The bottlenecks present themselves when a capacity group cannot start any production orders, because its successor repeatedly has no capacity available. At such moments, it is up to the foremen to reallocate flexible capacity in a useful manner. The plan meeting every morning at the central screen in the factory offers a good opportunity for the monitoring process and any upcoming problems regarding bottlenecks can be solved by discussing the availability of flexible capacity between capacity groups.

6.3 Tactical level

6.3.1 Objective

This level of planning is all about the balance between supply and demand. How much capacity do we require and how much do we have (left) to allocate. This is what we call workload balancing (Appendix C). The most important thing to remember is that this is the level where we determine what we are going to make, when it is due and how much capacity it requires.

6.3.2 Responsibility

The responsibility at this level lies with the (tactical) planner and Director of Operations. It is their duty to make sure that the workloads for the capacity groups match the available capacity.

6.3.3 Decision process and practical approach

The decision making process revolves around roughly allocating capacity over the capacity groups, signaling shortages and take action accordingly. These actions could be, for instance, temporarily expanding capacity, rescheduling production orders or outsourcing. Most important to realize; as long as the solution restores the capacity balance and it is recorded in a consistent and correct manner, the chosen solution is subordinated to the solution of the problem. Consider the acceptance of a new order and its subsequent shortage of capacity. If it is decided to solve this problem with one Saturday's worth of overwork, the extra capacity must be translated into resulting capacity on an operational level. In other words, on the screen of the capacity group, the extra capacity is visual for the employees and it is electronically visible for the capacity availability check. Right after the operational meeting on the work floor, we propose to have a tactical plan meeting at the office of the Director of Operations. In this meeting, we put the workload for all capacity groups on screen, and review the implications. Sales representation is preferable, but not necessary if there are no new potential orders. When a potential order presents itself, a tactical meeting should take place in the aforementioned location and setting. The customer is represented by the appropriate sales employee. During this meeting we review the current state of capacity, assess what the impact on capacity is if the orders is accepted and problems are identified, solved and translated into operational adjustments.

6.4 Strategic level

6.4.1 Objective

The first and leading strategic objective is to guarantee the continuity and sustainability of the company. A strategic initiative is acceptable if it contributes to the business. This means the initiative should lead to (i) long term profitability, (ii) improvement of the quality of product and/or work and (iii) (practical) execution. Production planning specific, it must find a profitable and sustainable match between market demand and factory wide ability to supply this demand. At all levels, every underlying objective and decision should be subordinate to this main objective. The complete list of objectives, in hierarchical order:

1. Continuity and sustainability
2. Shorter, more reliable lead times

3. Reduced, controlled operating cost

All decisions have consequences regarding objectives at the strategic level itself, but also for lower planning levels. Conflicting objectives are to be solved by the management of Company X.

6.4.2 Problem owner

The Director of Operations is the problem owner and responsible party with decision-making authority. Required information, support and feedback in the decision-making process is given by the (tactical) planner and the sales manager.

6.4.3 Decision process and practical approach

Demand-driven production; Company X produces finished products based on demand only (pull). In traditional MRP (push) production, machine efficiency is a leading objective. These two concepts are mutually exclusive, as machine efficiency is mostly dependent upon large batch sizes of aggregated demand. Company X would be manufacturing parts earlier than demand requires, directly defying the strategic decision to produce demand-driven. This means that capacity is unnecessarily taken up, leading to larger lead times. It conflicts with the objective to reduce lead times as well.

Clustered production; several finished products will be manufactured in standard numbers, based on the number of products that fit in the standard packaging for delivery. All required parts needed for this quantity are clustered together per type of raw material. A clustered order is send through the factory as a cluster, all parts put together in a cart.

Flexible capacity manufacturing; because Company X faces uncertainty in demand, it would like to apply flexible capacity to be able to process a peak in demand. Flexible capacity does not mean total capacity increases, but rather means capacity can be temporarily interchanged across departments to meet required capacity. To treat flexible capacity as an inexhaustible tool which requires no control mechanism would be wrong, rather it adds a new layer of complexity for the planning function and therefore this process requires some careful considerations starting at a strategic level:

- Document employee competences
- Determine total capacity per department
    - Regular capacity
    - Maximum capacity
    - Assessment of adequacy
- Training flexible employees
    - Personal incentive
    - Financial incentive
    - Maintain employee competences
- Performance measurement
    - Different employee, same quality

External capacity; what if the regular and flexible capacity are not sufficient? Define rules or guidelines to allow overwork, outsourcing or rescheduling. For instance based on preferred customer list.

Capacity and occupation of the planning department; if Company X wishes to develop the function of planning into something more than just a regulating function of short-term production, it should consider adding extra capacity to the department itself as well. This is also worthwhile to

avoid the function being dependent upon just one person. Furthermore, facilitating the department with appropriate knowledge and tools is a strategic matter as well.

6.5 Summary

For the solution from chapter 4 to be successful, it requires embedding within the organization. We already discuss how the approach is to be combined with a higher-level MRP schedule. In this chapter we expand on the subject of embedding by introducing the hierarchal framework, where we define a strategic, tactical and operational level. We aim at providing transparency and help understanding at which level a problem should be solved. The objective of a well-structured planning function is matching supply and demand, like required and available capacity. The added value is about supporting the manufacturing process to deliver the right product, the right quantity at the right time. The planning function at Company X is predominantly of an operational nature, we intend to add decision level to make the function more efficient.

We start bottom-up, with operational planning. This is where our new order release approach with capacity loops is applied. The objective is to complete the entire workload as quickly as possible. The responsibility at this level lies with the foremen. Bottom line is guaranteeing a timely completion of the workload within the time frame, e.g., by shifting resources between capacity groups. The (operational) planner has the responsibility to monitor to progress and signal problems. The decisions to be taken revolve around solving temporary bottlenecks. The daily plan meeting at the central screen offers a good opportunity to discuss the operational problems.

At a tactical level, the responsibility lies with the (strategic) planner and the Director of Operations. The decision process revolves around balancing the workload over the capacity groups; how much capacity is reserved and how much is left to allocate. Shortages are to be signaled, solved and translated into operational adjustments. Right after the operational plan meeting, we propose to have a tactical plan meeting. When a potential order presents itself, a tactical meeting should take place to review the current state of capacity and assess the impact if the new order is accepted. When a problem is identified, it is to be solved and translated into operational adjustments. Potential orders without at least roughly estimated capacity requirements should not be accepted.

The leading objective on a strategic level is ensuring the continuity of the business. Production planning specific, this means finding a profitable and sustainable match between demand and factory wide ability to supply this demand. At all underlying levels, the objectives and decisions should be aligned with and subordinate to the strategic objectives and decisions. The Director of Operations is responsible at this level of planning. Strategic objectives are translated into tactical and operational objectives to manage and control performance. At this level, it is also important to develop, maintain and support the tools with which the underlying levels are to perform their tasks.

7. Conclusions, recommendations and future research

The main objective of this research is to find an answer to our main research question. Before we can answer this question, we first answer the sub questions.

7.1 Conclusions

"What has been written in academic literature regarding the problems Company X faces?"

We can conclude that for an ETO/MTO organization like Company X, neither a pure push (MRP) or pull (Lean) approach is suitable. MRP planning and control is too deterministic and inflexible, whereas Lean lacks the ability to manage highly variable product and demand behavior and the associated information. The literature describes several possibilities where the two approaches can be successfully combined. The most promising combination we have identified is called POLCA, a vertically integrated approach. It combines push and pull on different planning levels.

To make the planning function more transparent and manageable, our study of literature has provided us with a hierarchical planning framework. This framework is meant to divided the large, complex problem of planning into more manageable sub problems to solve sequentially. We distinguish strategic, tactical and operational. Moving down the levels, the planning horizon decreases while the level of detail increases. Higher levels require less detailed information as opposed to the lower, more sophisticated methods. Regardless of a push, pull or hybrid approach and which planning method is used, integration is the most important to make it successful. This integration refers to the different levels and their objectives, as well as planning as a function. The function must be embedded within the organization and processes should be integrated with the other functionalities. To make the planning function easier to understand and create more support, we can focus on the level at which planning and an employee's activities intersect.

In the literature we find queuing theory to be suitable to predict lead time and explain why waiting time occurs and how it increases. We use it to partly answer to next sub question.

"What does the production process look like and what is the current performance?"

At the moment, the internal lead time of Company X is a total of 13.8 days over five departments. This lead time has a standard deviation of 6.7 days. 57% of the internal lead time is waiting time. Reducing active processing time is a possibility to reduce lead time, however less likely to increase the ability to control it. Another important opportunity is clustering similar process steps, thereby decreasing the number of unique steps in the routing. Each of these steps is scheduled backwards in time with an MRP off-set lead time, reducing the steps to schedule directly reduces lead time. Both opportunities are beyond the responsibilities of planning as a function. From a planning perspective, our best possibility to control and reduce lead time is to decrease the waiting time.

In order to control and reduce lead time, we have attempted to analyze the waiting time using queuing theory. Waiting and lead time can theoretically be explained by arrival intensity (λ), processing rate (μ) and the variation related to them. When this is the case, we can subsequently control it as well by manipulating these variables. Unfortunately, the analysis pointed out that this theory does not explain the occurrence of waiting time for Company X. The most important basis for this result is related to the characteristics of Company X's business.

Company X is an MTO/ETO supplier for OEMs. It means Company X does not have its own products. It primarily manufactures customer specific products on order. As a direct result, the upcoming demand, product mix and required capacity are difficult to predict, if not impossible. The inherent variability in this type of business and the flexibility Company X applies to meet the challenging demand, is most likely the reason why queuing theory does not apply.

In the end, we identify that the situation at Company X and the planning approaches described in literature leave us with a gap to overcome. In particular the type of variability characteristic for organizations operating in "unknown unknown" environments offers us a challenge to look for an interesting new approach.

"How could Company X implement an approach to control and reduce lead time?"

Based on the literature review, queuing theory and analysis of the production process, we conclude it is impossible to schedule and control the production process beforehand. The variability is too high. We are looking for an approach to cope with the inherent uncertainty Company X faces. Bottom line, we look for a method to:

- Assess required versus available capacity and schedule production orders accordingly (push);
- Ensure the schedule is finished on time, meanwhile achieving an efficient order flow (pull).

With Company X's ERP system, the first part of the approach is in place. The literature research identified the POLCA method as a suitable combination of a push-pull hybrid approach. The POLCA concept of multiple, routing-specific overlapping loops of capacity groups with capacity availability authorization, is a pull-based approach. It limits the workload in an operational setting to ensure timely completion. Meanwhile, it identifies which capacity groups have available capacity within the MRP off-set lead time, thereby creating a demand driven priority in the preceding queues to accommodate a flow in the orders. Both the push and pull condition are met, so we can conclude the approach answer our question how we can control production. With a simulation study of a hypothetical example, we show how the use of this approach outperforms the current situation of order release. It reduces lead time and increased the ability to keep up with the MRP schedule.

We propose to use the existing network of television screens for electronic authorization. Because Company X is a ETO/MTO manufacturer, the routings and loops differ over time. The use of electronic capacity loops on television screens helps us to cope with change. We also try to take advantage of the screens as much as possible by creating a dynamic, real-time system, capable of quick response. If available capacity is below the maximum workload for a capacity group, it signals availability. The screen of a capacity group visualizes the production orders in queue, prioritized on MRP earliest release and capacity availability at the succeeding capacity group.

We limit the allowable workload on a tactical level, i.e., apply a maximum to the allocation of resources for order intake. The simulated hypothetical situation cannot be used as a typical, representative situation for Company X, the choice for a maximum allowable workload remains an arbitrary one. Our recommended choice; plan against 80%, work at 100%. This 80% is based on academic literature. By allocating all capacity, we commit ourselves to more customer due dates. Our experience with the inherent uncertainty for Company X has learned us there is a realistic possibility that we will require extra capacity to keep up with committed due dates. The cost of extra capacity should not be forgotten. Neither should the loss of good will if customer due dates are not met. Therefore we recommend to make the trade-off in favor of lead time control (guaranteeing the MRP off-set lead time), rather than return on investment. Any unused capacity will not be wasted, as it can be put to good use to accept short term customer orders or solve backlog.

The simulated results on allowable workload do give us the insight of a trade-off on an operational level. A trade-off between enough production orders in queue for our system of capacity loops to make a good choice and restricting the amount of orders to reduce variability.

The operational pull approach limits the WIP, ensuring a controlled lead time. The push approach helps us with a more clear prioritization. Production orders are not started before the earliest start date, therefore no unnecessary effort is spent working ahead of schedule. Furthermore, the pull approach is a logical extension of the currently applied network of television screens and the demand priority is unambiguous. It fits well with the objective to find a transparent and understandable planning method. An additional advantage is the fact that we require no further investment to implement to solution, except for the time required the reprogram the television screens. Introduction of the concept has resulted in a generally positive attitude both at the shop floor level as well as on a management level. Because it is an extension rather than a radically different approach, the probability for successful implementation and acceptation is higher.

"How could Company X structure its planning function?"

Company X can structure its planning function by dividing the decisions over different levels. With this division over different levels, we can identify at which level a particular problem should be solved, who has the responsibility and which information is required to make a well-informed decision. We know planning to be a large, complex function. In the literature review we learned that planning can be benefit from division into strategic, tactical and operational planning. It is beneficial because it helps us in dividing the entire complex problem into more manageable sub problems. Subsequently, we determine the objectives on all levels and align them with objectives on a higher level by careful integration. This division of decisions over different levels enables us to create a more transparent and understandable decision-making process. When this is done correctly, we ensure that everyone in the organization understands what their objective is and how it fits within a broader perspective. We aim to create overarching objectives for horizontal departments and vertical organizational levels. This decreases the chances of sub optimization. The new order release approach with capacity authorization between capacity groups is an operational method and we concluded that this approach is not a stand-alone solution. With the use of our hierarchical planning framework, we can embed the solution within the organization at the right level (operational) and manage the performance and expectations.

7.2 Recommendations

With the answer on the research questions, the results from our simulation study and its proof of concept, we answer the main research question and present our recommendations for Company X.

"How could the planning function at Company X be improved so as to support the company's objective of controlling and reducing internal lead time?"

Company X could improve its planning function by dividing its production resources into capacity groups and connect these to each other in virtual capacity loops using the existing network of television screens. We recommend applying the push-pull hybrid approach for order release, to control the internal lead time between the capacity groups. Our simulation study has given us a proof of concept. It is possible to reduce the lead time, because the amount of waiting time is limited. (Four case studies in academic literature show a lead time reduction between 22 and 70%.) The double authorization approach will bypass capacity groups with a temporary capacity problem on an operational level, thereby avoiding unnecessary waiting time. We are careful to emphasize the words

temporary and operational. The recommended use of this order release method will not solve a structural capacity shortage. Nor does it guarantee the solution to an operational bottleneck to coincide with the solution to a tactical bottleneck, i.e., bypassing an operational bottleneck does not guarantee that the bypassed production orders are still completed according to our tactical MRP schedule. This also depends on adequate response, applying flexible capacity between capacity groups for instance. As with any planning method, its performance will not be optimal when it is not used properly. If we continue to schedule more work than the capacity groups can handle, our recommended approach avoids overloading and all related problems. It does not create a higher output. We require careful workload balancing on a higher planning level, therefore we recommend the use of a hierarchical planning framework.

When we purely concentrate on the ability of the approach to reduce lead time, we conclude that other factors are of equal importance. The number of unique processing steps in a routing is the determining factor as each step is scheduled backwards with one unit off-set lead time. Clustering short processing steps into one is especially effective to decrease lead time, as it cuts out multiple units of off-set lead time. The main objective to finish 80% of the workload within five work days, will depend as much on 80% of the work actually requiring five processing steps as it does depend on our new order release approach.

In summary, we recommend the following:

- Divide factory resources in capacity groups, electronically linked using the network of television screens;
- Apply the double authorization approach as an operational order release mechanism between capacity groups;
- Use MRP to review, balance and schedule the workload per capacity group (plan against 80% capacity, work at 100%);
- Use a hierarchical planning framework to integrate the operational method in the planning function and to take the function to a higher level;
- Add capacity to the planning function itself.

The implementation of the new approach presented in this research falls outside the scope of the thesis. This research and the results are transferred to Company X's Business Engineer to take the lead in the implementation project. The implementation will fall under the responsibility of the Director of Operations. A roadmap of the implementation steps is shown in Appendix E.

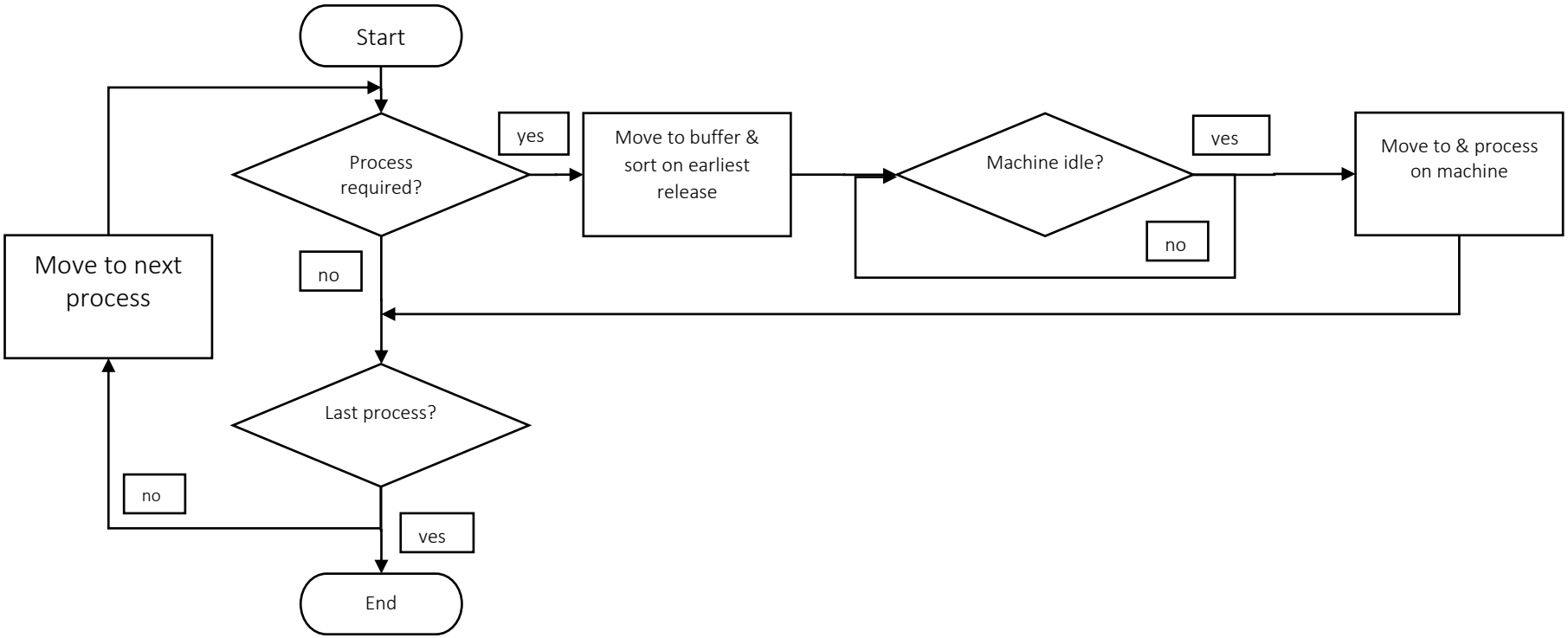
7.3 Future research

The focus of this research is on production planning and control. For ETO/MTO organizations, an important part of the process is of course the design of a product and the process of preparing it for release into production. This part of the process also provides opportunities to further improve the performance of Company X and contribute to the main objectives set by Company X's management.

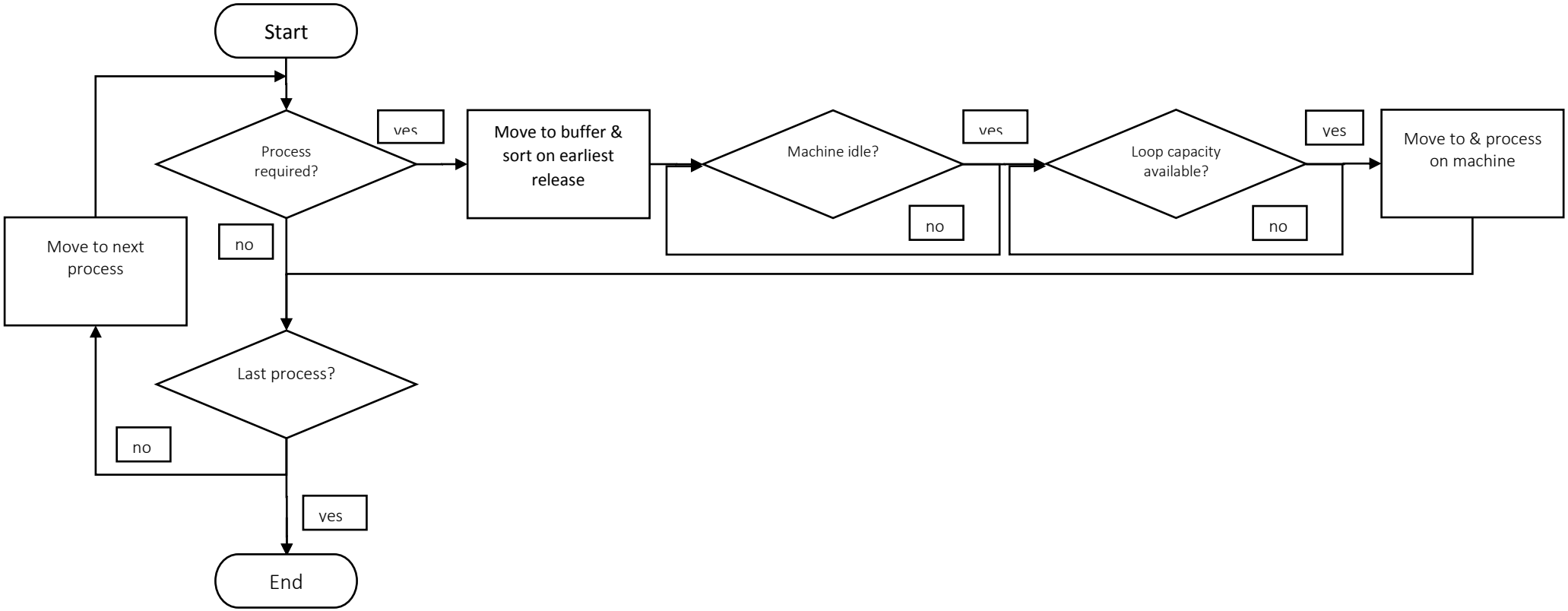
We recommend to invest time and resources in researching the possibilities to improve the engineering and pre-production phase. In an organization where the ability to design and produce everything the customer asks for, it is important to control how much time is spent in this phase and manage what the time is spend on. We regularly experience pressure on planned production time because the design phase exceeds the planned time. We can choose to engineer until the product satisfies the customer or devote extra time to engineer the product until it is easier to produce with less problems in production. In extension, we recommend to invest time and resources into the

possibility to convert initial work into repeat business as quickly as possible. Experience from the planning department learns us that Company X has a stream of repeat business which accounts for steady income at relatively low control and cost. Effort to turn an initial order stream into an easily manufactured, easily controlled order flow will likely pay off. We can think of design for manufacturing initiatives and such, to reduce the variance between products and improve overall quality. Or attempt to strengthen the customer relationship through (mutually) beneficial design changes and become first choice supplier.

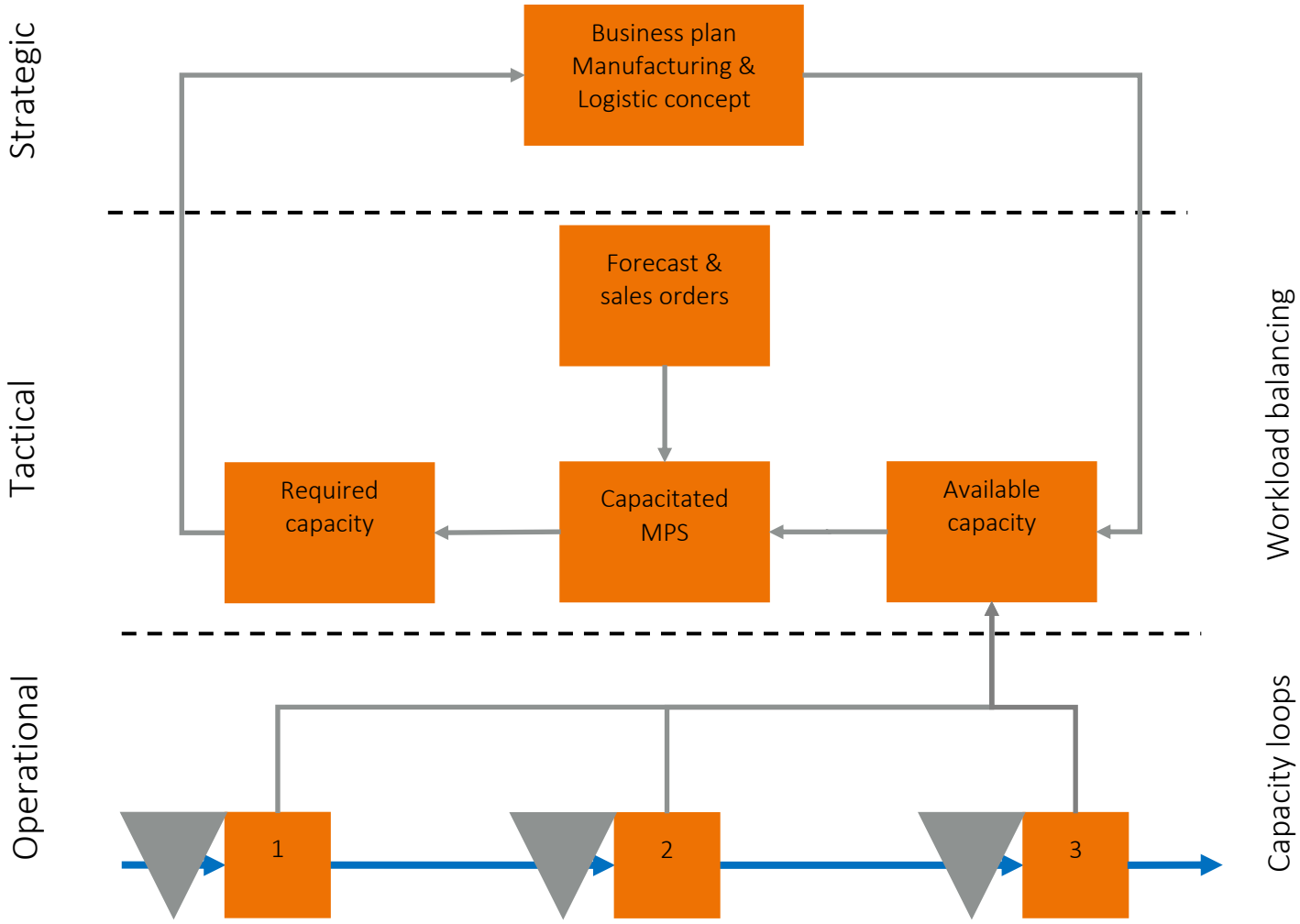
Appendix A: Order flow current order release



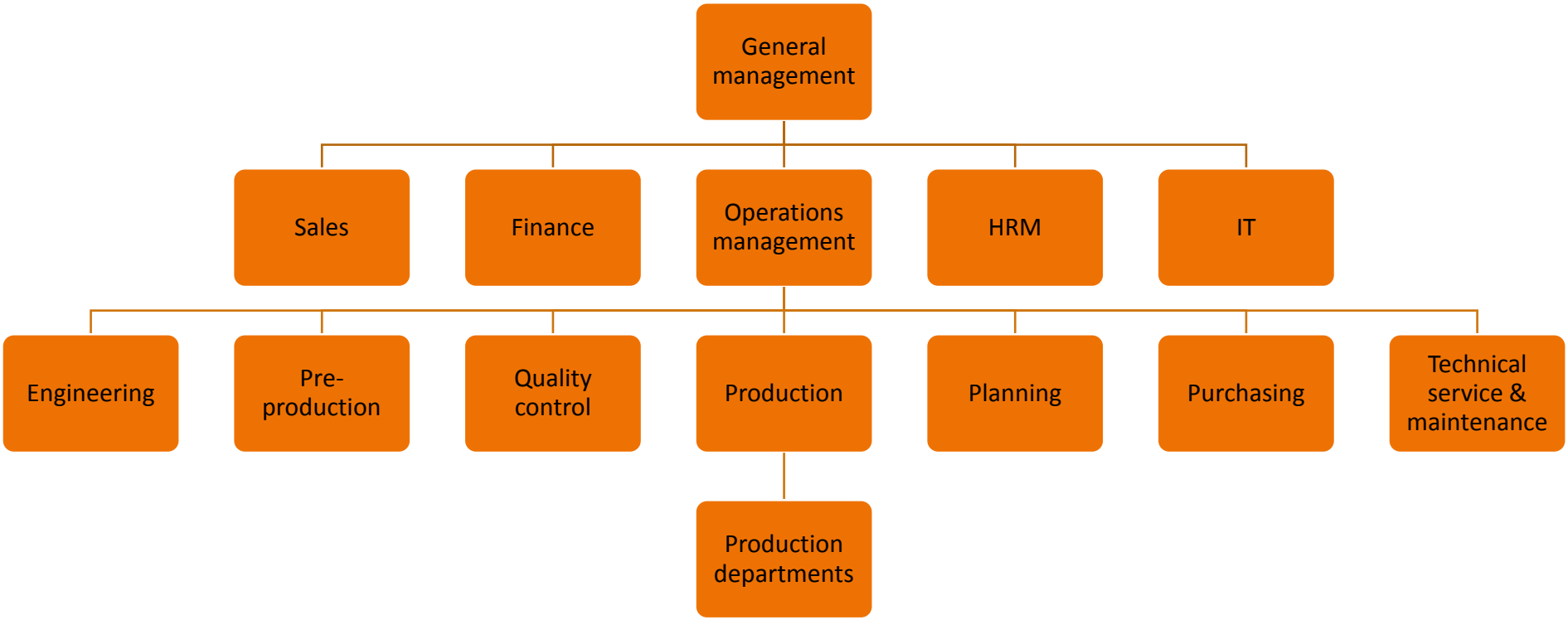
Appendix B: Order flow new order release



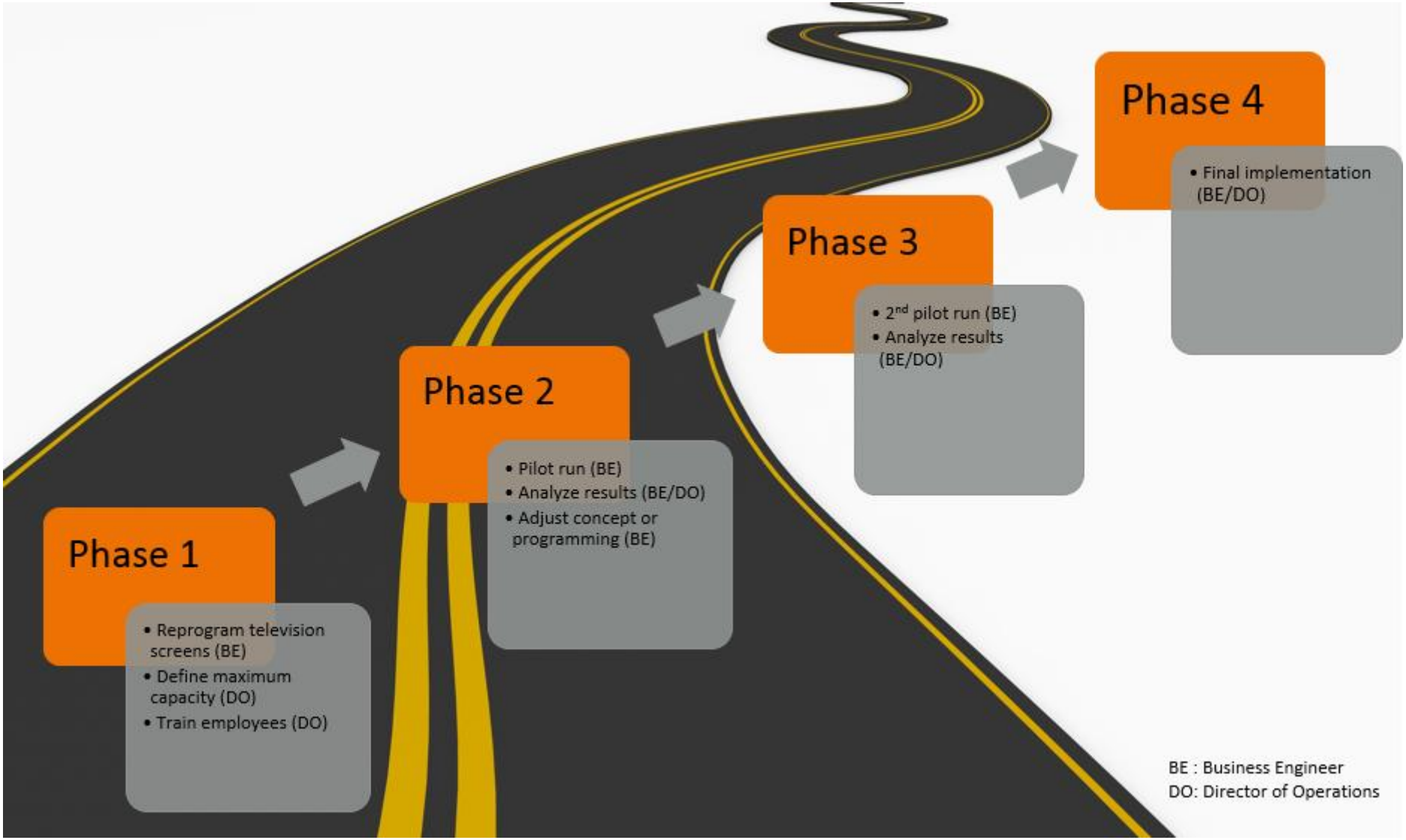
Appendix C: Hierarchical planning framework



Appendix D: Organization chart



Appendix E: Roadmap



References

Adenso-Díaz, B., & Laguna, M. (1996). Modelling the load levelling problem in master prodcution scheduling for MRP systems. *Internatonal Journal of Production Research*, 483-493.

Beach, R., Muhlemann, A., Price, D., Paterson, A., & Sharp, J. (2000). A review of manufacturing flexibility. *European Journal of Operational Research*, 41-57.

Billington, P. J., McClain, J. O., & Thomas, L. J. (1983). Mathematical Programming Approaches to Capacity-Constrained MRP Systems: Review, Formulation and Problem Reduction. *Management Science*, 1126-1141.

Bish, E. K., Muriel, A., & Biller, S. (2005). Managing Flexible Capacity in a Make-to-Order Environment. *Management Science*, 167-180.

Bott, K. N., & Ritzman, L. P. (1983). Irregular Workloads with MRP Systems: Some Causes and Consequences. *Journal of Operations Management*, 169-182.

Browne, J., Dubois, D., Rathmill, K., Sethi, S. P., & Stecke, K. E. (1984). Classification of flexible manufacturing systems. *The FMS Magazine*, 114-117.

Carlsson, B. (1989). Flexibility and the theory of the firm. *International Journal of Industrial Organization*, 179-203.

Chin, C.-S. (2009). Queueing Theory and Process Flow Performance. *Annual Conference of the International Group for Lean Constructions* (pp. 247-255). Taiwan: Pingtung.

Company X B.V. (2012a). *Procedure Planning.* Location X: Company X B.V.

Company X B.V. (2012b). *Werkinstructie Herstelmethoden in Company Z.* Location X: Company X B.V.

Company X B.V. (2014). *Werkinstructie uren scannen.* Location X: Company X B.V.

Corry, P., & Kozan, E. (2004). Meta-heuristics for a complex push-pull production system. *Journal of Intelligent Manufacturing*, 381-393.

Ebadian, M., Rabbani, M., Torabi, S., & Jolai, F. (2009). Hierarchical production planning and scheduling in make-to-order environments: reaching short and reliable delivery dates. *International Journal of Production Research*, 5761-5789.

Euwe, M. J., & Wortmann, H. (1997). Planning systems in the next century (I). *Computers in Industry*, 233-237.

Frank, B., Neumann, K., & Schwindt, C. (1997). A capacity-oriented hierarchical approach to single-item and small-batch production planning using project-scheduling methods. *OR Spektrum*, 77-85.

Gelders, L. F., & Van Wassenhoven, L. N. (1982). Hierarchical Integration in Production Planning: Theory and Practice. *Journal of Operations Management*, 27-35.

Gerathy, J., & Heavey, C. (2005). A review and comparison of hybrid and pull-type production control strategies. *OR Sprectrum*, 435-457.

Hans, E. W. (2001). Resource loading by Branch-and-Price Techniques (ch.1 of Ph.D. thesis). *Universiteit Twente*, 1-24.

Hans, E. W., Herroelen, W., Leus, R., & Wullink, G. (2007). A hierarchical approach to multi-project planning under uncertainty. *The international Journal of Management Science*, 563-577.

Hopp, W., & Spearman, M. (2000). *Factory Physics.* New York: Irwin/McGraw-Hill.

Houy, T. (2005). ICT and Lean Management: Will They Ever Get Along? *Communications & Strategies*, 53-75.

Jonsson, P., & Matsson, S.-A. (2002). The use and applicability of capacity planning methods. *Production and Inventory Management Journal*, 89-95.

Karmarkar, U. S. (1986, June). Working Paper Series No. QM8615. *Integrating MRP with Kanban/Pull Systems*. The University of Rochester.

Krishnamurthy, A., & Suri, R. (2009). Planning and implementing POLCA: a card-based control system for high variety or custom engineered products. *Production Planning & Control*, 596-610.

Kulatilaka, N., & Marks, S. G. (1988). The Strategic value of Flexibility: Reducing the Ability to Compromise. *The American Economic Review*, 574-580.

Law, A. M. (2007). *Simulation Modeling and Analysis.* New York: McGraw-Hill.

Liberatore, M. J., & Miller, T. (1985). A Hierarchical Production Planning System. *Interfaces*, 1-11.

Markus, M. L., Axline, S., Petrie, D., & Tanis, C. (2000). Learning from adopters' experiences with ERP: problems encountered and success achieved. *Journal of Information Technology*, 245-265.

McKenzie, E. (2009, november 20). *Lean vs Six Sigma: What's the Difference?* Retrieved from Ultimus Enterprise Solutions: http://www.ultimus.com/Blog/bid/33875/Lean-vs-Six-Sigma-What-s-the-Difference

Mehrabi, M., Ulsoy, A., & Koren, Y. (2000). Reconfigurable manufacturing systems: Key to future manufacturing. *Journal of Intelligent Manufacturing*, 403-419.

Moon, Y. B., & Phatak, D. (2005). Enhancing ERP system's functionality with discrete event simulation. *Industrial Management & Data Systems*, 1206-1224.

Moscoso, P. G., Fransoo, J. C., & Fischer, D. (2010). An empirical study on reducting planning instability in hierarchical planning systems. *Production Planning & Control*, 413-426.

Murthy, D., & Ma, L. (1991). MRP with uncertainty: a review and some extensions. *International Journl of Production Economics*, 51-64.

Nauhria, Y., Wadhwa, S., & Pandey, S. (2009). ERP Enabled Lean Six Sigma: A Holistic Approach for Competitive Manufacturing. *Global Journal of Flexible Systems Management*, 35-43.

Olhager, J., & Östlund, B. (1990). An integrated push-pull manufacturing strategy. *European Journal of Operations Research*, 135-142.

Öztürk, C., & Örnek, A. M. (2012). A MIP based heuristic for capacitated MRP systems. *Computers & Industrial Engineering*, 926-942.

Plenert, G. ((1999)). Focusing material requirements planning (MRP) towards performance. *European journal of Operational Research*, 91-99.

Portioli-Staudacher, A., & Tandardini, M. (2012). A lean-based ORR system for non-repetitive manufacturing. *International Journal of Production Research*, 3257-3273.

Powell, D., Alfnes, E., Strandhagen, J. O., & Dreyer, H. (2012). ERP Support for Lean Production. *unknown*, 115-122.

Powell, D., Alfnes, E., Strandhagen, J. O., & Dreyer, H. (2013). The concurrent application of lean production and ERP: Towards an ERP-based lean implementation process. *Computers in Industry*, 324-335.

Powell, D., Riezebos, J., & Strandhagen, J. O. (2013). Lean production and ERP systems in small- and medium-sized enterprises: ERP support for pull production. *International Journal of Production Research*, 395-409.

Riezebos, J. (2010). Design of POLCA material control systems. *International Journal of Production Research*, 1455-1477.

Riezebos, J., Klinkenberg, W., & Hicks, C. (2009). Lean production and information technology: Connection or contradiction? *Computers in Industry*, 237-247.

Slomp, J., Bokhorst, J. A., & Germs, R. (2009). A lean production control system for high-variety/low-volume environments: a case study implementation. *Production Planning & Control*, 586-595.

Steger-Jensen, K., & Hvolby, H. (2008). Review of an ERP System Supporting Lean Manufacturing. In T. Koch, *Lean Business Systems and Beyond* (pp. 67-74). Boston: Springer.

Subba Rao, S. (1992). The relationship of work-in-process inventories, manufacturing lead times and waiting line analysis. *International Journal of Production Economics*, 221-227.

Taal, M., & Wortmann, J. C. (1997). Integrating MPR and finite capacity planning. *Production Planning & Control*, 245-254.

Tempelmeier, H. (1997). Resource-constrained materials requirements planning - MRP rc. *Production Planning & Control*, 451-461.

Tenhiälä, A. (2011). Contingency theory of capacity planning: The link between process types and planning methods. *Journal of Operations Management*, 65-77.

Upton, D. M. (1994). The Management of Manufacturing Flexibility. *California Management Review*, 72-89.

Vandaele, N., Van Nieuwenhuyse, I., Claerhout, D., & Cremmery, R. (2008). Load-Based POLCA: An Integrated Material Control System for Multiproduct, Multimachine Job Shops. *Manufacturing & Service Operations Management*, 181-197.

Villa, A., & Watanabe, T. (1993). Production management: Beyond the dichotomy between 'push' and 'pull'. *Computer Integrated Manufacturing Systems*, 53-63.

Zijm, W. (1999). Towards intelligent manufacturing planning and control systems. *OR Spektrum*, 313-345.