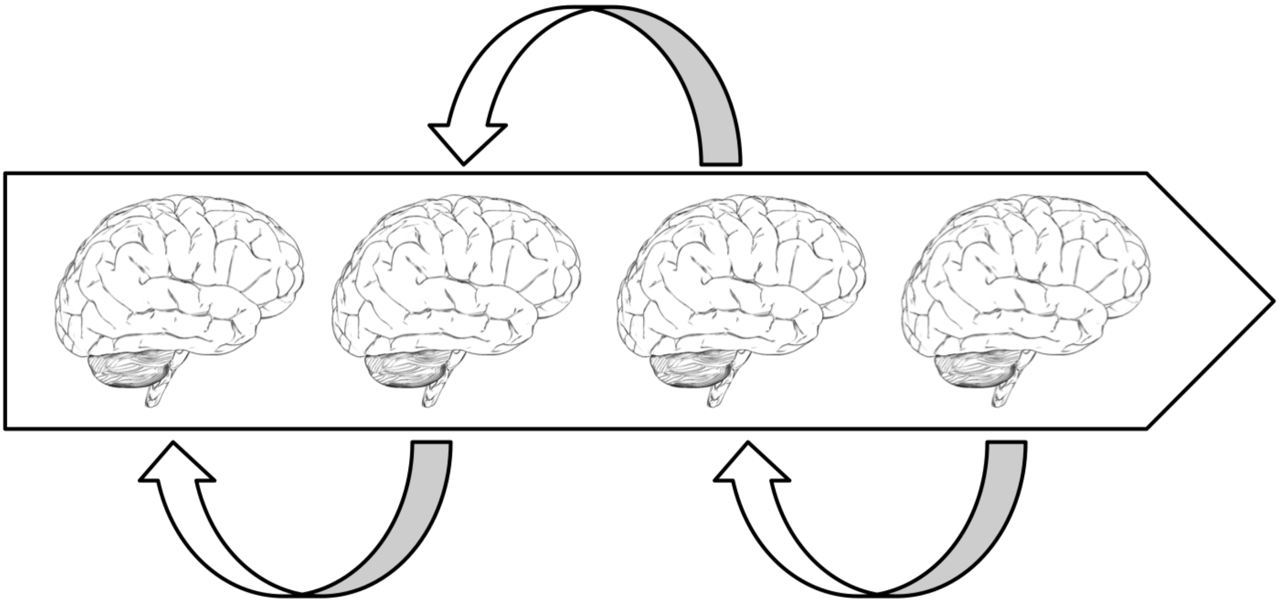

Achieving Pull-Based Control in White-Collar Settings

Reducing internal obstacles and adapting conventional control systems

Master of Science Thesis in the Master Degree Programme Quality and Operations Management

JONNY ERIKSSON
JOHAN SANDSTRÖM

Department of Technology Management and Economics
Division of Operations Management
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden, 2013
Report No. E 2013:070

MASTER'S THESIS E 2013:070

Achieving Pull-Based Control in White-Collar Settings

Reducing internal obstacles and adapting conventional control systems

JONNY ERIKSSON
JOHAN SANDSTRÖM

Tutor, Chalmers:	Ludvig Lindlöf
Tutor, company:	Måns Falk

Department of Technology Management and Economics
Division of Operations Management
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2013

Achieving Pull-Based Control in White-Collar Settings

Reducing internal obstacles and adapting conventional control systems

JONNY ERIKSSON
JOHAN SANDSTRÖM

Department of Technology Management and Economics
Division of Operations Management

Chalmers University of Technology

SE-412 96 Göteborg, Sweden
Telephone: + 46 (0)31-772 1000

Chalmers Reproservice

Göteborg, Sweden 2013

Abstract

The purpose of this thesis is to describe the preconditions needed to achieve a pull-based value stream in white-collar settings by contrasting this type of work to blue-collar settings. Existing systems for pull-based control in blue-collar settings are evaluated to assess their suitability and whether elements may be combined to increase general suitability in white-collar settings. Data was collected through studying literature and a case study was used to anchor and exemplify the findings. Interviewees comprised experts within the field and staff at Volvo Cars Customer Service as well as staff within the quality bulletin (QB) process, the company and cross-departmental sample process studied in the case.

There are three basic preconditions of pull: defined agreements, dedicated items and control. These preconditions set the stage for what needs to be present and improved. The critical principles that enable these preconditions are: leveling of the workload, standardization, and visual control. Conventional pull systems studied are: Kanban, CONWIP, Kanban-CONWIP Hybrid and POLCA. There have been successful attempts of using pull systems in white-collar settings with repetitive output. The most prominent discrepancies from white- to blue-collar settings are usually higher variability and task discretion, and thus higher complexity. According to some authors, “pure” pull systems are not even feasible when complexity is high. Others are of the belief that this complexity is due to the lack of standardization. Nevertheless, it is important to relax the assumption of order, in other words, to accept that some things have a certain amount of complexity.

A “pure” pull system, where parts are pulled from downstream demand and replenished upstream, is not possible in the QB process where output is non-repetitive. This is not an effect from non-physical flow, but from the unpredictability in output and demand. Since work arise from quality issues in the market, pre-defined solutions cannot be stored downstream in advance. The dynamic is similar to Engineered to Order processes where a degree of push and central planning is needed. To keep the process under pull-based control, work units of capacity should be pulled instead of products. Furthermore, depending on the complexity in terms of task-routing in white-collar settings, a pull system can have different levels of control and resource layout to allow for higher flexibility. Standardization is the most important effort in reducing the variability and complexity to allow for detailed control. Thus, by choosing pull system sensibly and/or by making appropriate changes to specific white-collar contexts, conventional pull systems can be as suitable for white-collar settings as for blue-collar settings.

The early stages of the QB process have higher variability and lower predictability than the later stages. A POLCA system is more suitable for the early stages, while a Kanban-CONWIP Hybrid system is more suitable for the later ones.

Acknowledgements

We would like to thank Ludvig Lindlöf, doctoral student at Chalmers University of Technology and our thesis supervisor, for being supportive and understanding. You have been a great guide from start to finish in this intense thesis. You have an eye for detail and a great deal of common sense.

We would also like to thank involved staff at Volvo Cars. Especially Einar Gudmundsson, (Vice President at VCCS), Måns Falk (thesis supervisor at VCCS), Mathias Adolphsson (you have delivered each week as promised), all the other brilliant people that make up the VCCS lean team, and the interviewees in the QB process. It has been a truly inspiring time for us.

Last, but certainly not least, we would like to thank all the expert interviewees that let us borrow their valuable time. Your input has been invaluable given our own limited experience.

Thank you and good luck moving forward.

Göteborg, 2013-06-16

Jonny Eriksson & Johan Sandström

Abbreviations

AL	Assignment Leader
CCIM	Critical Concern Investigation Manager
CCM	Critical Concern Manager
CCMT	Critical Concern Management Team
CMQ	Current Model Quality
CONWIP	Constant Work In Progress
ETO	Engineered To Order
FIFO	First In First Out
HL/MRP	High Level / Material Resource Planning
JIT	Just In Time
KISIT-VP	Knowledge Intensive & Staff Innovative Team Visual Planning
MBO	Management by Objectives
MSS	Marketing, Sales and Customer Service
MTO	Made To Order
MTS	Made To Stock
NSC	National Sales Company
PCO	Product Compliance Office
POLCA	Paired-cell Overlapping Loops of Cards with Authorization
QAT	Quality Action Team

QB	Quality Bulletin
R&D	Research and Development
SOP	Standard Operating Procedure
TDS	Toyota Development System
TMS	Toyota Management System
TMSS	Toyota Marketing and Sales System
TPS	Toyota Production System
VCC	Volvo Car Corporation
VCCS	Volvo Cars Customer Service
WIP	Work In Progress

Table of Contents

1. Introduction

This chapter describes the background of the thesis by discussing the importance of increased white-collar worker productivity. White-collar worker productivity is then contrasted to productivity increases made in blue-collar work. The background is followed by the purpose and research questions of this thesis. Finally, the scope of the thesis is discussed.

1.1 Background

”The most important, and indeed the truly unique, contribution of management in the 20th century was the fifty-fold increase in the productivity of the manual worker in manufacturing. The most important contribution management needs to make in the 21st century is similarly to increase the productivity of knowledge work and knowledge workers.”

– Peter F. Drucker, 1999

Historically, advances in productivity were often related to capital investments, developing new tools and technologies. Increases in manual worker (from now on referred to as blue-collar worker, see chapter 3.4 for further details) productivity was rare and potential improvements were assumed only to be achieved through harder work. The first man to work as a blue-collar worker and to study blue-collar work was Fredrik Winslow Taylor. He applied structured and scientific methods to enhance blue-collar worker productivity. Traditionally, blue-collar work was perceived as being only as good as the skills that the worker possessed. Taylor meant that there are no such thing as skill. Rather, there are simple and repetitive motions put together and what makes the work more productive is knowledge about how the simple and repetitive motions are organized and performed. The philosophy of Taylor has had the most significant impact on blue-collar worker productivity and is the predecessor of many succeeding philosophies that has continued to improve productivity, such as Scientific Management and Total Quality Management (Drucker, 1999). As his methods proved successful it was the main driver for economic growth in the developed countries in the early 20th century and quickly spread around the world. Still in the 21st century Taylor’s principles are important for developing the third world countries where the labor work force is a key to internal economic growth (Drucker, 1999).

The significant impact from the increase in blue-collar worker productivity can also be explained from the fact that blue-collar workers were the biggest group of labor in the developed countries. Taking a glimpse on the present situation the single biggest group in the work force consists of knowledge workers (from now on referred to as white-collar workers, see chapter 3.4). The ratio of white-collar workers to blue-collar workers was 2:1 already in 1980, as much of the blue-collar work is being replaced by automated equipment. Thus, the future prosperity of developed countries will mainly depend on the progress and productivity of the white-collar workers (Ramírez & Nembhard, 2004; Drucker, 1999). The potential for improving white-collar worker productivity is vast. Several studies have indicated very low white-collar worker productivity as compared to that of blue-collar workers. One typical case regards white-collar worker processes in the manufacturing and service industries, where time devoted to value-adding activities is often under five percent. In other words, more than 95 percent of the time is being wasted, as nothing of value is being added to the customer. Consequently, major reductions in cost can be achieved through the elimination of waste while freeing more time for value adding activities (Blackburn, 1992). Obviously, there is a big challenge ahead for developed countries to increase their white-collar worker productivity in order to foster economic growth and to be a strong competitor. However, basically the same challenges were faced when the work with increasing blue-collar worker productivity started in the beginning of the 20th century.

Taylor's principles are the foundation of Scientific Management, which was a popular way of increasing productivity in the beginning of the 20th century. Another philosophy for the same purpose is the post-war success of Japanese manufacturing, especially Toyota's production system, referred to as the Lean philosophy. Although there has been much discussion about whether Taylor's principles were an important element in the development of the Lean philosophy or not (Spender & Kijne, 1996), it is clear that Lean has been the most widely adopted of the two in the recent decades. The objective of Lean is to reduce waste and focus on activities that adds value to the deliverable of the process, thereby enhancing productivity. The approach is now being applied to white-collar worker and service operation settings more extensively (Liker, 2004; Marsh & Conard, 2008; Conant, 1988). The Lean philosophy is flexible in the sense that it can be utilized in many environments since its cornerstones are of fundamental nature. A lot of research has been, and still continue to, focus on development in health care settings and on the efficiency of product development processes in order to emphasize value-adding activities (Liker, 2004).

An area of less focus is the utilization of pull-based control in white-collar settings. Pull is an important element of the Toyota Production System (TPS) where the subsequent process withdraws work from the preceding process. Previous research has often focused on other areas of Lean such as 5S and Kaizen (Liker, 2004). A main element of TPS is the goal of reaching a Just-In-Time (JIT) state where all processes produce the necessary goods at the necessary time in the necessary quantities to eliminate various kinds of wastes lying concealed within a company. By reaching a JIT state, work in progress (WIP) is limited. By limiting WIP in white-collar worker processes, total productivity might increase (Marsh & Conard, 2008). In order to reach JIT and being able to deliver according to its premises, pull is a prominent technique for getting there as it incorporates a replenishment system producing only what the customer took away and when the customer takes it (Liker, 2004). There are several ways of achieving a pull-based flow in manufacturing. However, there is limited work done in determining how suitable these methods are in white-collar worker environments, producing non-physical “goods”. Some of the previous literature within the area includes Conant (1988), Feather & Cross (1988) Blackburn (1992), Swank (2003), Hamm (2005), Marsh & Conard (2008), Locher (2011) and Chen & Cox (2012).

Given the huge potential in increasing white-collar worker productivity using pull-based control, it is surprising to find that the benefits and drawbacks of doing so is not better charted. The methods for achieving a low-waste value stream within blue-collar settings are well tested. Is it possible to also utilize these methods in white-collar worker environments?

1.2 Purpose

The purpose of this thesis is to describe the preconditions needed to achieve a pull-based value stream in white-collar organizations by contrasting this type of work to blue-collar settings. Different existing systems for achieving pull-based flow in blue-collar settings will be evaluated to see whether one of them is more suitable, and whether some elements of the different systems may be combined to create an appropriate basis for implementation in white-collar organizations. Also, the thesis will discuss the general suitability of utilizing pull-based flows in these settings. The aim is to create a fundamental framework that may be used as a basis for continued work with more tangible application strategies.

1.3 Research Questions

To narrow down the purpose of this thesis and to make it more researchable, four research questions were formulated to focus the research and provide a basis for an evaluation of what literature and empirical data is needed and how such empirical data is best collected.

Q1: What are the enabling preconditions for achieving a pull-based flow in blue-collar settings?

Having mapped the preconditions, the continued work is focused around how white-collar work is different from blue-collar work with regards to these preconditions.

Q2: What are the main differences between blue-collar and white-collar settings concerning these preconditions?

The last two questions focuses on how to achieve a well-functioning pull-based control system in white-collar settings.

Q3: How suitable are conventional pull-based control systems in white-collar settings?

Q4: What can be done to increase the suitability of conventional pull-based control systems in white-collar settings?

1.4 Delimitations

This thesis will deal with the preconditions of a pull-based flow on a conceptual level, that is, it will not go into details around specific white-collar worker tasks but rather focus on principle differences between white- and blue-collar work. Other important areas of the Lean philosophy and Lean manufacturing, such as continuous improvement (Kaizen) and how to develop leaders will not be the focus of this thesis. However, these areas will be presented shortly to give the reader a better understanding of the Lean concept as a whole.

2. Methodology

This chapter describes what data was required and how this data was collected and analyzed in order to answer the research questions. Also, it contains a short description of the context that was used to make the research more anchored to reality. The last part of the chapter deals with potential research quality concerns and why these concerns might be considered serious or not.

2.1 Required Data

Answering the first research question (regarding the enabling preconditions for pull-based flow in blue-collar settings) is a matter of collecting and analyzing previous literature since there is substantial previous theoretical and practical work done, including Liker (2004), Liker & Meier (2006), Monden (1998), Womack & Jones (2003), Marsh & Conard (2008), Liberopoulos & Dallery (2000), and Hopp & Spearman (2004) to mention a few. In this sense, the first research question was answered by conducting a meta-analysis of previous literature. The three following research questions required new data to be collected since these questions, especially the third and fourth, involve subjects with little or no previous work done. To answer the last three questions, and to increase the chances of the results acting as a foundation for more practical future application, a case study was conducted. Since white-collar is a broad expression, the case findings enabled tangible comparisons between conventional pull systems and the example white-collar settings. Furthermore, data from this case gave insight into the complex reality of white-collar value streams and what is needed to capitalize on the benefits of a pull-based flow. Some quantitative data regarding the specifics of the value stream input was also collected. This data was collected to gain a better understanding of the mechanics of the value stream and the context.

2.2 Research Strategy

It is fair to say that this research is mainly deductive since it is essentially about testing a hypothesis; that pull-based control is beneficial in white-collar settings. That said, research is rarely fully deductive or fully inductive, but usually contains elements of both (Bryman & Bell, 2011). Given the limited initial insight of the researchers, the hypothesis and theoretical basis of this thesis evolved as time went by. This is an example of an inductive element.

Generally, deductive research is associated with quantitative methods. However, this is mainly a distinction made for practical reasons. Qualitative research may be employed to test theories, not just generating them (Bryman & Bell, 2011). For this thesis, data was gathered mainly through semi-structured interviews, i.e. qualitative interviews. There are several reasons for this; the limited initial knowledge of the researchers makes it hard to ask the right questions from the start (instead the interviewee needs to have enough room to develop his or her reasoning), the fact that a large sample is not desirable (the sample does not need to represent a larger population) and the practical implications of finding the right people to interview. Since it was not fully clear which people that were appropriate interviewees, snowball sampling was continually applied.

2.3 Choice of Context and Sample

As mentioned in chapter 2.1, a case study was used to collect data and to anchor and exemplify the findings. It is desirable to use a context in which the sought after results may have a great impact on competitive performance. With this in mind, few industries seemed better suited than the automotive industry given how highly competitive it is. Also, many actors within the automotive industry have come far in implementing Lean production in the production of physical goods. After all, Lean production and the Toyota Production System was developed for the automotive industry (Liker, 2004).

Since this thesis is about using ideas created for manufacturing in white-collar settings it is suitable to use a context that displays both production-like and non-production-like behaviors in their non-physical value streams. One specific department that is working a lot with Lean in their white-collar organization is Volvo Cars Customer Service (VCCS) within Volvo Car Corporation (VCC). VCCS has started a Lean journey that includes standardization of tasks, Lean leadership training, visual planning, value stream mapping etc.

2.3.1 VCCS

Marketing, Sales and Customer Service (MSS) is one of the main strategic departments within VCC. MSS is responsible for activities in the market, including communication and interaction with customers, partly through national sales companies, retail partners and importers. As part of MSS, VCCS is responsible for global logistics, Customer Service Process and Technical Support, Marketing & Sales (Parts business) and Accessories and Remanufacturing business, basically all activities taking place after the car is sold. VCCS launched a Lean program in 2010. The aim of the program was to improve quality, reduce lead time and increase efficiency throughout VCCS (25% higher efficiency) and at the same time increase customer and employee satisfaction. It should be noted that program is not a rationalization effort but aims at both increased efficiency and creating new business. Resources that are freed up as efficiency increases will be used to develop new business. The main reason for starting the program was to deal with increasing competition, not just from other automotive companies, but also from private service and repair actors on the market. These service and repair actors have moved from being local, privately owned garages to being multinational chains. One major difference between these private chains and VCCS is that the private chains can reinvest their profits into their organizations while VCCS's profits are reinvested into VCC, where it is mainly used for developing new cars.

The Lean program is focused around several areas and basically covers all aspects of the work done by VCCS.

2.3.1.1 From Skill to Method

VCCS is making an effort to standardize and methodize tasks. To have standardized tasks is fundamental to know what may be regarded as waste or not. This change from skill to method may be compared to the early efforts made in manual work. When Taylor began his work with manual worker productivity he started by breaking down task into its constituent parts, thinking of the tasks as a series of movements that may be optimized. This stood in contrast to the conventional rather romantic vision of the skill of the craftsman. Having defined methods is crucial when attempting to answer the question "What is the task?". This is a question that Drucker (1999) considers to be the most important to answer when dealing with knowledge worker productivity. In creating methods, VCCS hopes to be able to identify what tasks may be considered waste or misplaced. This is considered fundamental in order to enable a steady flow of tasks, focused on value adding activities.

2.3.1.2 Leadership

VCCS is currently developing a leadership development model based on the work by Jeffrey K. Liker and his 14 principles (see Liker, 2004). The model aims at creating leaders that are skillful coaches rather than skillful politicians and promotion will be based on understanding and ability to teach others. The VCCS leadership model follows the four-stage process presented by Liker and Convis in the Toyota Way to Lean Leadership (2012); commitment to self-development, coach and develop others, support daily Kaizen, Create vision and align goals (full Hoshin Kanri). These four steps are combined with the 14 Lean principles to create a set of questions and reflections at each level. The 14 Lean principles (somewhat modified) forms the Y-axis and the four steps presented by Liker forms the X-axis of a matrix (see figure 1). VCCS calls this model "9:9 Aspired Leadership" with 9:9 referring to the top right corner of the development matrix. In order to coach other, leaders need to at least reach level 5:5.

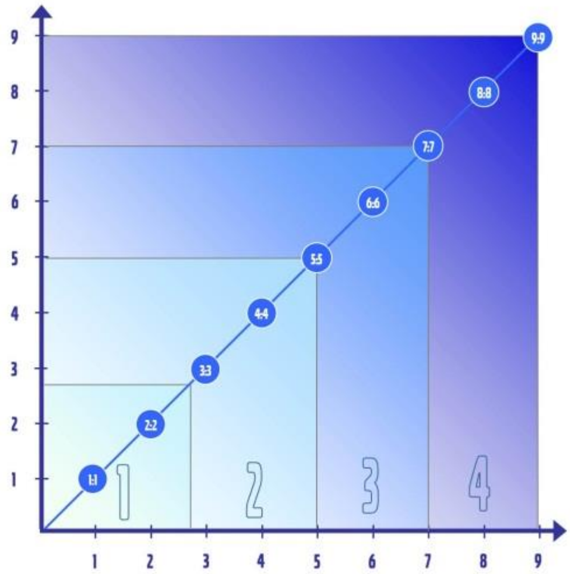

Figure 1. Arrangement of the VCCS 9:9 Aspired Leadership Model

2.3.1.3 Understanding the Value Stream

One of the key areas for VCCS is to gain better understanding of how value is created in the different value streams. In order to do this the current state needs to be mapped and substantial work is done with Value Stream Mapping (VSM) in different areas of the organization. VCCS have started by creating a map of what is considered to be the flow on a Meta level. Different areas are then mapped in greater detail and compared to the desired flow. The aim is to reduce the separation of roles and design value streams around the workers, creating teams that have enough internal competence to follow the value stream for as long as possible. This will give each individual a better understanding of how value actually is created. Less separation of roles will require a culture of coaching and continual learning since it will result in individuals learning new things and taking on more responsibility. A redesign of the packaging process within VCCS (putting spare parts into blue Volvo-branded boxes) is one example of this strategy in action. The packaging process originally involved 21 different roles and the front and end parts of the process were unconnected. The redesigned target condition reduces the number of involved roles by 60% and 80% of the steps are assigned to one person. This will create responsibility and accountability throughout the flow that, in turn, most probably will lead to increased quality and stability.

2.3.1.4 KISIT-VP

KISIT-VP is short for "Knowledge Intensive & Staff Innovative Team Visual Planning". VCCS believes in the importance of visibility and uses visual planning to a very large extent. Visual boards coordinating the daily work are used at various levels, from teams of workers to top management.

2.3.1.5 Hoshin Kanri vs. Management by Objective (MBO)

VCCS is describing the difference between MBO and Hoshin Kanri as the difference between setting targets at the top of an iceberg and setting targets on the top of a mountain. Instead of pushing down objectives set at the top VCCS wants to reach a state where objectives are set and decisions are made using both vertical alignment and horizontal coordination.

2.3.2 Sample: The Quality Bulletin (QB) process

This thesis is focused on one part of the context: the QB process. The purpose of this process is to solve quality issues from the market or internally. This basically involves filtering input, finding the root-cause of the issue, designing the necessary changes and preparing and launching these changes to the market. The process was deemed suitable since it is well defined and delimited and since it demonstrates various degrees of complexity.

2.4 Data collection methods

As mentioned above, the researchers had limited initial insight into the context. This suggests that it was favorable to use an abductive approach, i.e. an iterative way of finding new dimensions to the research and to continuously update data collection methods and analysis accordingly (Bryman & Bell, 2011). This holds true for basically all preferred methods of data collection; literature review, semi-structured internal interviews and unstructured external interviews.

2.4.1 Literature Review

The initial step of the research process was a more extensive literature study to broaden the theoretical basis of the thesis. Interesting fields of literature include classical Lean production, Lean office applications, fundamentals of pull systems, differences between white- and blue-collar value streams and some basic complexity theory. As mentioned above, the literature review was of abductive nature. The literature collection was focused around the digital and non-digital Chalmers library. To make sure that the literature was up-to-date and regarded as significant, references from fundamental works were used. Chalmers staff within appropriate areas was used to find these fundamental works and to give initial guidance.

2.4.2 Interviews

A great deal of knowledge and information was gathered using interviews. Interviewing is a suitable method for data collection since plenty of the know-how and expertise are inside the minds of people and not documented. Interviews were conducted both internally at VCCS and with external parties to get a broader understanding of the problem. In both cases, a qualitative approach was used for the sessions as the phenomenon studied is complex and partially unexplored. This implies that expert opinions may come in different forms why flexibility in the interviews is necessary. One method within qualitative interviewing is the semi-structured interview, which allows the interviewee to explore thoughts in different directions and thus providing insightful information. However, the interview should follow an interview guide with some specific topics to be covered. Another method is the unstructured interview, which allow for even more freedom for the interviewee to discuss a topic based on own experiences. In the unstructured interview there might only be one single question that should be considered initially, giving the interviewer a chance to respond to points that seems worthy of following up and lead discussions toward these topics (Bryman & Bell, 2011).

2.4.2.1 Internally at VCCS

The interviews at VCCS were of the semi-structured kind since there were a number of specific issues being discussed with interviewees; however, they were allowed to explore different directions within the topics. To give the interviews some structure, an interview guide was created. The guide consisted of a series of question mainly focused on how the value stream behaves between operations, i.e. the handovers. The questions were created using a generic process step visualization, with input, inbound inventory, main process, outbound inventory, output and supportive activities (see figure 2). The interview guide assisted in focusing on the critical factors that needed to be assessed with regards to the characteristics of the flow, such as the batch sizes of the tasks and frequency of the deliveries. Thus, there was an objective to, when possible, obtain quantitative statistics for some parameters to increase the understanding of the dynamics of the white-collar value stream. However, in many cases, the interview guide could not be followed to the letter since the desired information simply was not available. In those cases, the interviews were more unstructured. The interviews were primarily conducted with the leaders for the different departmental functions whom are part of the QB process since they normally have the most holistic understanding of the process and its interrelations. Building on the knowledge from the semi-structured interviews, a better understanding of the value stream was acquired.

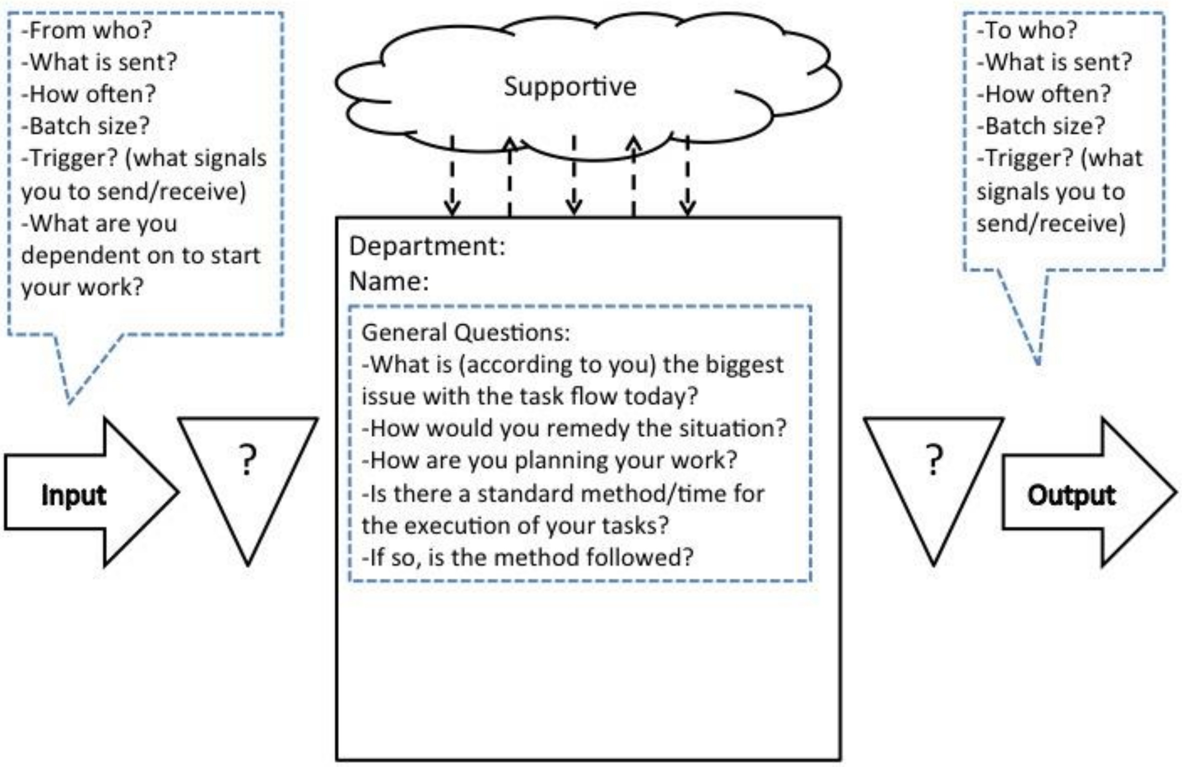

Figure 2. Template used when preparing the internal interviews

2.4.2.2 External Parties

The external interviews played an important part in increasing the knowledge of the topic and to get an understanding of what has been done earlier within the area. Since the external parties are of different background and can provide varying contributions to the research these interviews were unstructured. Given the varying contributions, unstructured interviews are appropriate since they provide the highest flexibility (Bryman & Bell, 2011). Thus, these interviews were more of a discussion and the response is seen as opinions and is evaluated thereafter. The input was used to point out critical areas to focus on in the empirical studies and also assisted with ideas towards the end, when synthesizing the data and making recommendations. The external parties consisted of consultants, coaches, professors and doctoral students with proven experience from working within the Lean area. A personal network of connections was utilized in order to identify key people from the four categories mentioned.

The external interviewees:

- **Dag Lotsander:** Independent Lean consultant since 2008. Before that General Manager within sales and aftermarket at Toyota Sweden where he was responsible for introducing Lean to the entire Swedish organization (previously only the production facility was utilizing Lean). Dag has held approximately 450 lectures and have had around 32 000 participants.

- **Lars Medbo:** Associate professor in Logistics and Transportation at Chalmers University of Technology. He is researcher, teacher and head of the Division of Logistics and Transportation. His research focuses on materials supply systems, materials handling and production flows in production system design. The work has resulted in development of new materials feeding techniques, layouts, materials flow structures and work structuring principles implemented in the Swedish automotive industry.

- **Ludvig Lindlöf (Thesis supervisor):** PhD student at the department of Operations Management at Chalmers University of Technology. Research focused on making product development processes and organization more effective, more specifically how the principles of Lean product development can be used in that context.

- **Sara Algestam**: Employed doctoral student at the Division of Logistics and Transportation at Chalmers University of Technology since 2010. Involved in the project Strategic Vehicle Research and Innovation, and her research focuses on pull-oriented material planning systems.

- **Stefan Bükk**: Lean Development coach with extensive experience. Focused on Lean Development but have substantial knowledge of most areas within Lean.

- **Tomas Helling:** Currently Lean consultant at Lean Communications in Norway. Used to be responsible for a process similar to the QB process at SAAB Automobile. Has extensive experience from implementing Lean solutions.

2.5 Data analysis

The data collected from literature was summarized to form a framework containing the main preconditions of pull as seen by the researchers. To perform this summarization the data was coded, i.e. the abstraction level was raised to be able to collect data into suitable categories. Examples of categories are standardization, visual control and leveling.

Data collected through external interviews were summarized (see chapter 4.2) and used to form a model for analysis and to draw conclusions. All involved external experts received the summarized data in order to validate what themselves and others had said.

The data collected by internal interviews at different stages of the QB process were analyzed for mismatches and contradictions. For example, what an upstream activity called their output was compared to what next downstream activity called their input. In this manner, the data could be validated throughout and existing question marks could be straightened out.

When all theoretical and empirical data had been collected and summarized, a model was created to evaluate different pull systems and how the QB process is positioned with regards to the evaluation criteria (see chapter 5.4). The model consists of the criteria deemed most important for evaluation and is of qualitative nature.

2.6 Research Quality Concerns

A result of the chosen qualitative method and non-probability sampling is that the transferability of results comes in question, that is, the extent to which generalization of results can be made. This is a common issue in qualitative research and there is no easy way to circumvent it (Bryman & Bell, 2011). Worth noting is that, according to Lincoln and Guba (1985), transferability is a matter of concern for the consumer of reports, i.e. it is up to the receiver of the report to determine if the results are transferable or not. This, however, requires a thick description, i.e. the description should explain the context in such detail that findings become meaningful to an outsider. The “thickness” of the description will probably be the limiting factor of the transferability of the results of this thesis. Since the data collected is not used to make predictions about the opinions or preferences of an entire population, but rather to look for certain properties of the value stream at a certain location, convenience sampling causes limited or no damage to the credibility of the results.

To make the results as trustworthy and credible as possible it is important to let both internal and external respondents verify the collected data before making final conclusions and to look for consistency and reoccurring statements from several respondents. One way of doing this is to let respondents from different steps in the value stream answer questions about the same activity, for example to ask the representative of a upstream activity to clarify certain aspects of their output and then ask the representative of the downstream activity to clarify the same aspects about their input and then look for mismatches. If mismatches are found, these need to be explored further by both involved parties, preferably together.

The thesis is focused on pull on a conceptual level and most interviewees have managerial positions since these positions usually provide a better process overview. However, not interviewing the people closest to the actual tasks of the flow may result in certain loss of detail.

3. Theoretical Framework

In this chapter, the relevant theory needed for understanding the problem area is presented. First, a general description of the Lean philosophy is described, leading to a more extensive presentation of Lean production and pull systems in order to understand the preconditions. Subsequently, differences between white- and blue-collar work is better understood. Related to these differences, task complexity and the assumption of order is shortly introduced.

3.1 The Toyota Production System and Lean Production

TPS is a proud legacy of the automotive manufacturer Toyota and is aimed at optimizing manufacturing processes. Their way of working has also been known as Lean production because of the Lean approach that Toyota has towards manufacturing. Toyotas production system was shaped in post-war Japan which was a relatively small and diverse market in terms of variety of vehicles. Because of this, the key to success was flexible operations. The tough market conditions made Toyota realize that a key to be flexible is to eliminate the waste that occur in the processes which help in reducing lead times, which is a key concept of TPS. Toyota found that short lead times and flexible operations provide higher quality, better customer responsiveness, better productivity and better utilization of equipment and space. A reason for the wide spread use of TPS amongst companies today is that they are now facing the similar market conditions as Toyota did in post-war Japan: the need for fast, flexible processes that give customers what they want, when they want it, at the highest quality and affordable cost (Liker, 2004). Becoming a Lean production manufacturer can be defined through a five-step process including; defining customer value, defining the value stream, making it “flow”, “pulling” from the customer and striving for excellence. Working in the Lean way obliges one to focus on getting the product to flow through the operations, ideally only in value-adding activities, by means of a pull system that throughout the operation replenishes only the customer demand and a culture of striving to improve continuously (Womack & Jones, 2003).

As mentioned, though it cannot be emphasized enough, the Lean philosophy is in a sense about eliminating the wastes in the operations that does not add value to the customer. In TPS, the very first question is to understand what the customer want from the process in order to really understand what is value and what is not. The customer could be both the end user, but just as much the internal customer being the next step of the process. The concept is not only applicable to manufacturing, but to non-physical processes and services as well. Though not all non-value-adding activities can be eliminated, the goal is to minimize the time spent on these activities. The Lean improvement is quite different to the traditional process improvements where companies often focus on a critical machine and try to increase uptime or decrease the cycle time. The traditional method could definitely improve that individual machine considerably, though still having a small impact on the overall value stream. A reason for this is that the particular machine is just one step of the overall process, and it will give a small overall contribution. The Lean method is rather to phase out the non-value adding activities, while the value-adding time is also reduced. This kind of improvement is often achieved through trying to create a one-piece flow, e.g. using a cell. The cell is a close arrangement of resources (people, tools or desks etc.) needed to execute the operation, and the item being processed is passed from one activity to the other in a one-piece flow, with minimum in-between inventory, space and lead time. The one-piece flow means making one unit at a time which corresponds to the customer's demand rate, called the takt time. As inventory is minimized there is no space for downtime as in mass production, meaning that quality issues become visible instantly (Liker, 2004).

The TPS is visually presented as a house (figure 3). The house is a good representation as its strength is dependent on all links that builds the structural system. If one link is weak, the house will not be stable enough to provide all the desired benefits. The goals of TPS is found in the roof and is the state of performance one should aim for. The roof is standing on two pillars, JIT and Jidoka. JIT is about minimizing inventory, and this is where the pull-based control system is introduced in the TPS. Jidoka is in a way supported by JIT as the intention is to resolve problems immediately as they arise in order to continue producing, thereby never letting a defective part pass through the system. In the center of the house there are people and a focus on their mindset. Only through the people's willingness to adapt to the new culture, seeing wastes and continuously improve can the TPS philosophy be realized. The house is standing on the firm foundation of TPS which is primarily made up of four elements. First, the production must be leveled, Heijunka, in order to ever reach a continuous flow through the operations. Second, there is a need for stable and standardized processes so that improvements efforts will be incorporated successfully. Third, for TPS to be successful it must have a visible management system which allows for all workers to see the process and to see where and when problems occur. Last, but most fundamentally, there must be a sense of the Lean philosophy among all employees. The full effects of TPS start with the people being ready to work in a new way, a new culture, and to contribute in their best possible way (Liker, 2004).

Figure 3. Toyota Production System visualized as house with its constituents (Liker, 2004)

3.1.1 TPS within Toyota

It is important to realize that Toyota is not just about TPS. As the name suggests, TPS is a production system and is not primarily designed for development work and other activities with high complexity. James Womack, the man who coined the expression “Lean production”, actually warned about taking Lean production upstream (Holmdahl, 2010). Toyota is made up of four interconnected systems; Toyota Development System (TDS), TPS, Toyota Marketing and Sales System (TMSS) and Toyota Management System (TMS) (see figure 4). Although the TPS is primarily designed for production, some authors advocate its use in less certain environments. Liker (2004), for instance, thinks that it is possible to use the principles of the TPS outside the factory floor since every process is repeatable at some level. It does, however, require some imagination and hard work. Value-streams are usually harder to map when dealing with non-physical products, but, using the customer as starting point, it should be possible. It is important to be able to separate repetitive and unique processes, and apply TPS-principles on the repetitive ones (Liker, 2004).

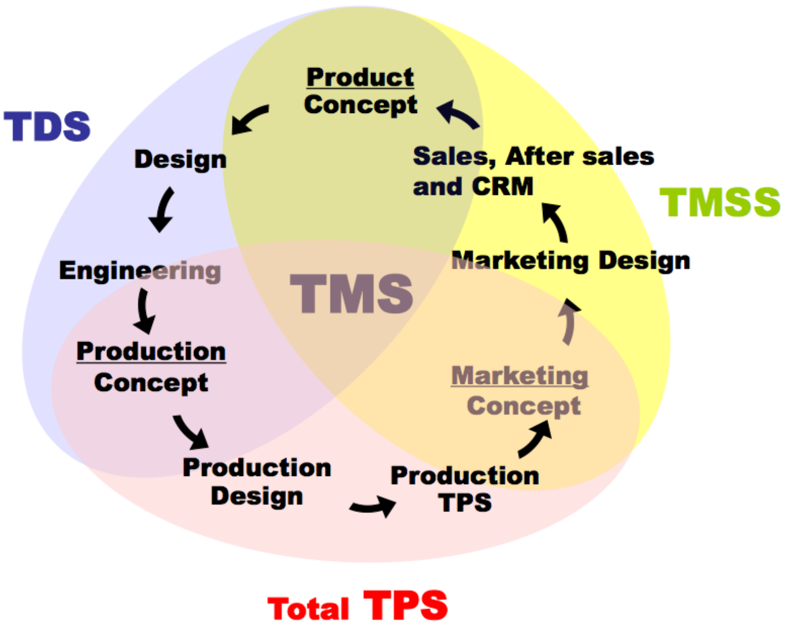

Figure 4. Toyota seen as interconnected systems, from Tanaka (2011)

3.1.2 The Toyota Way Principles

Having built up the basic knowledge of TPS and its background, it is appropriate to go a bit deeper into the enabling elements of TPS. The TPS can be explained through the principles necessary to reach the state of continuous improvement and excellence. The Toyota Way, which incorporates TPS, is not a toolbox that guarantees success, rather it provides means for continuously improving everything that is being done. The success depends solely on the workers and their attitudes towards the Lean culture. Thus, there are 14 principles that constitute the Toyota Way. Liker (2004) organized these principles into four categories, called the four P’s of the Toyota Way, see figure 5. The categories are Philosophy, Processes, People and Partners, and Problem Solving. A strong philosophy helps in focusing in long-term achievements, the right processes will produce the right results, value is added to the organization by developing people and partners, and continuously solving root problems will drive organizational learning (Liker, 2004).

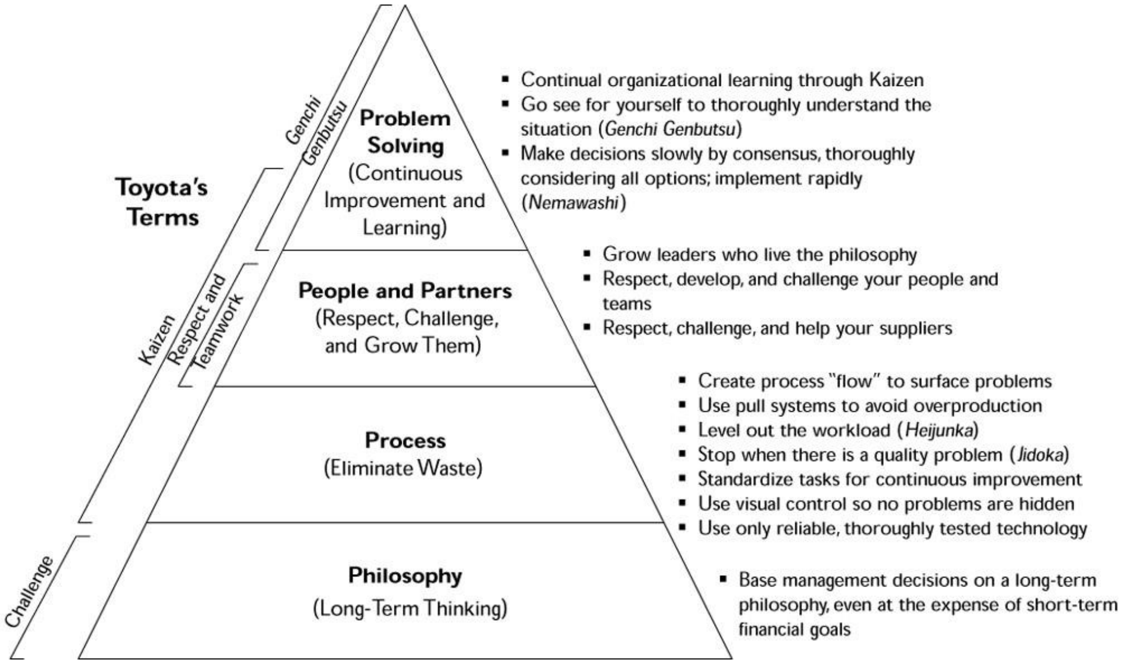

Figure 5. The 14 principles and how they relate to the four P's of Toyota according to Liker (2004)

Even though not all of the 14 principles are relevant for this thesis, they will be briefly summarized to provide a complete understanding of their importance and relations between one another. The principles described below are summarized from Liker (2004). VCCS have developed their own visualizations of the 14 principles (see figure 6).

First P: Philosophy

Principle 1: *Base decisions on a long-term philosophy, even at the expense of short-term financial goals*

Emphasize a philosophical purpose rather than short-term profit, which allow the organization to work and grow toward something else than just profit. The starting point should be to generate value for the customer, society and economy. Maintain and improve the skills that enable you to produce added value.

Second P: Processes

Principle 2: *Create continuous process flow to bring problems to the surface*

Redesign work processes to achieve high value-added. Make flow evident in the organizational culture as it is the key to continuous improvement and to developing people by letting problems surface right away.

Principle 3: *Use "pull" systems to avoid overproduction*

The downstream operation replenishment is triggered by the downstream customer who withdraws materials upstream when they want it and at what amount they want it. Minimize the work in process (WIP) by stocking small amounts of components and frequently refill what customers took away. This is the basic principle of JIT.

Principle 4: *Level out the workload (heijunka)*

Remove the overload on people and machines and eliminate unevenness. Leveling out the workload of all manufacturing and service processes rather than to stop and start work as batches are delivered.

Principle 5: *Build a culture of stopping to fix problems, to get quality right the first time*

Quality for the customer is the value proposition and the equipment should have the capability of detecting problems and stopping itself. Develop a visual system to alert the team that a machine needs assistance (*Jidoka* is the foundation for building in quality). Create an organization that solve problems and put countermeasures in place quickly.

Principle 6: *Standardized tasks are the foundation for continuous improvement and employee empowerment*

Use stable, repeatable methods everywhere to maintain the predictability, regular timing, and regular output of your processes. It is the foundation for flow and pull. Standardize best practices, but allow for creativity and individual expression to challenge the standard and improve upon it. Then implement the improvement into the new standard to ease learning for new people.

Principle 7: *Use visual control so no problems are hidden*

Develop simple visual indicators and systems to help people determine instantly whether they are meeting the standards or not and to support flow and pull functionalities.

Principle 8: *Use only reliable, thoroughly tested technology that serves your people and processes*

Use technology to support people, not to replace them. Often, the right technology that truly support the process is done manually at first, and then the technologies are added to aid. Reject or modify technologies that conflict the culture and that might disrupt stability, reliability and predictability. Quickly implement a thoroughly considered technology if it has been proven effective in trials.

Third P: People and Partners

Principle 9: *Grow leaders who thoroughly understand the work, live the philosophy, and teach it to others*

Promote leaders within the organization to maintain a sense of belonging. Promote leaders that are good role models for the company and who "lives" the company's philosophy and way of doing business. Strive to develop leaders that understand the details of the work so that he or she can be the best teacher of the company's daily operations and code of conduct.

Principle 10: *Develop exceptional people and teams who follow your company's philosophy*

It is important to build a strong and stable culture in the company where values are being shared and lived. Utilize cross-functional teams to improve quality and the flow through problem solving, empower people who use the company's tools. Educate individuals to work within the corporate philosophy continuously and to be better team players.

Principle 11: *Respect your extended network of partners and suppliers by challenging them and helping them improve*

Show respect for your partners and suppliers and make them know they are regarded as an extension of your own business. Show that you value them by setting challenging objectives and helping them to achieve these objectives in order for them to continually develop.

Fourth P: Problem Solving

Principle 12: *Go and see for yourself to thoroughly understand the situation (genchi genbutsu)*

Actions to solve problems and to improve operations should be based on observed facts from the source, rather than creating theories based on other people judgments or computer calculations. Refer to own personally verified data when discussing different issues. Executives should also go and see for themselves to deepen their understanding of the situation.

Principle 13: *Make decision slowly by consensus, thoroughly considering all options; implement decisions rapidly (nemawashi)*

Allow time for thoroughly considering all alternatives to solve problems. Once the path has been decided, move quickly down that path. *Nemawashi* refers to the approach of discussing ideas and thoughts with all people affected from a change or problem, and to get their ideas and to mutually agree on the path forward. *Nemawashi* is obviously time-consuming, however, as it helps broaden problem solving and getting a mutual acceptance of the decision, implementation of the solution can be rapidly executed.

Principle 14: *Become a learning organization through relentless reflection (hansei) and continuous improvement (kaizen)*

When a stable process is in place, continuously improve the process by identifying the root causes and implement effective countermeasures. Aim at achieving a process with minimum inventory as it makes waste visible for everybody. As waste is exposed, use the company's continuous improvement efforts (*kaizen*) to eliminate it. Allow time for reflection (*hansei*) at each milestone to openly identify shortcomings of the project, develop countermeasures to avoid the same mistakes again. Incorporate learning by standardizing best practices, to avoid time-consuming inventions for new projects.

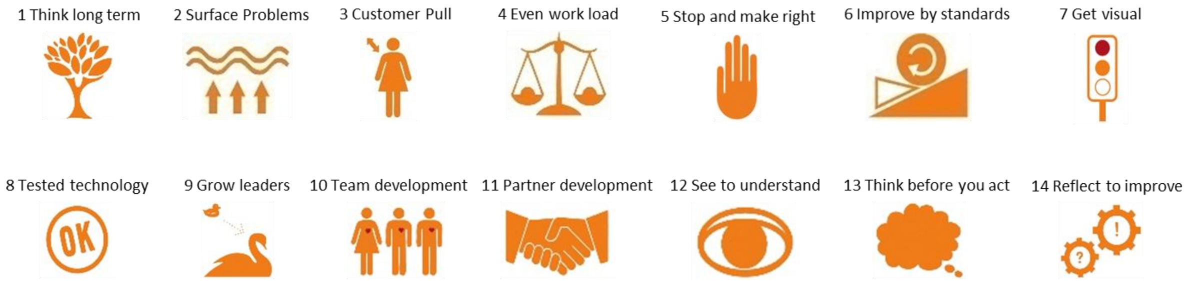

Figure 6. VCCS's visualization of the 14 Lean principles

All principles are in a sense related to one another and are necessary in order to achieve complete success for the Lean enterprise. However, there are some principles that are more interesting than others when investigating the enabling of a pull-based value stream. These particular principles are more or less related to JIT, which also is one of the pillars in the TPS house. The principles of more interest are principle 3, 4, 6 and 7, which not surprisingly are related the second P: Processes. JIT is the ultimate state for processes that aims at reducing inventory, achieving a controllable process and allowing for high quality. Additionally, the ideal state of the JIT production is the pull-based value stream where the customer gets what he or she wants, when and in the amount he or she wants it (Liker, 2004). To facilitate the pull system (*principle 3*), there are some prerequisites that must be put in place. First, standardization of processes (*principle 6*) is critical for it to be predictable enough, to know how long time different cycles take. Since a process is following predetermined sequences, the takt time helps in determining a standard quantity of WIP to allow the different process steps to operate simultaneously (Monden, 1998; Liker & Meier, 2006). Probably the most important condition for implementing a pull system is the smoothing of production, or leveling out the workload (*principle 4*), which refers to the process being able to meet the peak in the variance of demand. As the subsequent process withdraws items from the preceding process, this one must also be able to meet the variance in demand through preparation of inventory, equipment and manpower. Smoothing thereby enables the pull system to flow adequately (Monden, 1998). Another important element for pull is to have visual control (*principle 7*). The visual control system helps in visualizing the process and thereby highlighting problems and defects immediately (Monden, 1998; Liker & Meier, 2006). This allows for fast problem solving, which is in the interest of everyone and it might also increase the interest of employees to take part of problem solving. The extended interest enables components to flow better in the process. In the pull-based flow, it is critical that only good components are flowing. Letting defect products flow has a greater negative impact on a pull flow compared to a mass production where defect components can be put aside and work can continue on the good enough components (Monden, 1998).

3.1.3 JIT Production and Related Principles

As mentioned above, the ultimate state for a Lean process has the characteristics of the JIT system, both in manufacturing and in white-collar processes. The most important principles for achieving pull have been highlighted in previous chapter but each principle will be thoroughly explained below to understand what the prerequisites are for reaching JIT, or a pull flow.

3.1.3.1 The Pull System

Many often equate the word “pull” with “flow”. The similarity is that they are both concepts and they are also linked in the pull system. Flow denotes how material moves from one process to the other, while pull rather prescribe when material is to be moved and who determines whether it will be moved, namely the customer. A pull system could be explained as a replenishment system, which means that replenishments takes place when the inventory is getting too low (Liker & Meier, 2006). Think of the replenishment of milk; you would not want to buy a big batch of milk since you are not certain how much you will consume within the upcoming week. Rather, you go and buy milk when you realize that the milk is running out. Thus, the trigger in the pull system is that the inventory is low on a particular item and the inventory needs to be replenished. It is however possible to achieve flow using a push system, Liker & Meier (2006) identify three primary elements that distinguish pull from push:

1. **Defined.** A defined agreement with specified limits pertaining to volume of product, model mix, and the sequence of model mix between the two parties (supplier and customer).

2. **Dedicated.** Items that are shared between the two parties must be dedicated to them. This includes resources, locations, storage, containers, and so forth, and a common reference time (takt time).

3. **Controlled.** Simple control methods, which are visually apparent and physically constraining, maintain the defined agreement.

The push system normally has no defined agreement between the supplier and customer on the volume to be delivered, or when it should be delivered. Instead, the supplier works in his own speed and make sure to be aligned with his own predefined schedule. When the supplier is done, the material is sent (pushed) to the customer, whether it was requested or not, and the material is placed where there is space, so there are often not a dedicated location. As there is no definition or dedication between customer and supplier, there is no easy way to understand what to control or how to do it (Liker & Meier, 2006).

Deviation is an issue that may interrupt the flow. There are some primary causes of deviation when implied by the operator:

1. Imbalanced work cycle times that may be due to normal variation in work content, operator skill, or machine cycle times. Typically, the person with extra time will deviate.

2. Intermittent work stoppages due to lack of parts (information) or operators leaving the work area to perform additional tasks – such as retrieving parts or performing quality checks – machine failures, or correction of defects.

3. Intermittent work delays due to struggles with machines or fixtures, or overly difficult or complex tasks.

4. Miscellaneous issues such as “building ahead” to “buy time” for change-over, an operator leaving the line for some reason, or to stagger break or lunchtimes, or such.

A one-piece flow requires perfect operation time balance, which is not feasible in most circumstances. Even small natural variation in the work cycles would disrupt the one-piece flow. For work being done by humans this is especially hard as the tasks cannot be completed with exact precision each time. Since wait is waste (Liker, 2004), naturally the operator will add buffer to stay busy as variation occurs between process steps. However, it is important that the added buffer to compensate for minor variation is defined as the standard WIP between the different operations. Unless it is being defined as standard there will also be variation in the buffers, thereby implying more waste (Liker & Meier, 2006).

To be correctly interpreted, pull is the second best alternative when a true one-piece flow is not feasible. Toyota’s solution is the Kanban system (for a detailed description of the Kanban system refer to chapter 3.2.1), which is one way to achieve pull in the processes. An example from Toyota is in their assembly plant where stamped steel panels are welded together into a body. The stamping is a fast process; it takes about one second, while the welding operation takes 60 seconds. So a one-piece flow is not practical. Instead, a Kanban system is in place which withdraws panels from a supermarket (storage shelf) between the operations, and when a certain number of steel panels have been used a Kanban card goes back to the stamping process to order more panels. The ordering takes place when a certain trigger point has been reached, which could be defined as a minimum in-between inventory (Liker, 2004).

However, Toyota is not obsessed with implementing Kanban systems everywhere, the push system also has its place. Often, Toyota use push system, scheduling of orders from suppliers for example, when there are short lead times such as ordering each day rather than once a month. In many cases where a pure scheduling system is present, which is beneficial thanks to more advanced computer equipment, the process is still visually controlled through cards similar to the Kanban system (Liker, 2004).

3.1.3.2 Level Out the Workload

In general, when you try to apply the TPS, the first thing you have to do is to even out or level the production. And that is the responsibility primarily of production control or production management people. Leveling the production schedule may require some front-loading of shipments or postponing of shipments and you may have to ask some customers to wait for a short period of time. Once the production level is more or less the same or constant for a month, you will be able to apply pull systems and balance the assembly line. But if production levels—the output—varies from day to day, there is no sense in trying to apply those other systems, because you simply cannot establish standardized work under such circumstances.

- Fujio Cho, President, Toyota Motor Corporation

Leveling the workload, or *heijunka*, is an important element for making it possible to use a pull system. To level out the workload, there are three primary elements that should be minimized; these are illustrated in figure 7. *Muda* which represents waste in the process is often the primary focus of firms trying to work Lean as it is easier to work with, however, *Mura* (the unevenness) is of higher order in terms of cause and effect on smoothing. Minimizing the unevenness is fundamental in eliminating *Muda,* and also, the overburden on people and equipment (*Muri*) that could cause breakdown and defects, further destroying the possibilities to achieve an effective pull flow (Liker, 2004).

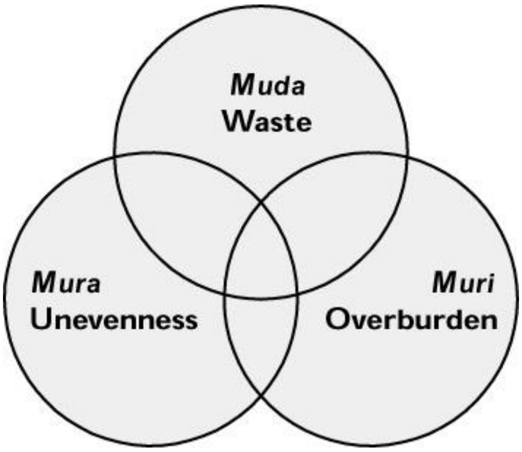

Figure 7. The three primary elements of waste according to Liker (2004)

The meaning of *heijunka* is really to level the product mix and product volume over a specific time period as it will make the process produce the biggest mix of parts every day, which is contradicting to the traditional batch deliveries (Liker, 2004). The idea is based on the fact that customers seldom order products in specific batch sizes, the consumption and ordering of the customers is more random than that. Smoothening a process also incorporates the highest possible flexibility and responsiveness in the process as it is being built to tackle changes in demand. Obviously, there will be some negative effect initially; paradoxes are a recurring phenomenon in TPS. Since there will be more changeover due to a greater mix of products, production time will be lost. Working leveled will eliminate the possibility to reduce this negative effect through batch production, instead, the only option is to improve the changeover process and make it standardized (Monden, 1998; Liker & Meier, 2006).

As you want to level out the workload and be as smooth as possible, there could still be a need for inventory to reach a smooth and even workload and not just any inventory, but the most expensive kind, the finished goods inventory. This may seem contradicting to Lean thinking with the aim of the smallest possible inventory, but in order to reach *heijunka*, there must be a safety level which allows the production to run smoothly even at high peaks when customer demand is shifting. The inventory should obviously be the smallest possible, but by allowing the small waste in finished inventory it is possible to eliminate a lot of other waste in the processes while enabling pull. Furthermore, leveling can be done in service operations as well as in manufacturing though it is more cumbersome due to the typical lower-volume environment. The main way to level the workload is in principal analogous; first the customer demand should be fitted into a leveled schedule. Thus, cases and tasks should be planned given a pre-defined schedule. Second, standard times must be established for performing the different kinds of work, as it will enable the possibility to level out the workload through more accurate predictions. To conclude, in order to have a smoothed production and leveled workload, it is important to foster it by having a pull approach and visual management so that problems are clearly, and to use standardization of tasks to control lead times (Liker, 2004).

In leveling the production it is also critical to decide on the takt time which will set the standard for the process performance and pace. The three main aspects that should be leveled includes (Liker & Meier, 2006):

1. Product *volume,* which is simply the quantity of a given product that must be produced in a specified period of time (the pitch).

2. Product *mix,* which is the proportion of the various models that are produced during the pitch increment, the quantity of A's, B's, C's and so forth.

3. Produced *sequence*, which is the order that the product volume and mix are produced. It may be model by model, such as A, A, A, B, B, B, C, C, C, or part by part, such as A, C, A, B, A, C.

The aspects are listed in order of difficulty in achieving leveling. As it might be hard to decide on the correct volume to smoothen down to, Toyota generally selects approximately 80 percent of peak demand. The gap between 80 and 100 percent can be covered by overtime, also the occasionally free time allows for better flexibility in the process (Liker & Meier, 2006). In situations where you have workers handling more than one machine which are normally 90 % loaded, the increases in demand can be dealt with by hiring temporary workers so that each worker is handling fewer machines, thereby reaching 100 % utilization of machine capacity. For this option it will be necessary to have machines that even new and unskilled labor can manage effectively within a few days. Another issue is when demand decreases, obviously temporary workers will be dismissed, but there are still excessive employed workers. Toyota's way of dealing with excess manpower is to let workers take a rest from the normal day-to-day activities rather than to produce unnecessary stock. Some examples of actions during downturns are listed below (Monden, 1998):

- Transfer workers to other lines for which demand increased

- Decrease overtime

- Use a paid holiday

- Conduct quality control circle meetings

- Practice set-up actions

- Conduct maintenance and repair of machines
- Manufacture improved tools and instruments
- Conduct plant maintenance and upkeep
- Manufacture parts previously purchased from suppliers

3.1.3.3 Standardized Tasks

Standard work sheets and the information contained in them are important elements of the Toyota Production System. For a production person to be able to write a standard work sheet that other workers can understand, he or she must be convinced of its importance.... High production efficiency has been maintained by preventing the recurrence of defective products, operational mistakes, and accidents, and by incorporating workers' ideas. All of this is possible because of the inconspicuous standard work sheet.

- Taiichi Ohno, the father of TPS

Standards play a key role in the TPS, without it the system would never work. Standardization is the very foundation for continuous improvement as the standard of a process makes it predictable and stable, thereby enabling correct analysis of the process and find improvements. If the process is not standardized and is varying uncontrollable, any improvement effort might just be aimed at another variation of the process, which will continue to vary. Furthermore, standardization must always take place to assure consistent quality and to ensure that improvements are being incorporated and execute by everyone. Standardization is in itself a continuous process, consisting of temporarily freezing processes but changing them as improvements arise (Liker & Meier, 2006; Liker, 2004). In TPS there are three primary elements of standardization, which provide a broad frame of standardization and guides the implementation of standards and what should be included (Liker, 2004):

- Takt time – time required to complete one job at the pace of customer demand
- Sequence of doing things
- How much inventory the individual worker need on hand in order to accomplish the standardized work

Monden (1998) presents the same main elements in a slightly different way, and connect them to something he calls standard operations. The first element is the cycle time, which is the concept needed to fulfill the goal of line balancing among all processes. This is an important part for effectively implement a pull system as well. The second element is the standard operations routine which specifies the standardized order the various operations should be performed by the worker. The routine will help in fulfilling the goal of working efficiently without wasteful motions, and is also important in implementing a pull flow. The third element is the standard quantity of WIP which specifies the minimum number of units necessary for the standard operation to be performed by workers, thereby fulfilling the goal of having a minimum WIP to eliminate excessive inventories and burden on workers. The standard quantity of WIP is also key in realizing a pull-based flow as it must be specified what amount of inventory that will be accepted throughout the value stream.

The cycle time, as defined by Monden (1998), specifies the time span available for producing one product. The cycle time is determined by the required daily quantity of output and the effective daily operating time. The effective daily operating time should not be reduced due to anticipated machine breakdowns, idle time when waiting for material, rework of products or for fatigue and rest time. Thus, it should be all available time to push for improvements (Monden, 1998).

Another measure to standardize in connection with cycle time is the completion time per unit, which specifies the time needed to refine a unit at each process and for each part. As the cycle time and completion time per unit are determined, the number of different operations each worker should be assigned to and in what order to perform these can be decided (standard operations routine). The next step is to determine the standard quantity of WIP. If the operations routine is in accordance with the order of process flow, only work attached to each machine is necessary in the ideal world (without concern to variation in demand, time to completion, etc.). On the other hand, if the operations routine is in an opposite direction to the order of processing, it must be necessary to hold at least one piece of work between machines (Monden, 1998). Another important aspect when introducing standardization is visual control, which will be discussed below.

3.1.3.4 Visual Control

Just having and knowing the standards is not enough; they must also be visual to everyone (Liker & Meier, 2006). Visual control is much a communication tool with the objective of providing immediate feedback to employees for them to do a better job. It is an important part of JIT production as the visual control offers just-in-time information to ensure fast and accurate execution of operations and procedures. The visual control may show where items belong, how many items belong there, what standard procedure to follow, status of WIP, and other types of information critical to the value stream activities. Possibly even more important, the visual control is a great tool for visualizing and identifying any deviations from standards and thereby is facilitating the flow in the value stream (Liker, 2004).

A visual control innovation from Toyota is the *obeya*, which means big room. *Obeya* is referred to as a “war room” where most of the visual management tools for a project are displayed and maintained by the responsible representatives from the different departments and functions taking part in a project. The idea is for the project to be easily monitored and provide fast response to possible difficulties arising during the project. The visual tools include status of each area and key supplier compared to schedule, design graphics, quality performance, manpower charts, financial status and other key performance indicators. The visual control tools in the room should be continuously reviewed by all team members in the project (Liker, 2004).

Most of the trusted visual control systems use lights or other type of signals to indicate an abnormality in the operations. Some of the common visual control tools are Andon, Standard operations sheets, Kanban cards, digital display panels, and storage and stock plates. Andon is the name of an indicator showing when workers stops the line or other issues that may occur in the process. As problems should surface right away, the operators can stop the process in the event of breakdown or delays at his or hers workplace. When this occurs, there is a call light being sent to the supervisor or maintenance worker and a red light will be turned on over the workstation. The board should be visible for everyone, and can indicate many different things relevant to the process, a common color code should be used that is known and practiced by all employees (Monden, 1998).

The standard operations sheet covers the three main elements of standardization mentioned in the previous chapter, the cycle time, operations routine, and the standard quantity of WIP. The sheet is posted at the workstation where each worker can easily see it. The worker can check the sheet to control if the work is being executed according to standards, if problems occur the worker must stop the line to resolve the problem. Kanban cards also serve as a visual control, if an item finds its way into storage without a Kanban it is a sign of overproduction, which should be investigated immediately. The absence of a Kanban is thus a signal for starting to look into the problem and to solve it. Furthermore, the Kanban indicates how many products are in the process, which helps in determining whether overtime will be needed. The production’s performance is also shown on digital display panels indicating the goals and the running count of units produced so far, or time spent of the allowed cycle time. The digital displays are seen be everyone so that all employees can help in keeping production on schedule. Store and stock indicator plates are placed over each storage location in the facility, indicating the standard quantity of stock. Each store plate is given an address representing the storage location, which is also stated on the Kanban. As a result the parts will be delivered to the right address and it helps in controlling inventory when the Kanban arrives to the storage location (Monden, 1998).

3.1.4 JIT and Queuing Theory

As have been described previously, pull is a technique for achieving better flow in the processes with the aim of being able to deliver JIT. At the same time, inventories should be reduced as much as possible while assuring high quality. The pull-based value stream is convenient for this purpose as it operates from a defined standard WIP between process steps and their buffers are only replenished with the same item that was removed.

If a push technique were to be utilized, capacity could be dimensioned to meet the demand over a period of time, however, as variation is inevitable, queues will build up unless investments are made in costly slack resources. Using queuing theory the two elements with variation would be arrival rate and service rate. A common way to think is that things will be done faster if they are forwarded in the process as soon as possible. However, as the system is getting heavier loaded it is also getting more sensitive to variation and inventories will pile up. Queuing theory says that the waiting time for a system will be considerably constrained as the utilization level is reaching 80 %. Figure 8 is an illustration of how patient waiting time at a hospital increases with the utilization level. The increase is almost exponential and reaches infinity when utilization gets close to 100 % (Palvannan & Teow, 2012). The analogy can be compared to a computer's CPU, when the utilization reach 100 % the computer becomes unresponsive. The same setup of capacity and total demand may still have different efficiency, since the variation in demand might be of different nature. In figure 8, curve A, B and C represents the same setup of capacity and total demand, but the variation is different.

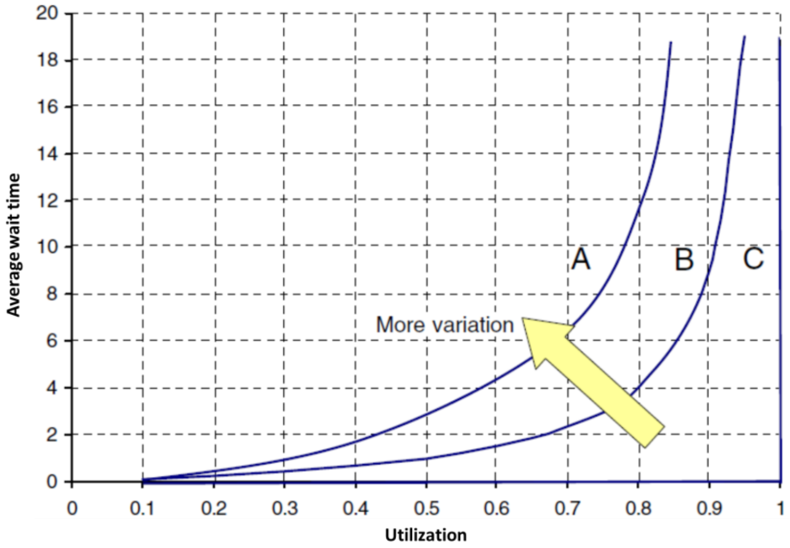

Figure 8. The waiting time increases exponentially as utilization rate increases (Palvannan & Teow, 2012)

Curve A has the highest variation and therefore can allow the lowest amount of utilization in order to offer the same waiting time as B and C. Curve C is the ideal state where there are no variation, the utilization can thus be close to 100 % (Palvannan & Teow, 2012).

Thus, having a push way of working may create queues in a production system and lead to high levels of waste. Furthermore, it is essential to understand the level of variation in a system in order to dimension the system for best efficiency and to be aware of potential effects from certain levels of variation.

3.2 Different Pull Systems

As has been described above, the pull system is important in the TPS to assure that only what the customers want is being produced. There are countless ways of implementing and facilitating a pull-based value stream in a process, though the methods mostly cited in literature include Kanban and CONWIP. Another interesting method, mainly invented for high variety production is the POLCA system. Each method in itself may have a variation in the set of rules controlling and facilitating the flow, depending on the characteristics of the environment, thereby creating new systems (Liberopoulos & Dallery, 2000). The different methods will be principally explained in the following chapters. A hybrid system of Kanban and CONWIP will also be described as it exemplifies a new dimension that can be established using a combination of rules.

3.2.1 Kanban

Kanban is the commonly used tool at Toyota for achieving the objectives of JIT production. Kanban literally means card, which is a visible record used as the means for communication, however, Kanban must not be a card for communication. The Kanban technique may just as well be executed via verbal command, a flag, a light, a hand signal or golf balls. Pull is in a way built into the Kanban system as a product is being held at a workstation until the next workstation in the process pulls the product from the preceding station (Esparrago, 1988). A Kanban system may have uncountable number of different designs and names for the signaling means, in order to fit the dynamics of the specific work environment. The most common design is to use two different kinds of Kanban, a withdrawal Kanban and a production-ordering Kanban. The withdrawal Kanban specifies the kind and quantity of the product that the subsequent process should withdraw from the preceding process, while the production-ordering Kanban specifies the kind and quantity of the product that the preceding process must produce. Monden (1998) explains the basic interactions between withdrawal Kanban and production-ordering with the following steps:

1. The carrier of a subsequent process goes to the storage of the preceding process with the withdrawal Kanban cards in his withdrawal Kanban post (receiving box) and an empty pallet (container). He does this at regular predetermined times.

2. When the subsequent process carrier withdraws the parts at storage A, he detaches the production-ordering Kanban cards that were attached to the physical units in the pallets (each pallet has one sheet of Kanban) and places these Kanban cards in the Kanban receiving post. He also leaves the empty pallets at the place designated by the preceding process people.

3. For each production-ordering Kanban detached, he attaches in its place one of his withdrawal Kanban cards. When exchanging the two types of Kanban cards, he carefully compares the withdrawal Kanban with the production-ordering Kanban for consistency.

4. When work begins in the subsequent process, the withdrawal Kanban must be put in the withdrawal Kanban post.

5. In the preceding process, the production-ordering Kanban should be collected from the Kanban receiving post at a certain point in time or when a certain number of units have been produced. It must be placed in the production-ordering Kanban post in the same sequence in which it had been detached at store A.

6. Produce the parts according to the ordinal sequence of the production ordering Kanban cards in the post.

7. The physical units and the Kanban must move as pairs when processed.

8. When the physical units are completed in this process, they and the production-ordering Kanban, are placed in store A so that the carrier from the subsequent process can withdraw them at any time.

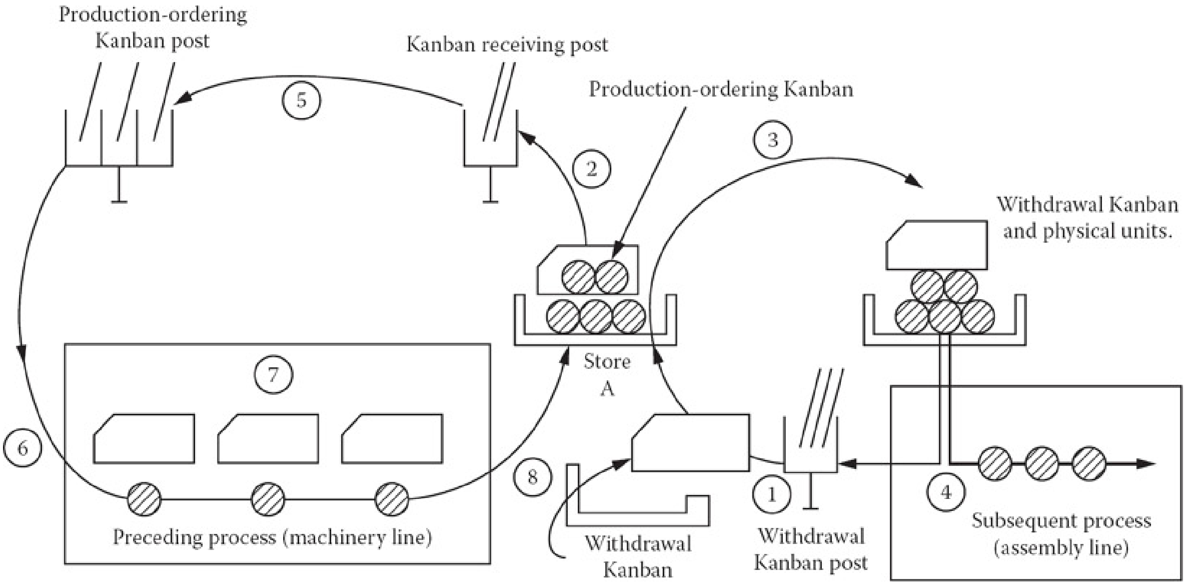

Figure 9. The interaction of withdrawal and production ordering Kanban in the Kanban system (Monden, 1998)

Having introduced the concept of Kanban and its principle interaction mechanisms, it must be noted that there are some prerequisites for a Kanban system (partly mentioned above) with the following conditions: leveling of the workload (principle 4), layout of processes, and standardization of jobs (principle 6). Furthermore, the system must follow a set of rules for it to work efficiently and for practitioners to get the desired benefits. The rules can partly be modified to different work environments for best fit, however, there are still a number of core rules that must be present and followed by everyone. Monden (1998) presents them as follows:

1. **The subsequent process should withdraw the necessary products from the preceding process in the necessary quantities at the necessary point in time.**

   The rule is also implying the following subrules:

   a. Any withdrawal without a Kanban should be prohibited.

   b. Any withdrawal that is greater than the number of Kanban cards should be prohibited.

   c. A Kanban should always be attached to the physical product.

2. **The preceding process should produce its products in the quantities withdrawn by the subsequent process.**

   The rule also implies the following subrules:

   a. Production greater than the number of sheets of Kanban cards must be prohibited.

   b. When various kinds of parts are to be produced in the preceding process, their production should follow the original sequence in which each kind of Kanban has been delivered.

3. **Defective products should never be conveyed to the subsequent process.**

   The rule also covers process, and not only products. A defective operation is a job for which standards are not fully attained, and inefficiencies exist in manual operations, routines, and/or labor hours. Defective operations are in the end also likely to produce defective products, therefore these defective operations must also be eliminated, and the standardization of jobs is thus an important prerequisite.

4. **The number of Kanban cards should be minimized.**

   Since the number of Kanban cards expresses the maximum inventory of a part, it should be kept as small as possible.

5. **Kanban should be used to adapt to small fluctuations in demand (fine-tuning of production by Kanban).**

   Fine-tuning is one of the Kanban system’s most remarkable features: its adaptability to adapt to sudden demand changes. The traditional scheduling companies has a central function for planning production and deliver the production schedules simultaneously to all departments, as changes comes in the planning department must revise the schedules and make new plans which has some lead time built into it. In the Kanban system, production schedules are not given to each preceding processes, each preceding process can only know what to produce based on the production-ordering Kanban cards. It is only the final assembly who receives a sequence schedule for a day’s production, so changes in demand is naturally incorporated in the line. If demand increases for the next month, since the change can be incorporated immediately, there is enough with a small incremental increase in the output for each workstation and day. This fine-tuning though can only adapt to small fluctuations in demand. Toyota recognizes that demand variations around 10 % can be taken care of by changing only the frequency of Kanban transfers, without revising the total number of Kanban cards. For bigger changes in demand, production lines may have to be rearranged to calculate a new cycle time and the number of workers may need to be changed. Otherwise, the total number of Kanban cards must be increased or decreased (may violate rule number four).

Another, slightly different, form of the Kanban system is the utilization of a bigger and more flexible storage point connected to more than two processes (one downstream and one upstream), called a supermarket. A supermarket is normally a line-side location to the production line where parts are sorted and stored. The supermarket has a fixed amount of inventory (raw material, WIP, finished goods). The fixed amount of inventory is stopping the upstream processes from replenishing when the supermarket is full and triggers it to start replenishing the supermarket as the downstream processes withdraws material from it (Cuatrecasas-Arbos, Fortuny-Santos, & Vintro-Sanchez, 2011).

3.2.2 CONWIP

CONWIP is a system that combines push and pull techniques in order to achieve the benefits from a pure pull system in situations where it otherwise would be difficult to actually implement a pure pull system (Marsh & Conard, 2008). As the goal of a pull system could be seen as minimizing WIP, the CONWIP (Constant Work-In-Process) could well be classified as one as it is a mechanism for limiting WIP (Hopp & Spearman, 2004).

CONWIP is supposed to be a more generalized form of Kanban, the three main characteristics that separate CONWIP from Kanban are (Spearman, Woodruff, & Hopp, 1990):

1. Use of a backlog to dictate the part number sequence
2. Cards are associated with all parts produced on a line rather than individual part numbers
3. Jobs are pushed between workstations in series once they have been authorized by a card to start at the beginning of the line

CONWIP, as Kanban, utilize trigger signals that can be either electronic or physical “card” depending on what suits the environment best. The card is attached to a standard container of parts at the beginning of a production line (the description will be related to a single production line), as the container of parts has been used in the end of the production line the card is removed and sent back to the beginning of the line where it waits in a card queue to be attached onto another standard container. The parts are in turn sent to the next production line or finished inventory storage. The standard containers of parts should contain the same amount of work in terms of time, so that each container has approximately the same processing time at the bottleneck. The cards thus travels in a circuit throughout the whole production line meaning that cards are assigned to the specific production line and not to a specific part as in Kanban. Part numbers are instead assigned to the cards in the start of the line, using a backlog list dictating the sequence of production of the parts. As work is needed in the first process step, a card is removed from the card queue and marked with the part number in the backlog for which raw material (or components) are available so that production can commence. The time of the part number match (part to be produced) is noted on the card as the *system entry time*, this time is used to maintain the queue discipline at all process step which is first in, first out (FIFO), meaning that the lowest system entry time should be processed first. The backlog list is the "planning function" of the CONWIP system and responsible personnel should maintain it. The backlog could be generated from a master production schedule and orders could also be added as received to the production line. Expediters could rearrange the backlog and add part numbers to it in order to ease fast prioritization. The most important criterion is that work can never be started without a card present in the queue, even if the first process step is idle, in order to limit and control the WIP. The principle dynamic of the CONWIP system can be seen in figure 10.

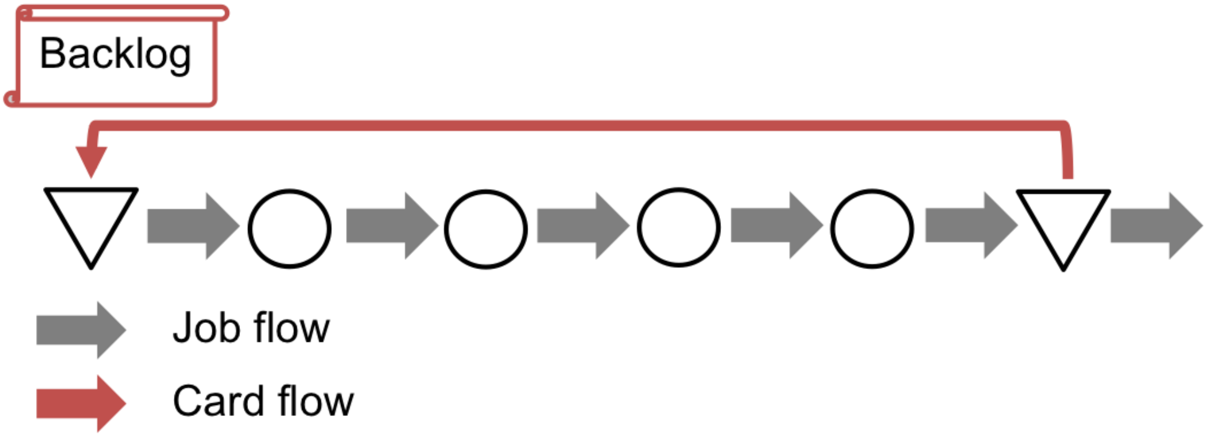

Figure 10. Dynamics of the CONWIP control system using a backlog as a planning function

Obviously, the CONWIP system has a number of parameters that must be calculated and maintained for the system to run effectively and to offer the desired positive effects (Spearman, Woodruff, & Hopp, 1990):

m,	The card count. Determines the WIP level.
q,	The production quota. The desired production quantity.
n,	Maximum work ahead. If $q + n$ is produced during a period, the line is stopped until the start of the next period.
r,	Capacity shortage trigger, as a function of the actual production up to a time t, A(t). The trigger function indicates that that additional capacity is needed. A trigger could be a constant allowable shortfall, i.e. not being able to produce enough during a certain time period. Thus, capacity additions are triggered at the end of a production time period (T), if $A(T) < q - r$. Another trigger could be a probability distribution function, e.g. based on A(t) and the current status of the

production line or other resources relevant for production (e.g. machine failure, personnel absence etc.).

These parameters should be set according to objectives and requirements of the company, including economical trade-offs. Increasing *n* and/or *m* tend to increase service level at the expense of raised inventory cost. Increasing *q* tend to increase expected revenues, given there is buyers for the increased volume, while decreasing service. The trigger function aids in determining service level and costs associated with additional capacity. The balancing of the parameters are subject to optimization, however, a careful initial approach is to set inventory parameters, *n* and *m*, higher than necessary to assure deliveries and then lowering these over time. The system will be constrained by the bottleneck, assuming there is sufficient demand, a correct issue of number of cards will maintain just enough WIP to keep the bottleneck busy at all times. If standard containers of work are piling up in the bottleneck station then cards are not arriving to the end of the line and new work will not be started, holding the WIP constant (Spearman, Woodruff, & Hopp, 1990).

CONWIP have been shown to be effective only for production environments with low process and demand variability. Low process variability in this case refers to that the steps to complete an output did not vary much from one output to the next and demand variability refers to changes in demand from one time period to the other (Marsh & Conard, 2008). A situation where CONWIP is beneficial compared to Kanban is for production lines producing many different parts, as there is not enough room to have standard containers of each part number in the factory, if there were, WIP levels would be higher than necessary. CONWIP solve this through the use planning function of the backlog and standard containers of work and by controlling the sequence of work (Spearman, Woodruff, & Hopp, 1990).

3.2.3 Kanban-CONWIP Hybrid

An issue with the CONWIP system is that it only limits the WIP for the whole production line, and not for the specific workstations within it. The problem is that there will be a big pile up of inventory at the bottleneck (or failed machine), which will not be removed for a long time, and the production line will continue to run until the cards are all stuck at the bottleneck. An option to avoid this is to introduce a limit for the inventory on upstream machines from the bottleneck, so they are not pushing too much work ahead. This is similar to the principle of Kanban where there are intermediate inventory limits that are controlled using Kanban cards, as in N1 and N2 in figure 11. As a product is made in the last production stage, N3, it is being sent to the finished goods buffer while the Kanban cards are sent to the first production stage to authorize the release of another part into the system, as in a CONWIP system (Bonvik, Couch, & Gershwin, 1997).

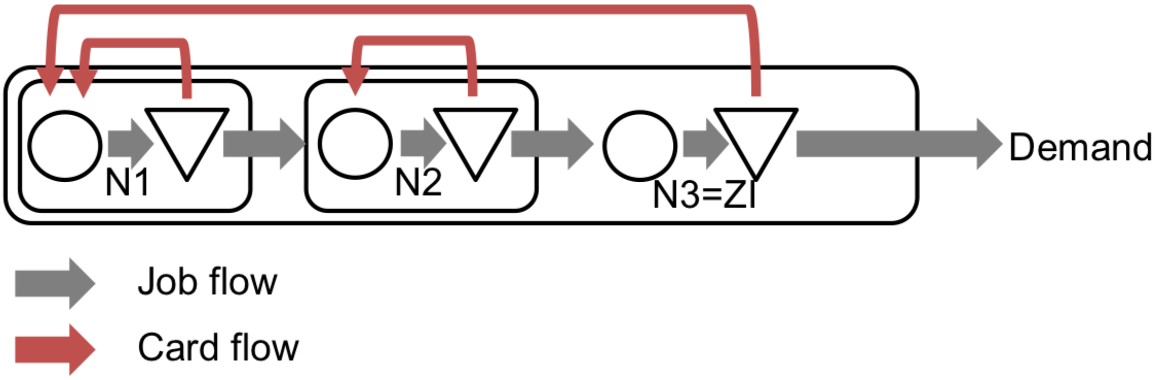

Figure 11. The dynamics of a Kanban-CONWIP hybrid, having cards between the different workstations to limit local WIP

3.2.4 POLCA

POLCA is short for "Paired-cell Overlapping Loops of Cards with Authorization". Like CONWIP, It is a hybrid push-pull strategy. POLCA was developed by Rajan Suri to effectively cope with high variety or custom engineered products (Suri & Krishnamurthy, 2009). According to Suri and Krishnamurthy (2009), there are some implications of using a pull/Kanban system when a company has a wide variety of products with highly varying demand or if they make custom engineered products in small batches:

- Proliferation of WIP inventories at each stage of the process
- There are no predefined finished outputs that might be stored
- Traditional pull/Kanban systems require set takt times, which can be impractical when products are custom or demand is highly variable

Traditional push or Material Requirements Planning (MRP) systems also have clear drawbacks including the creation of excess inventories and long lead times (Suri & Krishnamurthy, 2009). POLCA is a combination of pull and push that incorporates benefits of both systems. As the name suggests, the system is based on production in cells that are focused on subsets of the process. Because of the nature of the products being made, orders might have different demands and different routing within the cells. This is what makes neither level scheduling nor takt times applicable (Suri & Krishnamurthy, 2009). In the POLCA system, the routing of products is instead controlled by both centrally controlled release authorizations and production control cards called POLCA cards. The release authorizations are issued by a high-level MRP system, otherwise known as a HL/MRP system. This system does not interfere with the operational level but considers the cells to be black boxes and helps in planning the flow between cells by issuing a routing sheet for each order. When an order is received, the HL/MRP system creates release authorization times at each cell. However, the release authorization time only authorizes the beginning of the work, but the cell will not start work unless a corresponding POLCA card is available. The POLCA card is communicating and controlling movements between the cells, making sure that nothing is sent forward until there is free capacity ahead. Within the cells, the different stations have the freedom to use other procedures (for example a regular Kanban system).

To understand how the POLCA system works, imagine a production process with four basic steps: fabrication, painting, assembly and shipping (see figure 12). Within each step there are cells specialized on certain aspects of the activity. The routing is decided to be; fabrication 2 (F2), painting 3 (P3), assembly 1 (A1) and shipping 1 (S1) by the HL/MRP system and work is initiated in cell F2. This order will proceed through cell loops F2/P3, P3/A1 and A1/S1. Every pair of cells will have a corresponding POLCA card, for instance a F2/P3 card. The POLCA card will stay with the job during its way through both paired cells and then it will be sent back to the upstream cell. For example a F2/P3 card will be attached to the job when it enters F2 and stays with the job throughout the F2 cell and is then sent to the P3 cell and stays with the job throughout the P3 cell. Also, a P3/A1 card will be attached so that both the downstream and upstream card follows the product throughout the cell, hence the name *overlapping loops*. When the P3 cell finishes its operation, the F2/P3 card will be detached and sent back to the F2 cell. Note that the P3 cell can only send the job forward to the A1 cell if they have received a P3/A1 card from the A1 cell. The process continues in this manner until the order is delivered.

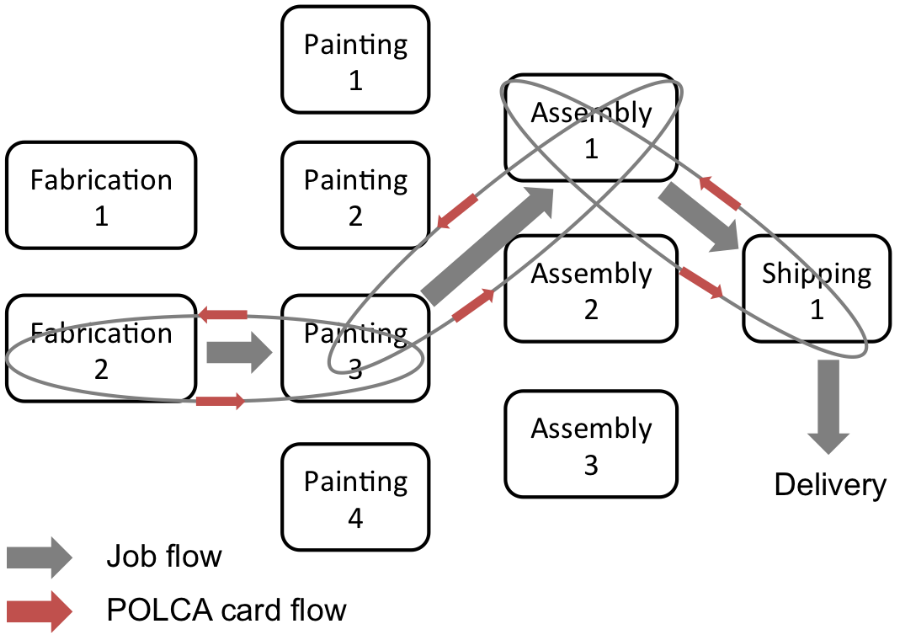

Figure 12. POLCA card flow through F2, P3, A1 and S1

The POLCA system assures that the work being sent forward has a receiving cell that is ready to start working (Suri & Krishnamurthy, 2009). Since a returning POLCA card signals free downstream capacity, the POLCA card is a capacity signal while the Kanban card is an inventory signal. If a POLCA card is not received, this means that some downstream cell is backlogged. However, a returning card does not say which job that should be started. Instead, the upstream cell looks at the list of authorized jobs from the HL/MRP system. By only controlling a higher level (cell) the POLCA system allows the lower level (within cells) to be flexible to workload variations. By issuing authorization times, the system also prevents avoidable inventory that is not associated with specific orders. So, a card might be received but if there are no authorized orders, no job will be started.

In a Kanban system, Kanban cards tightly couples workstations. The POLCA cards flow in longer loops that allows for more flexibility. A Kanban system needs to be tuned to a specific takt time and can only handle rather small variations from the determined rate (around 10 %, see chapter 3.2.1). By having longer loops, the POLCA cards can be used to adjust inventory levels to deal with variations in demand or product mix. Each cell can see both the upstream and downstream workload and can react to variations by adjusting their own capacity.

Suri and Krishnamurthy (2009) present four main phases for implementing a POLCA system:

- Pre-POLCA assessment.
    - Conduct a needs assessment to see whether any of the involved cells require lead time or capacity planning before implementing POLCA
    - Check the prerequisites: a planning system that can obtain rough capacity lead time estimates of different cells and a cellular organization
    - Set goals and metrics for evaluating the implementation
- Design the POLCA system
    - Identify POLCA loops by analyzing the routing for different products
    - Compute release authorizations
    - Determine the amount of work that a POLCA card represents
    - Design the POLCA card and document the POLCA procedure
    - Compute the number of POLCA cards in each loop:

        - $\#_{A/B} = (LT_A + LT_B) \times (NUM_{A/B}/D)$

          With:

          $\#_{A/B}$ being the number of A/B cards

          LT_A being predicted lead time for cell A

          $NUM_{A/B}$ being the total number of jobs between A&B

          D being the planning period (in days)

- Launch the POLCA implementation
    - Determine a POLCA champion
    - Train and educate all operators involved in the implementation
    - Have frequent reviews and management support
- Post implementation evaluation
    - Track key metrics including throughput from different cells and product lead times
    - Measure qualitative metrics like stress level, morale, communication and employee satisfaction

3.3 Summarizing the Requirements of Pull

Resulting from the theoretical framework is a number of key insights with regards to pull-based value streams. Pull is essentially a technique for moving products from one workstation to the other. The objectives of the pull system is to assure deliveries to the customer just the way he or she wants it, while minimizing the WIP and related costs. Since pull is a technique, it can be applied to basically all kind of operations, however the effectiveness and desired performance of the pull system requires a set of conditions. First, there must be a clearly defined agreement between the two operations sharing a refill/withdraw-action that they are connected to one another, furthermore the agreement must also cover what product mix to be shared, the sequence of the model mix and volume limits for the products. Second, all items and resource needed for the interaction between two parties must be explicitly dedicated to them. This includes personnel, storage locations, and a common reference takt time, for example should personnel not be working on other stations or putting items on different storage locations. Third, there must be a built in control that is easy to see and use between all interacting parties. Having the control mechanism visual and physical, it supports in maintaining the definition and dedication between parties.

These three general preconditions set the stage for what needs to be present and improved. Going slightly deeper into details; the critical principles that are connected to the three preconditions are smoothing of production, standardization, and visual control. The smoothing of production is essentially important in leveling out the product mix, the volume, and the sequence of producing the product mix to make the demand more predictable. This is one piece of the puzzle in being predictable. Another part of the puzzle is standardization that mainly refers to three key elements. First, there needs to be a standard takt time in order to balance the production line in the best way. Second, there must be a standardized operations routine, defining in what order the different operations are to take place and also in what way the tasks should be performed. Third, there must be a standard level of WIP specified for each station. The last part of the puzzle is to control the flow and making it visual for all employees. Visual control encourages people to take corrective actions as well as it provides a quick built-in feedback loop. Being able to incorporate the smoothening and standardization in an organization, together with a visual control system, will enable the successful implementation of a pull system.

3.4 Differences between White- and Blue-Collar Work

Historically a white-collar has been an office worker and a blue-collar has been a manual worker (Hopp, Iravani, & Liu, 2009). In this thesis, for simplicity, knowledge workers and white-collar workers are considered to be the same. Other authors, such as McNamar (1973) and Ramirez and Nembhard (2004), also equate knowledge workers and white-collar workers. Hopp, Iravani and Liu (2009) are contrasting white- and blue-collar tasks by the use two dimensions: Intellectual vs. physical and creativity vs. routine. These two dimensions may be visualized in a diagram (see figure 13).

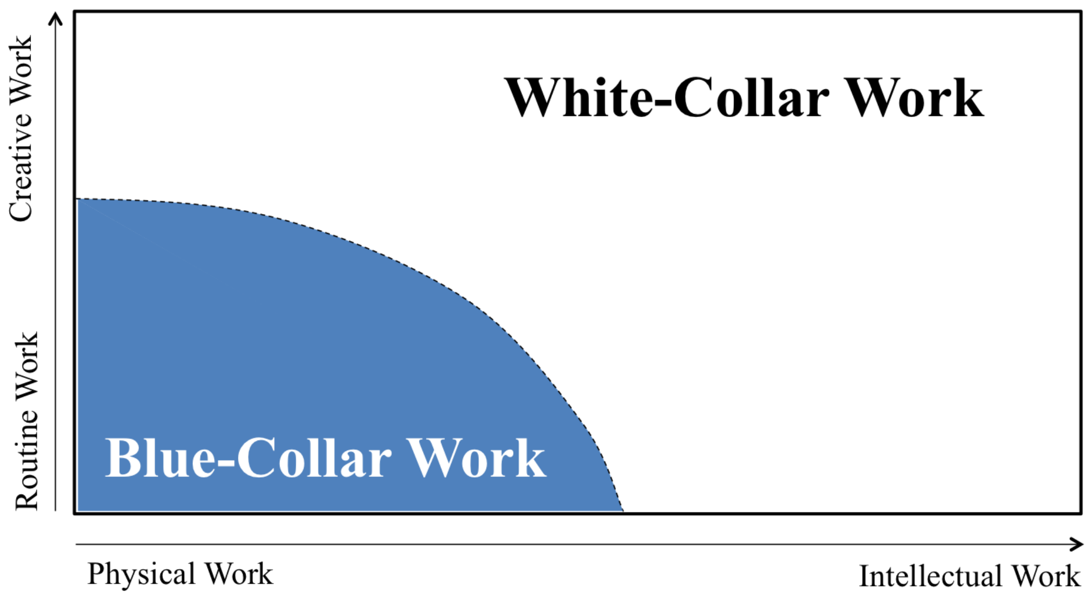

Figure 13. Visualization of the difference between white- and blue-collar work according to Hopp, Iravani and Liu (2009)

Defining the difference between white- and blue-collar work along these dimensions means that there is a wide variety of tasks considered white-collar work. The boundary between white- and blue-collar work should be considered inexplicit since, by using this definition, there is no such thing as pure white-collar work and pure blue-collar work. In fact, almost any kind of work consists of both white- and blue-collar tasks (Ramírez & Nembhard, 2004).

Gonzales-Rivas and Larsson (2011) describes the difference between classical factory work and white-collar work along two other dimensions: *Invisibility* and *elusiveness*. They focus on the use of electronic information and how this reduces visibility as compared to a factory or a paper-based office. This, in turn, has implications to what may be considered waste. For example, excess travel time is considered waste in the factory but is not nearly as important in the electronic office since electrons move so fast that distance becomes irrelevant. Figure 14 visualizes a spectrum “from atoms to bits”. It should be noted that Gonzales-Rivas and Larsson (2011) defines a knowledge worker as “a heavy computer user”.

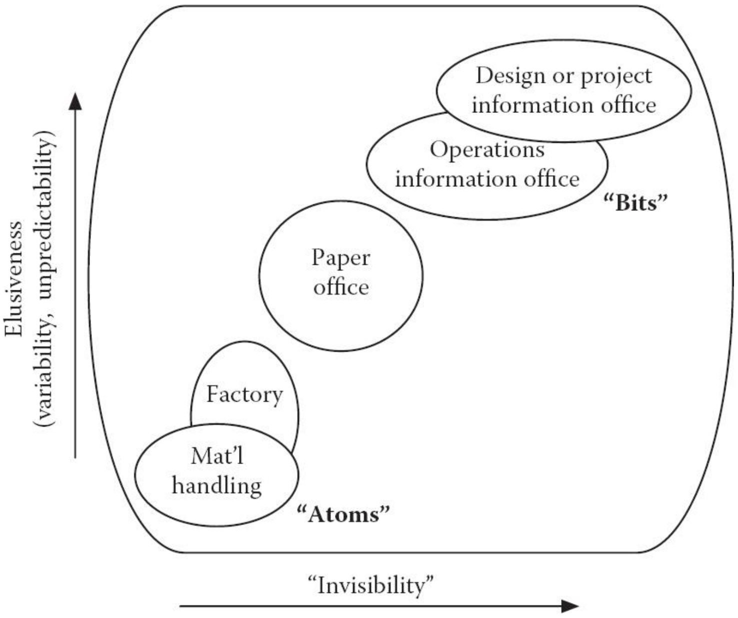

Figure 14. Representation of the difference between the factory floor and information office, from Gonzales-Rivas & Larsson (2011)

Similarities between white- and blue-collar settings include queuing behavior, that is, entities pile up while waiting to be processed by a worker with finite capacity. This in turn means that variability and high utilization will cause queues. In white-collar processes, a high WIP level is not about holding costly physical goods as in manufacturing, but time delays until the output reaches the customer (Marsh & Conard, 2008). Higher WIP levels will result in longer total processing time (response time) according to Little's law (Little, 1961). Hopp, Iravani and Liu (2009) discuss the similarities and differences between white- and blue-collar work in three levels: individual, team and organization. The main differences between blue- and white-collar work at an individual level as presented by Hopp, Iravani and Liu (2009) include:

- White-collar tasks generally demands more accumulated domain knowledge and are rarely identical

- White-collar tasks relies more on knowledge-based resources

- Learning is slower but more important in white-collar tasks

- It is difficult to measure output from white-collar work systems

- White-collars are more likely to involve self-generated work rather than just work received from the outside. This requires a higher degree of creativity

- White-collars tend to have more discretion over processing time and often uses personal judgment to determine if a task is complete

- Incentives are more critical in white-collar systems than in blue-collar systems (especially non-pecuniary)

The fact that white-collars usually have more discretion over processes is a fundamental difference that may have implications when applying results from blue-collar research on white-collar settings (Hopp, Iravani, & Liu, 2009). Blue-collar workers typically have well-defined, straightforward tasks and research usually assumes that workers have no or very limited flexibility. Also, some of the value created in a white-collar process might be impossible to detect upon completion, so-called latent value. Activities identified as waste in production, such as rework and loopbacks, may be considered creation of new knowledge that might be very valuable in development work (Holmdahl, 2010). This makes it hard to determine what may be considered waste and what may be considered value adding. All in all, these factors make it hard to measure actual performance of white-collar processes. For example, some of the conventional performance measurements, used in blue-collar settings, such as utilization, becomes difficult to apply in white-collar settings since the high amount of discretion makes it possible that all white-collar workers are 100% utilized. Instead, the key issue is how time is allocated, not how busy individual workers are.

According to Hopp, Iravani and Liu (2009) learning and incentives becomes even more important on a team level. Apart from this, they present some additional differences that requires consideration at a team level:

- Interdependencies are more important and complex among white-collar team members

- Behavioral issues are more important within white-collar teams

Since blue-collar tasks usually are well defined and routine, interdependencies within blue-collar teams are often simple and explicit (Hopp, Iravani, & Liu, 2009). White-collar team members on the other hand, often relies on frequent interactions to gain the necessary information to be able to cope with the complex and loosely defined tasks they are faced with. Behavioral issues are important in both blue-collar teams but even more important in white-collar teams. Trust is especially important in white-collar teams since interdependencies are more complex, processes and outcomes are uncertain and outcome measurement is ambiguous.

Looking at an organizational level, Hopp, Iravani and Liu (2009) discuss the temptation of modeling white-collar organizations as flow networks, as commonly done in blue-collar organizations. One big issue with doing so is that white-collars working with non-routine, intellectual work constantly needs to acquire additional knowledge dispersed in the organization. This means that there is a flow of information, often complex and informal, in addition to the formal workflow (Huberman & Hogg, 1994). In other words, there are both deterministic and probabilistic links between teams in a white-collar organization. Adler et al. (1995) argue that these systems may be represented in stochastic networks and models a product development organization in this manner. Since there are both formal and informal networks present in a white-collar organization, Hopp, Iravani and Liu (2009) suggest some additional issues that are important to consider when studying white-collar work at the organizational level, including:

- Proven methods from blue-collar settings, such as standard operating procedures that does not take knowledge and information as inputs, cannot be used directly on white-collar settings

- New, flexible systems are needed in order to control the flow and work assignment in white-collar settings

Since white-collar work usually involves creativity and intellectual content it is generally ill-suited to use control mechanisms developed for blue-collar work (Hopp, Iravani, & Liu, 2009).

3.4.1 Control Systems in White-collar Settings

Control systems usually take the form of either process-based control or outcome-based control (Hopp, Iravani, & Liu, 2009). Process-based control usually requires standardized work and is the traditional choice to control blue-collar systems. White-collar settings are usually better controlled by outcome-based control, but some authors (Nidumolu & Subramani, 2003; Turner & Makhija, 2006) argue that, if designed correctly, process-based control might increase performance in white-collar systems. Nidumolu and Subramani (2003) found that standardized performance measures and high discretion in the actual tasks improved performance in the software development firms that they studied. How well suited it is to use process-based control in white-collar settings seems to depend on how well the trade-off between discretion and standardization is balanced and the features of the knowledge involved in the white-collar work.

The features of knowledge can be divided into:

- Codifiability: If knowledge is codifiable it is rather simple to break down the process into parts and standardize

- Completeness: If knowledge is complete, all the necessary knowledge is available to the worker, meaning there is less uncertainty involved.

- Diversity: Indicates the breadth and relatedness of knowledge. If knowledge is less diversified, more standardization may be applied

In other words, if knowledge is codifiable, complete and less diversified process-based control is feasible in white-collar systems. If the opposite is true, outcome-based control might be the only option (Hopp, Iravani, & Liu, 2009). Information is another factor that differentiates white- and blue-collar control systems. Information usually flows in a sequential manner in blue-collar settings. White-collar processes are usually more dependent on information management and flexibility since information flows both sequentially and reciprocally. This makes coordination more complex.

3.5 The Cynefin Framework

Snowden and Kurtz presented the so-called Cynefin framework in 2003. The framework is what they call a “sense-making framework”, which basically means that its true usefulness is not as a basis for logical arguments or empirical verification but for its effect on sense-making and decision-making (Kurtz & Snowden, 2003). The framework describes five possible states: *known* (simple), *knowable* (complicated), *complex*, *chaos* and *disorder*, with disorder being placed in the middle (see figure 15). By contrasting between these different domains, and reflecting on to which domain a certain situation “belongs” to, more effective strategies might be used. The two right hand domains, *known* and *knowable*, is what Kurtz and Snowden calls ordered states and the two left-hand domains, *complex* and *chaos*, are called un-ordered states.

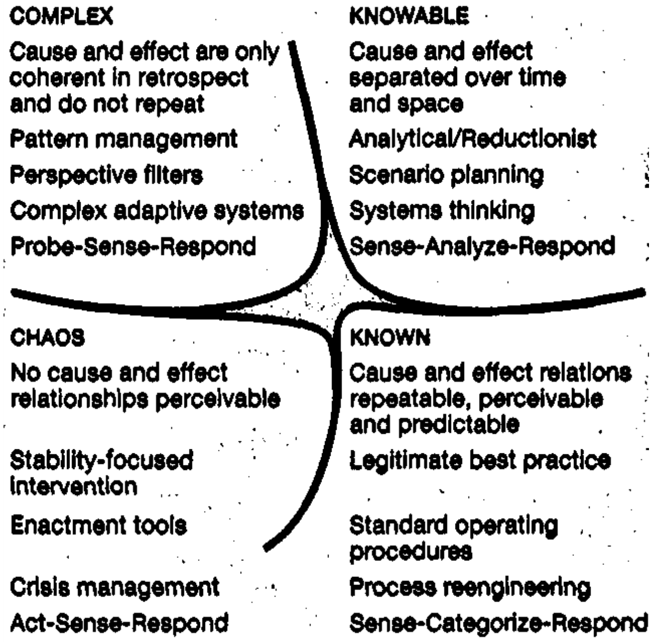

Figure 15. The Cynefin framework (Kurtz & Snowden, 2003)

In the known domain, cause and effect relationships are known and often linear (Kurtz & Snowden, 2003). Predictions may be made and situations are repeatable. Because of this, best practice may be used. Also, in the known domain, focus is on efficiency. In the knowable domain, cause and effect relationships exist and are stable but they are not fully known or understood. Knowledge about the relationships might for example be confined to a specific group of experts. In the knowable domain, focus is on methodologies that might find and define cause and effect relationships. Everything that is situated in the knowable domain might be moved to the known domain by mapping these relations. In the knowable domain, structured techniques are desirable but all assumptions need to be examined and challenged. If wrongful assumptions are in place this may result in false conclusions that are difficult to discover. Good practice may be used in the knowable domain, but there is no real best practice.

In the complex domain cause and effect relationships between agents exists, but the number of agents and the number of relationships cannot be analyzed (Kurtz & Snowden, 2003). Patterns may be perceived but not predicted, that is, patterns might be recognized in retrospect, but not foreseen. Patterns may repeat, but there is never any guarantee that they will continue doing so. Therefore, there is a risk of misjudging what seems to be logical patterns and to take improper action. It is important to realize that the tools and methods successfully used in the known and knowable domains do not work in the complex domain (Kurtz & Snowden, 2003). In the chaotic domain, the system is turbulent and there are basically no relationships between cause and effect. Trying to apply best practice is of no use and it may be the reason that the system became turbulent in the first place.

The final domain is that of disorder. This domain is crucial for understanding why the Cynefin framework is effective (Kurtz & Snowden, 2003). Different people tend to think of their situation as belonging to one of the other four domains depending on where they are most comfortable. For example, some people may assume that they are in the knowable domain and strive to map cause and effect relationships because they are most comfortable with that situation. Others may think of the same situation as complex because that is their preference. The more important the issue, the more people seem to pull in the direction of their preference. The domain of disorder is basically about not knowing what domain the situation really belongs to. Increased collaboration may lead to a decreased size of the disorder domain.

One of the basic functions of the Cynefin framework is to relax the assumption of order by pointing out that the assumption of order holds true only for the known and knowable domains (Kurtz & Snowden, 2003). In other words, by not assuming that all systems are or can be ordered, resources may be saved. Much of the development of management science has been rooted in assumptions that systems are ordered and that by applying enough time and resources cause and effect relationships will be known. Just consider how often best practice is used to improve specific contexts. By relaxing the assumption of order managers can realize that not all effective solutions are efficient solutions, in other words, that less energy might be spent achieving the same results when the means suits the context.

3.6 Applying the Lean Principles in an Office Environment

Even though there are many success-stories regarding the implementation of Lean in manufacturing, administrative processes within manufacturing companies often struggle to get the same results (Locher, 2011). In some companies, their Lean service work is limited to implementing "5S" techniques while they fail to implement key Lean concepts such as standard work, flow and pull. According to Locher (2011) the usual suspects for making these organizations fail is high variability of the work, multitasking, unpredictability of demand and the creative nature of the work. However, the companies themselves usually create these pitfalls by having lack of standard work, processing work in batches and so on. According to Blackburn (1992), the batch size in white-collar environments is the scope of the problem that is being solved or the amount of information to be processed. According to Gonzales-Rivas and Larsson (2011) implementing Lean becomes a bigger challenge the more elusive and invisible the work gets (see figure 14). Blackburn (1992) also stresses lack of process traceability as a major reason why implementing Lean in white-collar environments might be difficult. Chen and Cox (2012) highlight high variation, less foundational information (current state information), lack of literature references, lack of cooperation between departments and lack of directive from top management as major reasons for why it is difficult to apply Lean production principles in an office environment. Although there are some clear challenges in implementing Lean in white-collar processes, all of the above mentioned authors think that these challenges can be overcome, and that the potential rewards outweigh the efforts of doing so.

3.6.1 Standard Work

Many of the principles around standard work are the same for white-collar work as it is for blue-collar work (see chapter 3.1.3.3). Basically, it is the best-known way to perform an activity (Locher, 2011). It makes the process perform consistently and therefore ensures that the output is of consistent quality. If no standard work is defined it is impossible to identify when nonstandard conditions arise and to trigger actions to correct and improve. Documentation of standard work should be visible and simple and should, according to Locher (2011), not to be confused with Standard Operating Procedures (SOPs). SOPs should describe the process in much greater detail compared to standard work and it may be used for training purposes. Standard work should be used as a reference for someone who has already been trained. A common argument against the implementation of Lean in white-collar organizations is the high level of variability, but much of the variability experienced in white-collar organizations is due to lack of standard work (Locher, 2011). Therefore, this argument introduces a logical paradox.

The main elements of standard work are:

- A definition of the task to be performed (*what*)
- *How* to perform the task
- *Why* the task should be performed
- The expected *time* to complete the task and the *timing* when the task needs to be completed

Standard work can be applied to any repetitive process, including those performed in white-collar environments (Locher, 2011). In fact, since there is usually little or no previous standard work in white-collar environments the benefits might exceed those found in manufacturing.

3.6.2 Creating Flow

Locher (2011) presents three alternative ways of approaching flow in white-collar environments: combining activities into one role, creating a continuous flow and performing activities in parallel.

When combining activities, the key is to minimize the number of hand-overs and by that minimizing the chance of queues forming. One example of combining activities is the redesign of the packaging flow at VCCS described in chapter 2.3.1.3. Queues are especially likely to form in white-collar environments since the work usually involves multitasking. One possible disadvantage of combining activities is that it can require a lot from single individuals if the activities to be combined is complex. The combination will require a lot of learning and the activities need to be chosen such that it is plausible for one person to perform.

When creating a continuous flow the number of people performing each task and the number of tasks usually remains the same. However, the people involved in the process are working together, often collocated, approaching one-piece flow. It is key to have balanced work between the different people involved. Some of the benefits associated with this kind of flow are reduced lead times due to minimal or zero queues and improved quality due to more direct and timely communication and reduced processing time (Locher, 2011). Reduced processing time is often due to reduced non-value-adding work that has become a natural part of the task. When communication and understanding increases, this non-value-adding work can be identified and reduced. A common argument against this approach is that management becomes more difficult due to reduced control by functional managers. However, the reality is that the organizational charts usually have little to do with the process flow, on the contrary, they might actually impede flow. Instead, a value stream manager should optimally manage the process. If this change is not plausible, it is possible to organize by value stream without changing the reporting structure but it requires the support of existing functional managers. Another argument against this approach is less interaction between people with similar knowledge and skills, or "specialists". This might be a real disadvantage if not managed properly. A possible simple solution is to have periodic meetings and forums within the functions.

The third way to create flow as presented by Locher (2011) is to perform activities in parallel, in other words, concurrent processing. Many white-collar processes cannot be performed in parallel. If they are to be performed in parallel, managing the process will most likely be more difficult since convergence points need to be balanced. If concurrent processing is possible and the management difficulties are dealt with, it can provide breakthrough results in lead time reduction.

Locher (2011) presents six steps to designing a flow system in white-collar environments:

1. Identify the activities involved
2. Determine the demand rate (or takt time) of these activities
3. Determine the resource requirements by dividing the process time by the takt time for each activity
4. Identify roles, including standard work
5. Determine the need for training. Once the required training is completed, the envisioned flow system can be implemented
6. Develop visual management techniques that will be used to manage and sustain the flow

Typical results of a successful flow system implementation have included (Locher, 2011): Lead time reductions of 50 % to 90 %, process time reductions of 20 % to 40 %, reduction of defects by 25 % to 75 % and increased employee satisfaction.

3.6.3 Pull

Pull in white-collar organizations is a mean for controlling the flow of resources based on demand or consumption. In this case, resources are information or people. According to Locher (2011), the general rule is to first apply flow concepts and then apply pull where it is deemed necessary. However, there are exceptions and pull can often be less complicated to implement since it usually requires less physical reorganization.

When discussing pull in white-collar settings, Locher (2011) separates two basic (combinable) forms of pull systems– supermarket and sequential. Supermarket systems have storage points after each process that holds a certain amount of each output. When the downstream process consumes a certain kind of output the upstream process replenishes the storage according to a set of rules. Sequential systems also have storage points located after the process but not necessarily one for each output, that is, the content of the queue can vary over time. The upstream process replenishes the storage point according to the status of the queue and a set of rules.

Sequential pull systems tend to be more suitable for white-collar value streams than supermarket systems since it is usually not practical (or possible) to have a given amount of every type of information the process can produce on storage (Locher, 2011). In sequential pull systems, a limit is set on each queue. When this limit is reached a decision needs to be triggered. The limit can be a maximum or minimum limit. When a maximum limit is reached it typically means that the upstream process should stop producing a particular kind of information since the current level exceeds demand. When a minimum limit is reached it typically means that upstream process should return to producing a particular kind of information since the downstream process is waiting for it. In a sequential pull system it is possible for the downstream process to decide what work it wants from the queue. Therefore, clear rules that govern the desired sequence along the value stream needs to be established. Examples of these kind of rules include FIFO and “due date”. Rules like these will allow the flow to be basically self-managed.

All pull systems have some common characteristics (Locher, 2011):

- **Visible queues.** The more information that is electronic, the harder this point gets. Often, if a process is depending on electronic information, electronic tools need to be developed.

- **Defined limits for queues.** When defining these limits, organizational goals for finishing the particular process and the required lead time (ultimately set by the customer), needs to be taken into consideration. The total amount of WIP needs to be limited to reach these goals. Also, the amount of WIP should not vary significantly since this means that the lead time throughout the value stream will vary as well. The key is to establish a suitable unit of measure for the queue.

- **Defined rules for what happens when these limits are met.** These decision rules might be to relocate resources, ensuring that a certain sequence is followed etc. The important thing when determining these rules is to consider what level of flexibility that is required and possible. Flexibility can be hampered by lack of standard work and cross-training. Low flexibility will result in limited possible decision rules. Standard work and cross-training therefore is crucial to implementing a pull system in a white-collar environment.

- **Use of worker managed, visual signals.** The desired decisions needs to be visually triggered somehow. In Lean production, Kanban cards are a common way to trigger a decision. Kanban cards however, does not need to be actual cards but might as well be other forms of signals.

In order to implement a pull system in an office environment, Locher (2011) suggest the following steps:

1. Identify where queues are expected to form
2. Identify means to make queues visible
3. Establish limits for queues
4. Define rules for queues (actions when limits are met as well as preferred sequence)
5. Train people in the pull system and initiate
6. Monitor the system

Benefits of successfully implementing a pull system includes greater predictability, easier management, increased flexibility and increased cross-training (Locher, 2011).

According to Marsh and Conard (2008), successful application of "pure" pull systems is limited to processes with repetitive output. There are some success stories from when pull systems have been implemented in white-collar processes with repetitive output, including Conant (1988), Feather and Cross (1988), Blackburn (1992), Swank (2003) and Hamm (2005). The nature of many white-collar processes, with high variability and low output repetitiveness, makes it interesting to compare it to Made-To-Order (MTO) manufacturing. In MTO manufacturing the output is usually customized and non-repetitive. According to Stevenson et al. (2005), traditional pull systems have been proven not suitable in MTO manufacturing with high routing variability and lack of repetition. When the output is customized, or made to order, a "pure" pull system is not feasible since it is not possible, or desirable, to stock customized products

(Marsh & Conard, 2008). It has been shown that push/pull-hybrid systems are the most effective in MTO manufacturing. CONWIP and CONWIP hybrid systems (see chapter 3.2.2 and 3.3.3) are best suited when process and demand variability is relatively low. POLCA (see chapter 3.3.4) seems to be a workable solution for MTO manufacturing. Although all of these systems have both push and pull elements many authors, such as Hopp and Spearman (2004) and Marsh and Conard (2008) would classify them as pull systems since WIP levels are limited.

Traditionally, white-collar tasks are pushed through the organization, with the prioritization often being the delivery's due date (Marsh & Conard, 2008). However, the due date of white-collar tasks is often "as soon as possible" which results in increased multitasking, i.e. more WIP and a slower response time. Marsh and Conard (2008) had interesting results when testing a "POLCA-like" system for pull delegation of non-repetitive white-collar work. In their study, subordinates pull tasks when they have spare capacity and by doing this, they control their WIP level. The pull system they tested showed a significant improvement in response time but no improvement in quality as compared to a traditional push delegation system. The response time decreased by 9%, which in practice means a reduction of about 22,5 days per year and employee. One important result of the study was that there was no significant impact on stress or satisfaction, meaning the response time improvements is basically without "cost". It should be noted, however, that the test was conducted in a laboratory environment using students and that it included only one workstation (not a sequence of several workstations).

3.7 Summarizing Lean in White-Collar Settings

There are several definitions of what white-collar work really is. Most authors recognize non-manual work, i.e. non-physical work, and high variability as important factors. The model presented by Hopp, Iravani and Liu (2009) is a good representation of what is meant by "white-collar worker" in this thesis. That is, a white-collar worker is occupied with primarily non-physical work and might have both routine and creative tasks.

White-collar work differs from blue-collar work in several ways. Some authors have chosen to distinguish along the dimensions mentioned above, others along the dimensions "elusiveness vs. invisibility". Summarizing these and other views, the most prominent differences include:

- Higher routing variability
- Higher demand variability
- Higher output variability
- Less repetitive tasks

- Non-physical and low visibility of products
- Higher task discretion
- Stronger interdependencies (need for trust)
- More difficult to measure output (including latent value)

Traditional Kanban-based pull systems may yield huge benefits in manufacturing with repetitive outputs. Also, there have been successful attempts of using pull systems in white-collar settings. However, these white-collar settings have had repetitive outputs and generally low complexity.

Outcome-based control systems are usually used to control settings that are complex. However, process-based control systems may be used if modified to enable the increased flexibility needed. Two main things need to be considered to do these modifications; the nature of the knowledge involved in the process (codifiability, completeness and diversity) and the trade-off between discretion and standardization in white-collar work. According to some authors, “pure” pull systems are not even feasible when complexity is too high. Others, like Locher (2011), is of the belief that this complexity is due to the lack of standardization. Thus, he thinks that this argument presents a logical paradox. However, it is important to relax the assumption of order, in other words, to accept that some things have a certain amount of complexity. Even if things may be standardized and complexity is decreased, there is always a “cost” associated with doing so. In some cases the same or higher level of performance may be achieved at a lower cost by accepting some complexity and to choose strategy accordingly.

4. Empirical Findings

In this chapter, the context is described in greater detail based on data collected in interviews and discussions with staff at VCCS and VCC. First, the organization and dynamics of the QB process is described in greater detail. Then the QB process is put into perspective by describing its place within the main VCCS value stream, the so-called meta flow. After this follows a description of how the context fulfills or does not fulfill the preconditions of pull, as described in the theoretical framework. The chapter ends with key take-outs from interviews held with external experts, used to support the analysis and discussion.

4.1 VCCS

As introduced in the methodology chapter, Volvo Cars Customer Service (VCCS) was the context in which the emphasis of the empirical studies took place. The particular process of focus has been the QB (Quality Bulletin) process which deals with quality issues from the market or raised internally. The process will be presented in principle, highlighting the different interactions and ownerships of a quality case. A major part of the investigation is focused on the deliveries between the process steps, the standardization and the control system since these are related to the preconditions of pull.

4.1.1 Organization and Dynamics of the QB Process

The QB process is made up of a series of steps (see fig 16) that involve several different parties, amongst them, VCCS. The description of the QB process below is deliberately simplified to keep out some details, in part because of the sensitive nature of the work performed in the process and in part because there is no real need for greater detail. The issues handled in the QB process are sorted into different categories depending on the nature of the issue and what sort of action that is needed:

- QBA: Active. Not safety and/or non-compliance issue but still active action
- QBL: Limited Action. The issue (affected cars) has not reached the customer yet
- QBR: Recall. Safety and/or non-compliance issues. The affected cars are recalled from the field.
- QBS: Service Campaign. Solution is put in place at the next upcoming service.
- QBP: Policy. The customer gets extended warranty on the affected components.

Many of the QB process steps are gates that act as filters for deciding which quality issues are deemed critical (see figure 16). Input to the process comes from several different places; warranty claims made by customers, internal audits and tests etc., and authorities. The raw data input is filtered for reoccurring issues that might be critical to health and safety, law and policy compliance or customer satisfaction. The issues that are deemed potentially critical (around 40% of the total input) are presented in so-called pre meetings. These meetings are held at Current Model Quality (CMQ), the Product Compliance Office (PCO) and Volvo Cars Manufacturing (VCM) depending on what input that is being processed. At the pre meetings, once again, the data is filtered. Around this time, root-cause analyses are performed and the scope is defined (concerned models/markets etc.) to be able to correctly assess the issue. The issues that are deemed critical at the pre meetings (around 50%) are collected and presented in a forum of senior personnel called Critical Concern Management Team (CCMT). At this point, both the root-cause of the issue and a potential solution is presented and the decision to be made is whether to prepare the solution for market introduction or not. When an issue has come this far, it is uncommon that they are not deemed critical at the CCMT (around 5% is dismissed).

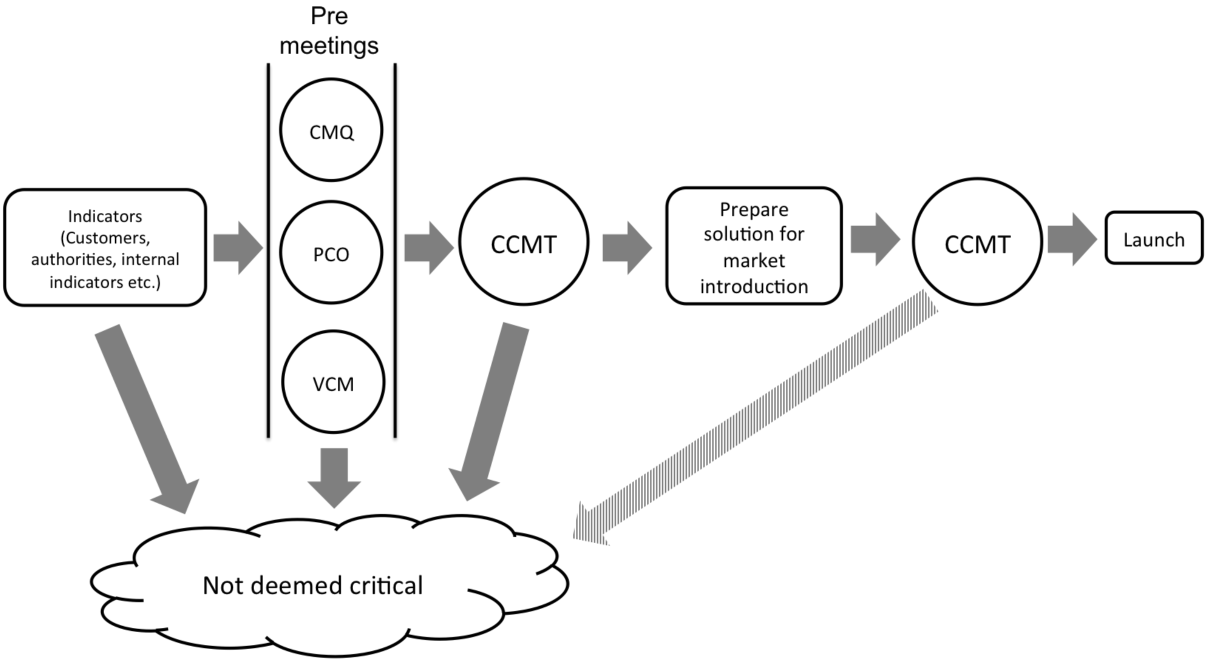

Figure 16. Simplified visualization of the QB process

After the CCMT review, the chosen solution needs to be prepared for market introduction. These preparations involve different activities depending on what kind of solution that is to be implemented; in other words, the routing of the flow differs depending on what is needed. For example, a simple software update typically involves fewer steps than changing physical components.

A major coordinator in the preparation process is the Assignment Leader (AL) team. They use resources normally dedicated to market preparation of new products to prepare both physical and non-physical part of the solutions. The non-physical activities include creating a part structure, determining and describing methods for service operators, determining what time it will take change the part/implement the solution, and to prepare information such as updated user manuals etc. The physical activities include ensuring supply of parts, packaging, warehousing and distribution. When the market preparations are completed, a final CCMT review is held. It is very rare that items are dismissed at this stage but changes to the solution or market preparation might be suggested. The items that pass this point are launched to the market.

When discussing the QB process and its filtering stages it should be mentioned that issues that are filtered out are not necessarily forgotten. Alongside the QB process there is a lessons learned flow with the purpose of making sure the indicator input not deemed critical to customer satisfaction, compliance or health and safety, is used to make better products. After all, this is often valuable customer input.

4.1.1.1 Roles and Responsibilities of the QB Process

The QB process involves many different functions within VCC and it is not absolutely clear who owns what part of the process and where responsibilities start and end. There are many different people and departments involved in investigating the problem, developing the solution and preparing the solution for market introduction and therefore, involvement seems rather ad hoc; people are involved when their knowledge of the issue is required. There are, however, some more prominent roles at the various stages, see figure 17. As the figure suggests, at any given point in the process, there might be several roles involved simultaneously. Therefore, it is difficult to define clear handovers in the process. Instead, work is often done in parallel and different roles are, as mentioned before, involved when needed.

Along the time axis in figure 17 there are estimations of how much time the different process steps might take. Even if this estimation varies from 8 weeks to 29 weeks, it has become clear that in reality, the process time can vary even more.

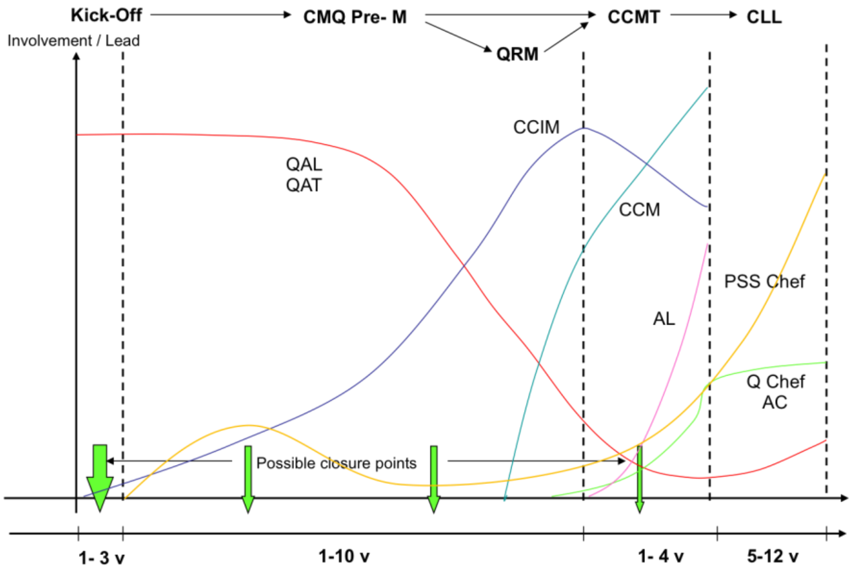

Figure 17. VCCs own interpretation of the involvement of different roles: QAT (Quality Action Team, part of CMQ), CCIM (Critical Concern Investigation Manager, part of CMQ), CCM (Critical Concern Manager, part of PCO), AL (Assignment Leader, part of VCCS), PSS (responsible for design of specific areas/components of a car, part of R&D), Q Chef AC (Quality Manager)

4.1.1.2 Complexity of the QB Process

Complexity is very high (low or non-existing predictability) in the early stages of the QB process (see figure 18). The parties that represent the interface towards the indicators, especially the external sources, have literally no idea whether something critical will appear the next day. One of the few times that they can expect increases in workload is when new products are introduced since this is usually coupled with frequent (but usually simpler) initial concerns. Complexity decreases as the data is filtered and investigations into the root-cause of issues are carried out. Complexity is still quite high when the root-cause have been determined and the scope have been set since there is usually still no solution to the issue. As the solution is being developed, complexity decreases. When a solution is formed and has been approved, complexity is considerably lower than initially and the process enters a more production-like stage. Although complexity is lower, the preparation process is subject to both routing and output variation. Depending on what kind of solution and bulletin class (e.g. QBR or QBL) that is being prepared, the preparation process may vary from days to several weeks. Extensive experience and gut feeling seems to be the dominant tools for predicting lead times, resource usage and potential criticality throughout the QB process.

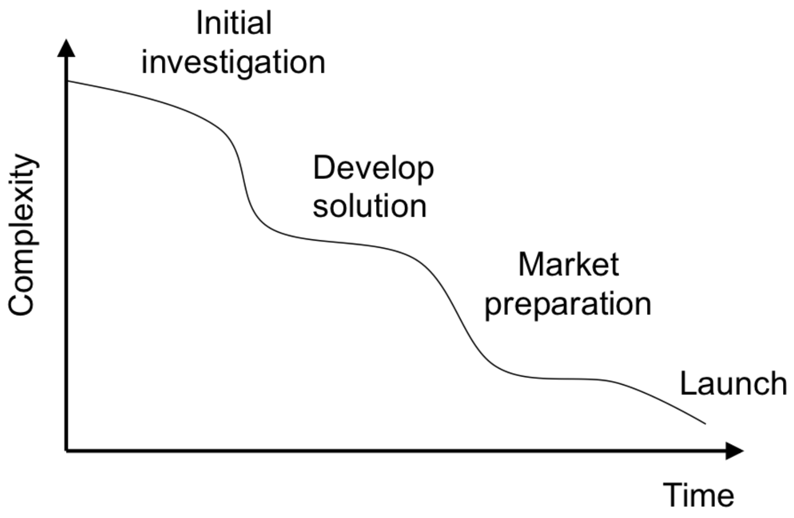

Figure 18. Principle figure describing how complexity decreases as work progresses

Approximately 95 % of the inflow to CCMT is derived from CMQ, why a comparison between these two process steps is representative for the QB process. Measuring the amount of items entering CMQ and CCMT over time gives a clear image of how both mean volume and deviations from the mean, i.e. volume variability, is considerably lower at CCMT than at CMQ (see figure 19). It should be noted that the data visualized in figure 19 was collected at the same point in time for both CMQ and CCMT and since there is a highly variable delay between these two stages a single data point does not represent the same items. In other words, a week with zero items entering CMQ does not imply a week with zero items entering CCMT.

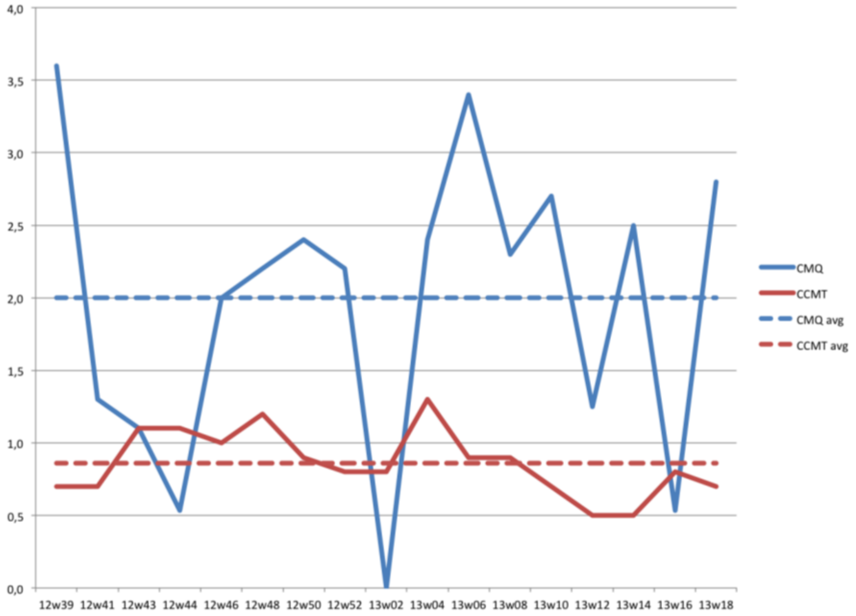

Figure 19. Normalized amount of items entering CMQ and CCMT between 2012w39 and 2013w18. Measured every two weeks

4.1.2 QB within the VCCS Meta Flow

As mentioned in chapter 2, VCCS is making efforts to better understand internal value streams. This work has resulted in insight into how unnecessarily complicated the current state actually is. To cope with this, VCCS developed a meta map describing a future state (see figure 20). The map is a guide when mapping value streams and decreases the risks of getting lost.

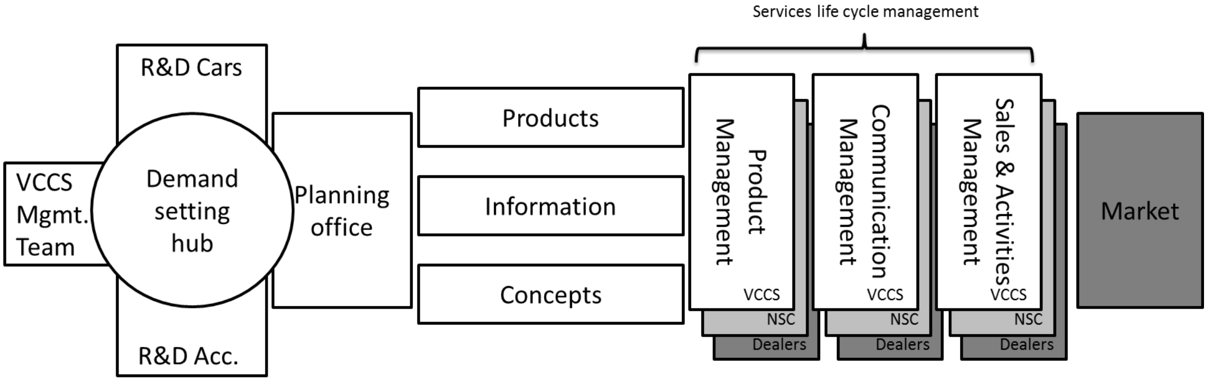

Figure 20. A slightly simplified version of the VCCS meta map

The map highlight different parts of the value stream, mostly not categorized by traditional functions but by actual activity. Within each of the map's sections there might be several departments or functions involved. What VCCS calls services life cycle management (product, communication and sales & activities management) is visualized in three dimensions: VCCS, National Sales Companies (NSC) and dealers. Visualizing these three dimensions aids in understanding that the value stream also flows in this direction. The sections called "Products", "Information" and "Concepts" represent what VCCS likes to think of as a factory. "Products" represents the handling of physical products: packaging, warehousing, distribution etc. "Information" represents the preparation of information related to aftermarket: breaking parts down to spare part structures, determining and explaining methods for changing the spare part, setting the time of executing the method and preparing information to operators and customers (see figure 21). Within each part of the information section there are specialized cells divided into different parts of the car (see figure 22). Delving deeper into each specialized cell there is even further specialization. These are the resources used by the AL team when preparing a solution to a quality issue for market introduction. However, new product and accessories development is the major input to this part of VCCS. "Concepts" represents the development of concepts being sold to dealers in order for them to increase efficiency and customer satisfaction. The "Products" and "Concepts" sections are not the focus of this thesis.

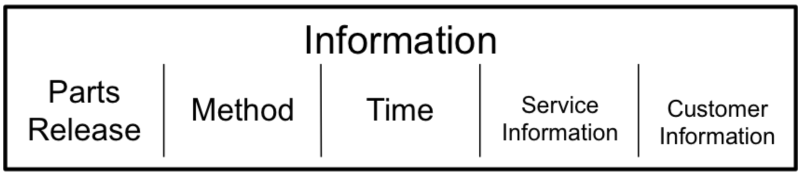

Figure 21. The main steps involved in the "Information" section

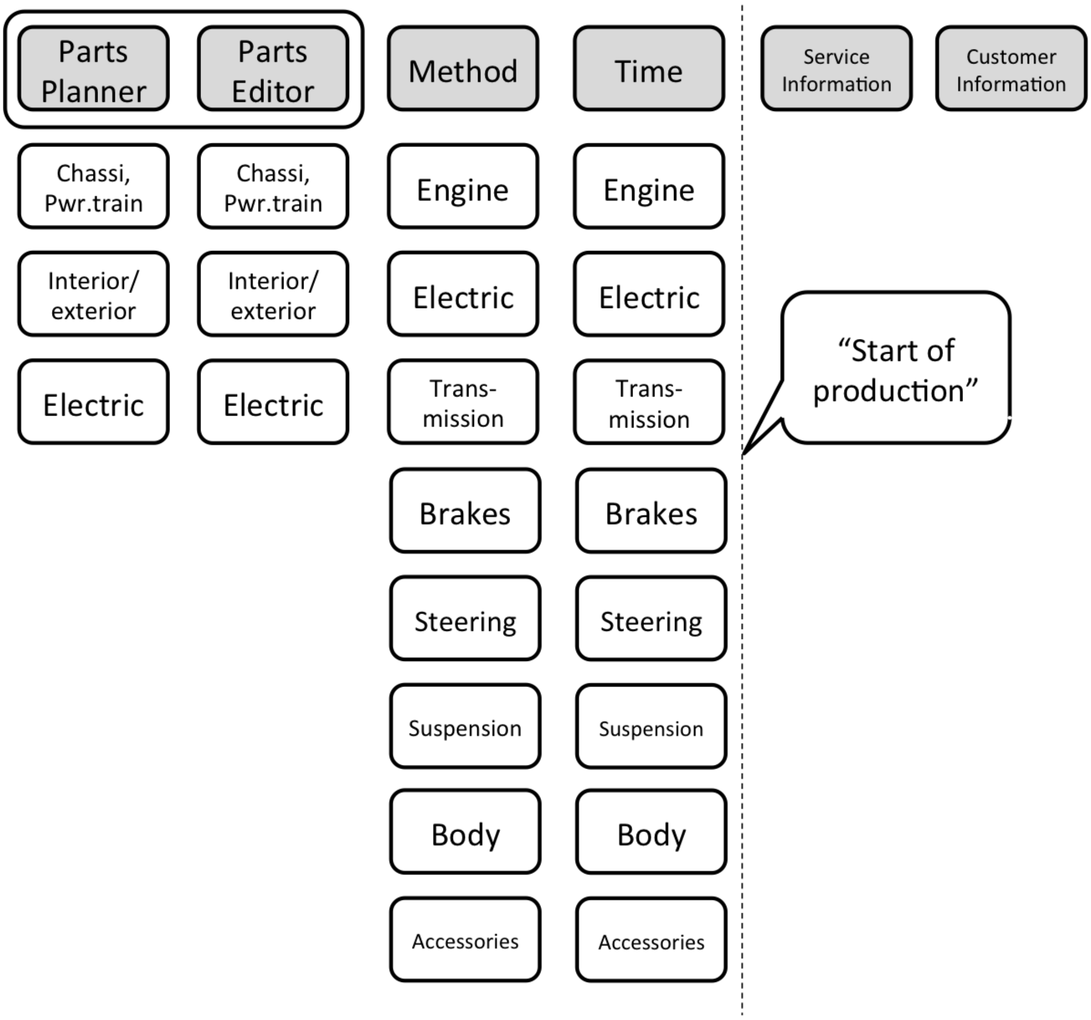

Figure 22. Detailed view of the ”Information” section

The major part of the QB process is situated in the “Product Management” section. Here, CMQ, PCO, R&D and AL is tending to the product life cycle by improving quality issues. Issues that are not solved at the dealer level are sent to the NSC level. VCC will solve the issue if the NSC cannot. Once a solution has been developed the AL team will prepare the solution for market introduction. Today, this is done by grabbing resources from the “Information” and “Products” sections with additional tasks. By doing so, the normal flow of preparing new cars and accessories development is disturbed. To deal with this, and to generally improve the flow of tasks, VCCS have envisioned a hub of sorts (see figure 20) with the purpose of guiding and controlling process input.

4.1.3 Fulfilling of Pull Preconditions

Having explained the QB process more thoroughly the next step is to dig deeper into details to understand the critical elements for incorporating a pull-based control system. Each element will be assessed in the following chapters.

4.1.3.1 Defined Agreements

Today there are no clearly defined agreements between the process steps being executed in the QB process. To start with, there are no surveillance on the storage points between the process steps, rather, cases are just being passed downstream as they arise from either the market or internally. A QB process manager from PCO claims that there are really no choice when issues arise from the market, the cases must be pushed through the process in order to solve it as fast as possible to maintain customer satisfaction. The consequence is that cases pile up in big buffers in front of downstream process steps when there is a big demand from the market (many quality issues in the market). Furthermore, there may be quality issues that are more critical than others to customer satisfaction, and possibly safety. These cases are given the highest priority and the perception is that they must be pushed through the process, implying that downstream process steps may have to put the present work aside. Obviously, this leads to an uncontrolled volume of cases in the process. There are no defined limits pertaining to the volume in the process. Instead, all cases needed to be solved find their way through the system by upstream personnel trying to plan and even out the burden on the downstream process steps. Referring back to queuing theory it means that the waiting time is increasing for the system. This is confirmed by an AL team manager who claims that when there are too many cases to handle, the throughput time gets higher for each case. The AL team manager compares to times when demand is low and time is allowed for detailed planning. During these periods, the throughput time is considerably shorter for comparable cases. When PCO delivers cases to AL it is done through the interaction of two system. PCO uses Lotus but the cases arrives into a system called SPIE, which will be further explained in chapter 4.1.3.3.

When a case arrives to AL and the market preparation commences, there is one assigned AL member that is responsible for the case and all possible interactions needed before delivering a solution to TIE, see figure 23 for a visualization of the AL process. This person requests work from different functions, such as Method and Time. The workload on AL is thus passed on to the other functions of VCCS. The main tool for interaction and deliveries within the AL process is SPIE, which will be further explained in chapter 4.1.3.3.

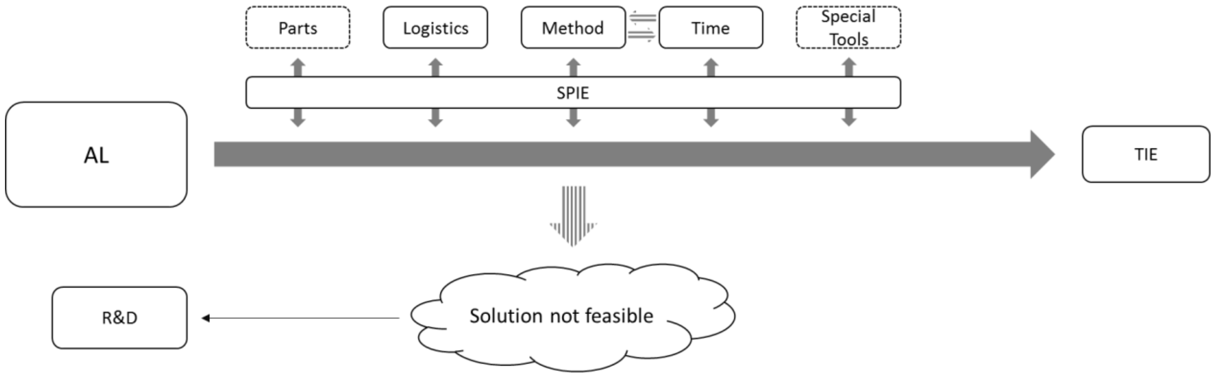

Figure 23. The AL process and the most common interactions during the market preparation. A finalized solution is delivered to TIE which is the channel to the market and solutions not feasible are sent to R&D for redesign

Looking at the product mix through the process, there is no defined agreement either. This, however, is expected since all cases are unique in a way and may require a variety of different competences. Thus, each case may require a different way of solving the problem which might alter the needed interactions with other competences. This was claimed from a PCO manager working in CCMT. The situation is unlike a production setting where there normally is a predefined product offering requiring the same interactions for every produced product. Even though the cases in themselves are unique, the AL team manager claims that methods and interactions are almost always the same, both for the complete QB process and within the AL process.

The same reasoning applies to the sequence of model mixes between two process steps. The sequence of case mix is random and one case is not the other alike, why there cannot be a defined agreement on the type of cases to be delivered on specific times.

The lack of defined agreement is also clear when it comes to deadlines for handling the cases. PCO is the owner of a case throughout the whole QB process, even when AL starts working on the preparation for market launch. For AL to plan the work efficiently, due dates for the cases is needed. However, the experience is that there are no clearly defined deadlines, rather they vary during the processing time which leads to uncertainty at the AL team. Furthermore, too many cases arrive to AL with the need to be completed immediately.

There is no clearly defined agreement on the number of cases that each member in AL should have, nor how big a workload each member should handle. Yet, there is an experience-based suitable workload internally decided within AL of around five cases per person. This number is though fluctuating and can differ depending on the scope of the cases. The reason for having more than one case is that there is always built in waiting time for each case, since there are other functions that works with the cases too. This internal arrangement is though not communicated to the rest of the QB process.

Cases that AL is preparing for launch to the market has reduced complexity, as the majority of investigations and problem solving has been done. Therefore only a few cases are rejected during the AL process, around 2-3 % according to the AL team manager's experience.

4.1.3.2 Dedication between Parties

Since the process is a white-collar one, the storage of WIP is primarily electronic documents which are stored in different databases for the complete QB process. Cases are first stored in a system called Lotus that is used by CMQ and PCO. However, there are two separate databases for CMQ pre-meeting and for PCO which is used in CCMT. When the case has been prepared in CMQ pre-meeting and a decision has been taken to forward it to CCMT, information must again be typed in manually in the database dedicated for CCMT. When a decision on market preparation is taken at CCMT the case information is transferred to the database SPIE, which is used by AL and its interacting functions, refer to figure 23. As cases are delivered to SPIE from Lotus, they are deleted from Lotus. If the AL team manager consider the input to be insufficient and thus needs to be complemented, PCO must revise their work and then refill all information manually before it can be sent to SPIE again. Today, there is no organized follow-up on delivery performance from CCMT to AL. A team manager at AL though claims that rework must take place the majority of times.

Resources, in terms of personnel, are not specifically dedicated between two process steps. This is partly due to the fact that workers are not only working on one case at a time, leading to a need for different kind of interactions and handovers in the process. This is especially true in the early phases of the QB process when the investigation and solving phases take place. The dedication between AL and the QB process (primarily PCO) is not perfect either, resources at AL are disturbed by other issues that arrives to them. An AL manager believes it is due to the fact that AL members generally have a broad network within Volvo Cars from previous engagements. Thus, there is noise into the process disrupting the regular work.

The lack of clear dedication is also an effect from the lack of consistent routing in the QB process. As mentioned in the last paragraph, the interactions and handovers needed gets clearer the further the investigation goes. Changes to the routing may still happen during the development of the solution as well since you do not know what problem you will meet, a CMQ pre-meeting manager claims. Thus, within the different functions of the QB process there seems to be quite complex interactions initially. More detailed mapping would be needed to understand these interactions thoroughly. The routing is though getting less complex further downstream. When a case has reached AL the working process is fairly straightforward. There are still other functions that may be included, for example Parts if the solution to the market includes a new part to the car. The Parts department make a breakdown of the parts structure for Method to know what parts that need a new method for repairing. All the other possible interacting functions, such as special tools or Chemistry, are known after the method has been set. After this point, the way forward to market release is completely known.

A common reference time, or takt time, should be defined between process steps to strengthen the dedication. This is not the situation for the QB process, rather cases are expected to be solved in the fastest possible manner as was explained before. Since there are no standard takt time to adhere to, the control becomes very complex. Historical data shows that AL handles around 80 cases per year. For Q1 and Q2 in 2013 there will be a total of 24 cases, meaning that the total number most likely will be lower this year. The AL team manager claims that it is due to the fact that there has not been any major market introduction lately, the filtering of cases has become more efficient why fewer cases are released to AL and also since the capacity to handle cases are deficient. Within the AL team, there is another main responsibility which is to handle technical journals. These tasks are less important and can be done when demand is not at its peak.

4.1.3.3 Visual Control

The absence of visual control is significant in the QB process, just as for the typical white-collar process the work is being sent electronically and there is no visual method for illustrating inventory levels. The AL team is about to introduce visual planning, as part of VCCS's Lean initiative, focused on showing what each person is working on for the moment and what the deliverables for the task are. The visual planning should also show the status of the tasks, e.g. to illustrate for everybody if assistance is needed. Still, there is no visual method for showing the inventory levels for each process step, rather the cases in queue are stored in databases. Furthermore, the visual planning tool suggested is not easily seen by all workers, if one worker indicates a possible problem using a post-it note it might not be seen by the other workers until they pass by the story board and are observant. Thus, there is no immediate feedback to workers from the visual planning tool, figure 24 illustrates a typical story board at VCCS. There is also no visual control for monitoring the performance of the workers, this is essential to avoid hidden performance deficiencies. The time it took to perform a task is noted on a post-it note. The post-it note for the completed task should then be stored in a personal book, but it is still not visually clear how the performance changes. Furthermore, there is no automatic stoppage and correction of the process if the cycle time is longer than the estimated time. This is partly due to the fact that the estimated time is just an estimate, and not a clearly specified standard for the cycle time. Furthermore, there is no visual tool or sheet for the workers to validate that their work is being done good enough according to standards. Thus, there is once again a lack of immediate feedback.

To conclude the findings on visual control, today it is more directed towards showing status of the work in order to create a discussion and to help aligning tasks towards the overall targets of the team, department or company. Also, the visual control is about clearly seeing what the team members are working on.

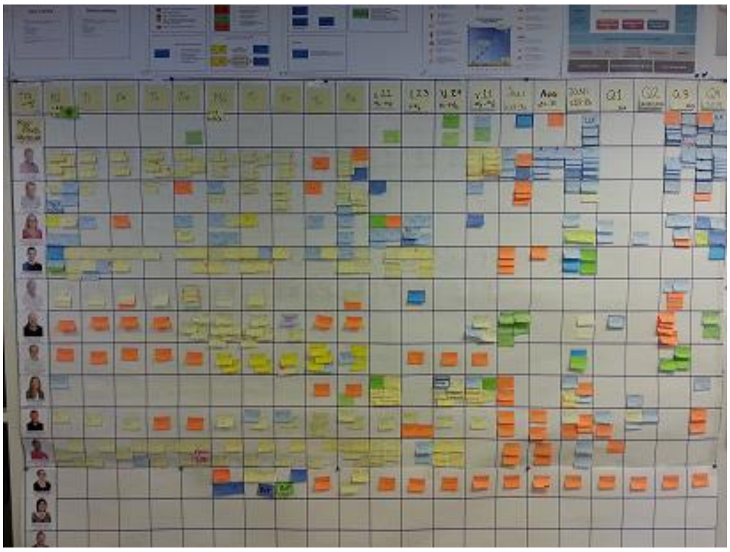

Figure 24. The visual planning tool at VCCS, with daily planning for the upcoming two weeks and weekly planning for the rest of the year

The means for sending cases is the use of the systems Lotus and SPIE, as was described in chapter 4.1.3.1. When for example PCO passes cases to AL and into SPIE, there is an electronic signal via email but no visual signal to the AL team managers. As the AL team manager start looking at the case, there is no automatic signal back to PCO saying that it has been received or planned for. The signal from AL that they will start working on the case is done manually through email and is confirmed through a start-up meeting together with concerned parties. As the work within AL commences the main interaction is the use of SPIE where boxes are ticked to indicate whom must work on a case. An AL team member facilitates this and assures that the concerned functions are working on the cases. The trigger between the functions is SPIE, but nothing is visualized for the workers.

4.1.3.4 Leveling of the Workload

The work being done by the AL team cannot be considered smooth. Work arrives to the team when quality cases are considered as important to be solved by CCMT, which can take place at any time. The consequence is that volume, task mix and sequence of working on the tasks is completely uncontrolled, as was mentioned in chapter 4.1.3.1. The lack of control leads to temporary overburden on workers during peak periods, and occasionally calm periods when demand is low. The difficulties in smoothing is expected considering the environment as the solution and launch of a quality case is made on demand, thus, there cannot be a finished goods storage. Utilizing the finished goods storage is a common approach in smoothening the production on a level that can still deal with peak demand. That approach does not apply to this environment as it is more of a MTO production, where products are made after the orders are placed. Moreover, the variation in the nature of the tasks also contribute to difficulties in smoothening. The consequence is that it is difficult to level one “production line”, since there is really no fixed production line due to varying routing.

4.1.3.5 Level of Standardization

Referring to the important elements of standardization presented earlier, it is insignificant in the QB process. Firstly, there is no defined takt time for the cases through the process, and there are no defined cycle times for the process steps. Rather, tasks related to a case are expected to be executed as fast as possible. The exception is AL where there are defined cycle times for the different process steps, these are based on the experience from senior personnel. Secondly, there is no firmly incorporated operations routine. The routing is complex in the early phases, while it gets more straightforward at AL, what is also critical is the need for iteration. Since it is not a manufacturing of the same products each time, there is a need for iteration between process steps from time to time which makes the system more dynamic and complex. Regarding the standard operation for doing the work (procedures), there is no standard sheet or similar tool to assist the worker in how to execute the job. There are checklists of what things that should be done before a delivery can take place, but now how. The exception is within the AL process were functions such as Method and Time have standardized methods for how to execute the work. Thirdly, there is no explicitly defined standard quantity of WIP for the different workers. Rather the workload on the workers is decided through “gut-feeling” and experience, as a PCO manager expresses it. The ad hoc planning of capacity leads to unbalance and uneven inventory between workers. Given the process dynamics, the consequence is that time to completion is unpredictable and varies a lot. There is a standard sheet for what information should be delivered to AL in order to start the market preparation. The experience from the AL team manager is though that the information is insufficient and not filled in correctly most of the times.

The experience from the AL team leader is that the effective time spent on a case is about 20 hours, but a whole process takes around 40 hours due to waiting. However, the process takes around 8 weeks due to parallel projects. The hours for a case are diluted on that time.

4.2 External Expert Interviews

Interviews have been conducted continuously. The interviews were rewarding and gave many important insights and allowed the validity of concepts and conclusions to be verified. As these interviews were of the unstructured kind the main insights gained is grouped and shortly summarized below. These and other insights have been evaluated and consolidated, and are incorporated continuously. A list of contributing experts is found in chapter 2.4.2.2.

4.2.1 Summary of Key Take-Outs from Expert Interviews

Several experts recommended caution when trying to use methods and tools designed for production for systems without the corresponding characteristics. In other words, not to force a production-like system onto a setting which is not production-like as it might hinder actual value-adding activities.

Variability was put forward as the biggest difference between white-collar value streams and conventional manufacturing. Variability is causing the processes to be more complex. When variability is high, i.e. a complex setting, pull becomes problematic since controlling in an upstream direction requires predictability. In the cases when complexity is high, tools such as PDCA or LAMDA cycles are more suitable than firm process control. Conventional, top-down, task and process standardization becomes very difficult (and even counterproductive) when complexity is high. Instead, a viable way of standardizing is bottom-up. In other words, letting the people performing the actual task standardize at the lowest level possible and then work on integrating these “micro-standards” using rules for interfaces. One example of bottom-up standardization is the alphabet: a set of around 26 standardized parts accompanied with a set of (grammatical) rules and allowed combinations (words). These 26 standardized parts can create a practically infinite variety of information. If, instead, standardization had taken place at sentence (or sequence of sentences) level, the flexibility of the system would be drastically reduced. Several experts agree that there should be some trigger point along the QB process where variability is low enough for predictions to be made with high probability. It is possible to visualize and communicate predictions before this point but the probabilities associated with these predictions should be clearly stated.

“Pure” pull does not need to be the only way to achieve a low waste, well-coordinated value stream. Push and pull/push hybrid systems needs to be taken into consideration as well. Achieving a demand-controlled flow (pull) might be done by setting prioritizations according to ultimate customer demand. This can be considered a high-level pull system that is not interfering with operational control.

Visualization becomes even more important when the value stream at hand has high variability. However, the visualization itself is not the biggest challenge to achieve. Therefore, making the value stream more visual might be seen as a high reward, low effort deal. To achieve pull, predictability is key. Hence, upstream activities and status needs to be visualized so that downstream parties may increase predictability.

Leveling is key to create continuity. Leveling product volume is not the biggest challenge, it can be done in three basic ways: inventories (especially finished goods), long-term forecast planning combined with daily prioritization and routing decisions, and through deliberate excess capacity. However, product mix is more difficult to level. Furthermore, it is crucial to have a determined takt time in order to know the status of the process. In other words, in order to realize how the pull-based process is performing, a reference time is needed.

The customer must be prioritized at all times and what the customer demands should be weighed heavily when the characteristics of a system are set. In other words, it is key to realize what is actually adding business value. Also, it is important to know what actually flows. In other words, what is actually delivered to the customer? If it is information that is being produced, it is important to realize that identical information, unlike physical goods, can be at several locations at the same time. This may open up to new possibilities for how to coordinate a value stream. Focus should be put on interfaces, i.e. handovers and handshakes between different parties and how this could be made more efficiently. Interfaces will become more troublesome the longer the physical and psychological distance between the involved parties.

At SAAB Automobile, a process similar to the QB process utilized cross-functional teams responsible for the whole process, basically owning the current item from customer input to market implementation. This enabled a quick and rich communication between roles, basically eliminating handovers. The number of items corresponding to each team was limited according to empirical data (continuously updated) and experience. The mean lead time for items proved to be 16 weeks. This mean value was used to determine capacity. However, capacity utilization was controlled through prioritization such that no more than 80% of full capacity was utilized due to queuing behaviors above 80% utilization (see chapter 3.1.4). There was also a “separate” flow for items deemed particularly critical. Even though a typical item required around 16 weeks (with substantial deviation) before implementation, a target was set to inform the NSC within 24 hours. This information could rarely be about a suggested solution. Instead, it usually was about informing the NSC about the plan moving forward towards a solution. Every team was attending a common weekly meeting to set and/or change priorities. A fundamental prioritization tool called “Profits” was implemented. Prioritizations were based on overall targets: first customer demand (i.e. criticality) and second, cost for the company. If prioritizations are constructed from targets concerning amount of accomplished items per time period, this might prove fatal for total process effectiveness since real customer need might be disregarded. Visual planning was utilized in order to have a controlled status for each item.

5. Analysis

In this chapter, the theoretical and empirical data is brought together. A simple model describing the relationship between pull, its preconditions and related principles is presented. Then, the methods associated with pull is contrasted to complexity and different parts of the QB process is discussed with regards to relative complexity. Complexity is then broken down to form a model for evaluating different pull systems and their suitability for the QB process.

As described in chapter 4, there are three basic preconditions of pull: defined agreements, dedicated items and control. To enable these preconditions it is common to utilize the principles of leveling, standardization and visual control. In other words: pull is enabled by the three preconditions and the three preconditions are primarily enabled by the principles of leveling, standardization and visual control. This relationship is visualized in figure 25.

5.1 Complexity and the Principles of Pull

To utilize the principles, a certain degree of predictability is required, in other words, complexity is hampering the utilization of the principles (see figure 25). This applies especially to the ability to standardize and level effectively. As for visual control, it might be more challenging to design a visual control system for complex situations, but there are no obvious fundamental reasons for why visualization might not work. Some of the most obvious reasons for why complexity hampers standardization and leveling will be discussed below.

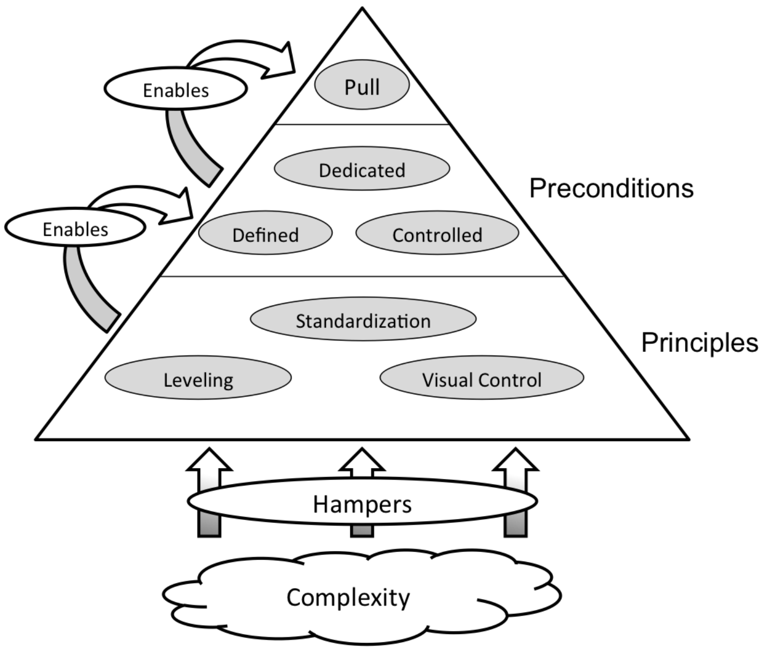

Figure 25. Relationship between pull, its preconditions, enabling principles and complexity

5.1.1 Complexity and Standardization

It is usually possible to standardize processes at a certain level. After all, at some level, all processes could be considered repetitive. The important issue to consider is whether it is desirable to do so or not. When the situation is complex, patterns may be recognized in retrospect and they might even repeat themselves, however, there is no guarantee that they will. Therefore, standardizing with respect to those patterns might even be harmful.

There are three main elements that need standardization: the takt time, the sequence of doing things, and the amount of inventory (for each operator) required to accomplish the standardized work. To standardize takt time, demand needs to be reasonably predictable. Otherwise, the takt time will be set using mean demand over a certain period of time, which might vary significantly from actual, real time demand resulting in either substantial under- or overcapacity. Also, to standardize takt time, the internal process should be stable and predictable for the takt time to be efficiently incorporated. To standardize a sequence of doing things require a known and limited product mix. Otherwise, the routing and specific operations required by any unique case will result in a unique sequence. To standardize the amount of inventory necessary for each operator requires the output to be fairly repetitive. In situations such as MTO or Engineered to Order (ETO) the output does not exist until an order has been made. Also, there is no guarantee that a similar output will be produced again. Therefore, when output is non-repetitive, it is difficult to think in terms of standard component inventory in the classical sense.

5.1.1.1 Possible Ways to Standardize in Complex Situations

As mentioned in chapter 4.2.1, there are alternative ways of standardizing. Instead of standardizing top-down, the process may be broken down into its smallest constituents and be standardized. These standardized elements needs to be brought together, by focusing on rules and interfaces for combining the elements, to form the standardized process in its entirety.

5.1.2 Complexity and Leveling

A common way to enable leveling of the workload is to introduce inventories, especially finished goods inventories. Simply put, it is a price worth paying to have smoother production. However, when output is non-repetitive and predictability is low it might not be feasible with an actual finished goods storage. Again, think of ETO products: they do not exist until specifications are set and an order has been made. The whole idea with introducing inventories to enable smoothing is to cope with variability. The more unpredictable the demand, the higher inventory level is required in order to achieve smoothed production. If demand is highly unpredictable and output is custom, leveling by finished goods inventory is not plausible.

An essential concept of leveling is takt time which is a way to deal with variable ordering, thus producing a defined product mix according to takt time and not acting on customer orders. This strategy is though not feasible in a process like QB since there are no raw material warehouse to produce from, as long as there are no quality issues, there is no work to be done.

5.1.2.1 Possible Ways to Enable Leveling in Complex Situations

If finished goods inventory is not an option there are alternative ways of enabling leveling. One is to deliberately dimension a process for overcapacity. A common rule of thumb is to have 20% overcapacity. It is, however, difficult to dimension capacity in the classical sense in complex white-collar situations since value-adding time might not always lead to proportional results. For instance, a designer that has 10% value-adding time and 90% “waste” might be more productive than a designer with 100% value-adding time. It is not the amount of hours spent on complex tasks that are important, but the impact and result of the work done. In order to talk about dimensioning capacity in the classical sense, there needs to be a high degree of standardization. Another way of achieving smooth production is to combine long-term and short-term planning as discussed in chapter 4.2.1. This will require predictability at some level to enable long-term planning of capabilities. Takt time as part of smoothing is not applicable in the conventional way, instead another type of reference should be set. It is possible to set a takt time for each case, i.e. a deadline, clarifying time to completion. It can be used as reference in order to identify deviations and to learn and develop from the insights.

When the situation is complex it seems a “pure” pull system, a basically autonomous system that demand controlling all movements throughout the whole process down to individual operator level, is not plausible. That said, not all white-collar processes are complex.

5.2 Different Levels of Complexity throughout the QB process

The earlier stages of the QB process are generally more complex than the later stages since they are of investigative and developing nature. Investigations made by CMQ and the development of solutions by R&D generally belongs to the complex domain of the Cynefin framework (see figure 26). It is possible to make efforts to move these activities into the knowable and known domains. However, when activities are of investigative or development nature, variation can be a key ingredient to developing successful solutions. In other words, there is a risk of losing initial width of possible solutions and, by that, decrease the ability to find the most suitable solution for a specific case.

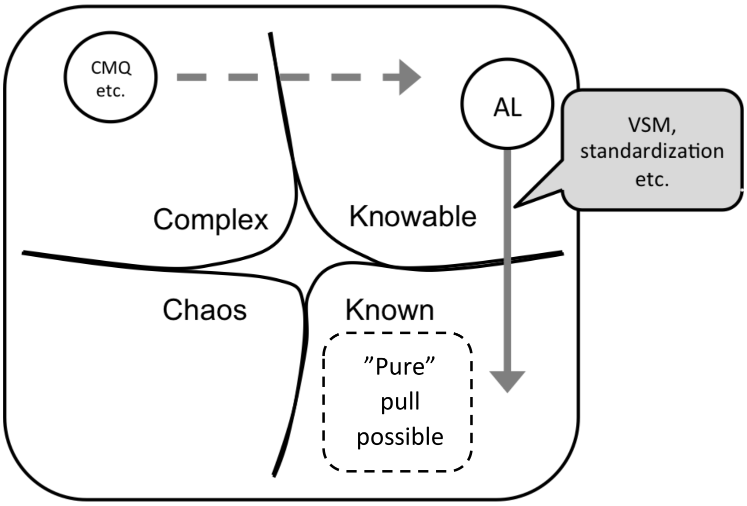

Figure 26. Different levels of complexity within the QB process

By conducting investigations and developing solutions, the continued process becomes more and more predictable. Once a solution is developed and decided upon, complexity is considerably lower. This is when the preparation for market introduction is started. The preparations coordinated by AL therefore generally belong to the known or knowable domains. Here, decreasing complexity by introducing standards, performing VSM, visual planning and control etc. is more efficient. Once cause and effect relationships are found and understood, they can be trusted to hold true moving forward.

5.3 Complexity and Variability

From theory and expert interviews it is clear that much of the complexity often associated with white-collar work is not so much about the output being non-physical, and therefore less visible, but about high levels of variability. To make the concepts of complexity and variability more manageable and useful it is divided into internal and external variability:

- Internal variability is basically a measure of internal ability, in other words, internal process variability. Think of it as the ability of the process to always produce identical output at the same quality and time when all input and external factors are static. Process variability may be lowered by introducing efforts such as standardization.

- External variability is divided into two sub-categories: Volume variability and product mix. It is difficult to directly influence external variability, instead, it may come down to actively choosing not to provide certain output or accepting that some customers will have to wait. In some contexts, like many of the items in the QB process, this is not a viable option.

- Volume variability is the variability of external demand with regards to volume
- Product mix is the variability of external demand with regards to type of output demanded

It is important to understand the difference between these three types of variability. If process variability is high, output may vary in terms of quality and timing even though both volume variability and product mix may be low. High volume variability offers different challenges as supposed to high product mix. You may, for instance, produce only one kind of output, with an extremely stable process and still have heavily fluctuating volume demand. This will result in difficulties in capacity planning and smoothing (may need large inventories to cope with the volume variability). On the other hand, if volume variability is very low and the internal process is stable but the product mix is high you face different challenges. If product mix is high, the process will need to be flexible and process routing will vary. Specialization is therefore hampering flexibility and standardization becomes more difficult.

Each of these three types of variability may cause the predictability of the situation to be limited. The situation becomes increasingly complex if more than one type of variability is present at the same time.

5.4 A Model for Evaluating Different Pull Systems

The different pull systems described in chapter 3.2 are prepared to cope with different levels of complexity. The different systems and their suitability was compared to two broad stages of the QB process: QB1 representing the investigative and developing beginning of the QB process (CMQ and R&D), and QB2 representing the preparation work coordinated by the AL team. These stages and the different systems were plotted against the three types of variability: process variability, volume variability and product mix, to determine how suitable they were in different situations. The evaluation and ranking of the pull systems in relation to QB1 and QB2 were executed subjectively based on acquired knowledge from theory and expert interviews.

5.4.1 Criterion for Evaluating the Different Pull Systems

The different pull systems were evaluated along one dimension at the time. The dimension in focus was thought of as “high” and the other dimensions as “low” or “non-existing”, thereby isolating one dimension. For example, when the systems were evaluated along the dimension “Process Variability” a scenario was imagined with high process variability and low or non-existing volume variability and product mix. The systems were ranked on a scale from one (1) to four (4) with one corresponding to low suitability and four corresponding to high suitability. In this case, “suitability” means how well a system is capable of delivering the highest possible customer satisfaction in a predictable manner. Throughput time, amount of WIP and average waiting time for work are among the most common and cited measures of performance for this kind of control system. These measures strongly correlate with delivered quality and customer satisfaction and is therefore the basis for evaluation (Sendil Kumar & Panneerselvam, 2007). Controllability is another important factor since low controllability may result in choked value streams with low quality, long throughput times and waiting time as consequence. If a system was deemed highly suitable for high variability along a dimension it does not necessarily mean that the system is unsuitable for a low variability setting along the same dimension. However, a low score along a dimension does mean that the system is generally ill-suited for high variability settings. In other words, the systems were primarily assessed for suitability in a high variability settings, not how ill-suited they are in low variability settings.

It should be noted that a high score, e.g. four, does not necessarily equal the "best" result since the comparisons are relative. Instead, what is considered the best result depends on each specific context. If the context has high variability, a high score is the most suitable. The two coarse parts of the QB process, QB1 and QB2 (as explained in chapter 5.4) was placed along with one corresponding to low variability and four corresponding to high variability. To clarify, QB1 and QB2 were scored along different scales than the pull systems. This was done in order to find overlaps between the QB process and the different systems.

5.4.2 Positioning the Different Pull Systems and the QB Process

Before the different pull systems and the stages of the QB process could be plotted in the model, each of them were evaluated according to the criterion described in the previous chapter. The justification for each rating is described below and the complete evaluation is presented in table 1.

Table 1. The relative suitability of the pull systems

System	Process Variability	Product Mix	Volume Variability	
Kanban	2	1	2	4 = Most suitable given high variability
CONWIP	1	2	1	
Kanban-CONWIP Hybrid	2	3	2	
POLCA	4	4	4	1 = Least suitable given high variability
QB 1	3	4	4	
QB 2	2	2	2	

5.4.2.1 Positioning With Regards to Process Variability

The lowest ranked system with regards to process variability is the CONWIP system. The rationale is that CONWIP only limits the total WIP for a production line. Thus, the system is sensitive for bottlenecks and for high process variability there is a high risk for accumulating piles of work in front of the bottleneck. As the system is based on standard containers of work in terms of time, it will also be more uncertainty in estimating the processing time for each part. The consequence is that the rest of the operations stop and all WIP is collected in front of a single workstation. Referring back to queuing theory, a drastic increase in waiting time is expected and it will incorporate a high inertia in the system slowing down the process of reaching a steady flow. The raised pressure on the bottleneck will obviously have a negative impact on the throughput time, furthermore it will negatively influence the quality of the work due to excessive pressure on the workers in that process step.

A slightly better system would be the Kanban-CONWIP Hybrid system since it also limits the WIP for each processing step. Accordingly, the negative long-term effect from the bottleneck will be reduced since it can only be as big as the internal inventory limit. When the process performance increases, faster recovery to a steady and efficient flow is possible.

The Kanban system is also given a score of two. The Kanban system will not run efficiently if process variability is high since it is very dependent on a standardized takt time. If, for example, the cycle time for a workstation varies a lot the whole system might be stopped while waiting for that workstation. The only way to maintain activity in the system is to increase the acceptable WIP through increasing the allowable number of cards in the system, which would break against Kanban rule 4. The logic is similar for Kanban-CONWIP hybrid.

The highest ranked system in this scenario is the POLCA system. POLCA consists of cells incorporating more than one process step and competence, or more resources (workstations) for the same processing step, and the interaction is taking place between the cells. Thus, individual variation in processing time is diluted within an autonomous cell and the overall effect should be smaller as WIP can be maintained on a lower level, avoiding excessive queuing and stress on workers. A problem with POLCA is the reliance on standard units of time and capacity, which gets harder to estimate when variation is high, compared to traditional Kanban were physical products are pulled. The issue is of minor importance though, given the other benefits from the system.

In QB1 there is a built in uncertainty in terms of interactions and routing since the investigation and developing stages might include different people depending on what problems are encountered. Thus, the time consumption can be expected to be more variable. In QB2 the possible alternatives for interaction are fewer, and even though there might still be some problem solving and uncertainty it is not to the same extent as for QB1. Thus, the variability is smaller in QB2.

5.4.2.2 Positioning With Regards to Product Mix

The lowest ranked system is the Kanban system. The traditional Kanban uses physical components for controlling the pull system, together with the Kanban cards. If the product mix is very high, it means that there must be storages between each workstation that can store every variant of products or components. This would not be feasible given a very high product mix, unless there are several different production lines with many storage points, which is also unrealistic. Furthermore, if the products were in the form of information it would still not be possible to store the solutions downstream given a high product mix, since each solution might be unique. The argument holds true even for physical products. In Kanban, the production flow and takt time is normally set for a certain production line, but given high product mix there would be a problem of routing. Either it would require highly flexible production equipment or the possibility to adapt the production lines each time, which is not feasible.

A significant improvement to the Kanban system is the introduction of CONWIP or Kanban-CONWIP Hybrid. These systems utilize standard containers of work, measured in time, which means that the variation of parts is less important as long as they has an estimated processing time. These systems also have an element of “push” which is beneficial when product mix is high since unique solutions cannot be stored downstream. However, the systems are not “perfect” since a high variation in product mix still requires a flexible production system due to different routings, which is more critical in a machine setting while it is reasonable to be more flexible in a white-collar production. If the production capability were not flexible enough, the production lines would have to be reorganized constantly which is not practical. The only difference between CONWIP and Kanban-CONWIP hybrid is the risk of excessive choking at the bottleneck. This is likely to happen since the more unique products that arrive, the harder it will be to estimate the time consumption. Compared to only a few products, after a while the learning and experience will lead to more accurate estimates. Therefore, Kanban-CONWIP hybrid is ranked higher than CONWIP.

The POLCA system uses a time plan with standard times for different operations and capacity cards for the pulling sequences. The standard times are rough estimates though, since the cells dilute small errors. As the POLCA system is set up and ready to interact with all possible workstations the possible change in routing due to high product mix does not concern the system.

For a high variability in product mix, first faced by QB1, the routing and interactions might differ for many of the incoming cases. The complexity is also a consequence from the fact that workers in QB1 do not possess the same level of knowledge and experience. If more workers could handle a broader spectrum of cases, the impact would not be as significant. The increased complexity is likely to affect throughput time quite significant. As the cases arrive to QB2, the reasoning is similar as for process variability. Since the cases have been processed, investigating root cause and working out a solution, the complexity has been considerably reduced. Moreover, since the routing is more certain within QB2 the product mix variability is not as significant within that context.

5.4.2.3 Positioning With Regards to Volume Variability

An important assumption for the volume variability is that it does not affect the routing of the product, since it could still be a single product with high volume variability.

Increased volume demand means that a system must work faster, in case of CONWIP (as for the other systems) it means that the number of WIP cards must be increased to allow higher volumes through the system. The alternative is to say no to customers. A risk when trying to increase the workload is that the system gets overloaded, it means that the bottleneck will be critical for the system. In the case of CONWIP it means a big risk for pile-up in front of the bottleneck, thereby stopping all the other machines with no inventories attached to them. The consequence is the same as described in 5.4.2.1, probably leading to decreased quality.

An improvement from the CONWIP system would be Kanban or Kanban-CONWIP Hybrid. For Kanban-CONWIP Hybrid there is still a limit on the WIP for each machine, leading to a more stable stoppage and thereby faster start-up of the system as the bottleneck gets going during low volume periods. The same rationale exists for Kanban where there are limits to WIP for each workstation.

The POLCA can adapt better to volume variability thanks to the paired cells. Firstly, the paired cells take up a bigger proportion of the work, which means that burden is always diluted within the cell. As the cells consist of more than one worker the excessive pressure is not on one worker’s shoulders, probably decreasing the effect and allowing for higher quality delivered to the customer.

The high variability in volume is first faced by QB1. The high variability of cases are first taken care of by QB1, before the critical ones are passed along to QB2. Normally, there are a number of cases per each critical case. For a higher volume into QB1, a bigger part of these must be evaluated before a decision can be made whether it is necessary to continue working on them or not. Thus, the true volume into QB1 solving phase is not known until after some initial processing. When cases arrive to QB2 there has accordingly been some filtering and changes over a time period are not as significant (refer to figure 19). Thus, volume variability is less evident in QB2 as compared to QB1.

5.4.3 Visualizing the Suitability of Pull Systems

The relative suitability of the different pull systems as evaluated in the analysis model is visualized in figure 27.

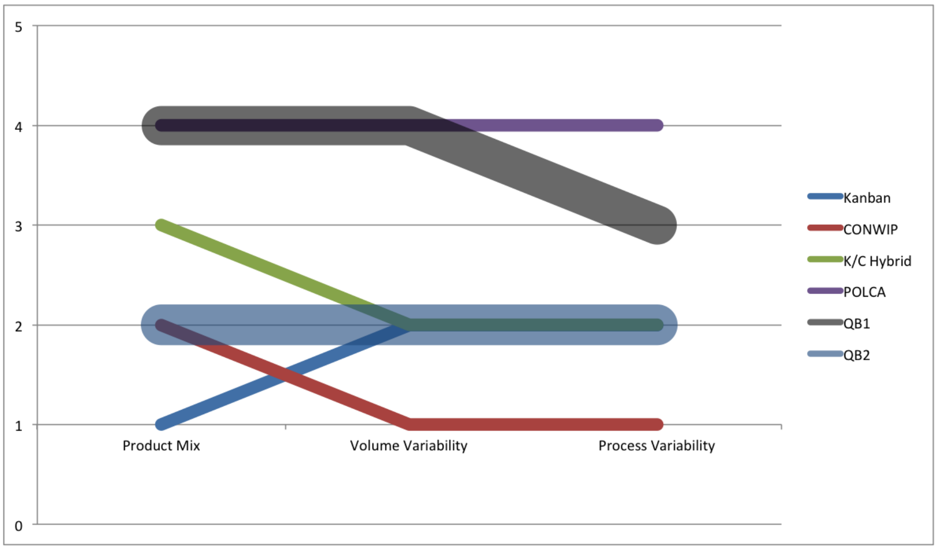

Figure 27. Two-dimensional representation of overlaps between QB and the different pull systems

5.5 How the Differences between White- and Blue-Collar Work Affect the Possibility of Pull

The differences between white- and blue-collar work present some challenges when it comes to achieving the preconditions of pull. Since white-collar workers generally have a higher degree of discretion, agreements tend to be less defined. This combined with the fact that white-collar workers usually have stronger interdependencies makes trust an important issue. Introducing a higher degree of standardization might reduce discretion and interdependencies. The risk of doing so, however, is a less flexible process. Since output measurability is generally lower in white-collar processes, deliveries between process steps might be less defined. When standardizing white-collar tasks, it is important to take the required knowledge into account as input. High variability will generally result in varying routing and resource requirements, therefore, it could be difficult to have fully dedicated resources. In other words, large variations in routing, type of product (mix) and volume make it difficult to have standardized, dedicated resources. Most white-collar products are non-physical. This will require more from control systems in order to make it visible and easy to use. However, compared to the other preconditions of pull, making white-collar processes visible should not be the biggest challenge.

In order to utilize process control in settings with high complexity, two things need to be considered: the nature of the knowledge involved and the trade-off between standardization and discretion (as mentioned above). If the nature of knowledge involved is hard to codify, incomplete and diverse, standardization is only feasible on a rather high abstraction level. For this scenario, outcome-based control is likely to be more suitable.

5.6 Limiting Process Variability in White-Collar Processes

The differences between white- and blue-collar settings are much related to internal process variability. Given the model used for analyzing the different pull systems and relating them to white-collar settings, the parameter that can be affected internally is the process variability. Lowering the process variability would make the environment more receptive to pull-based control systems, as illustrated in figure 28.

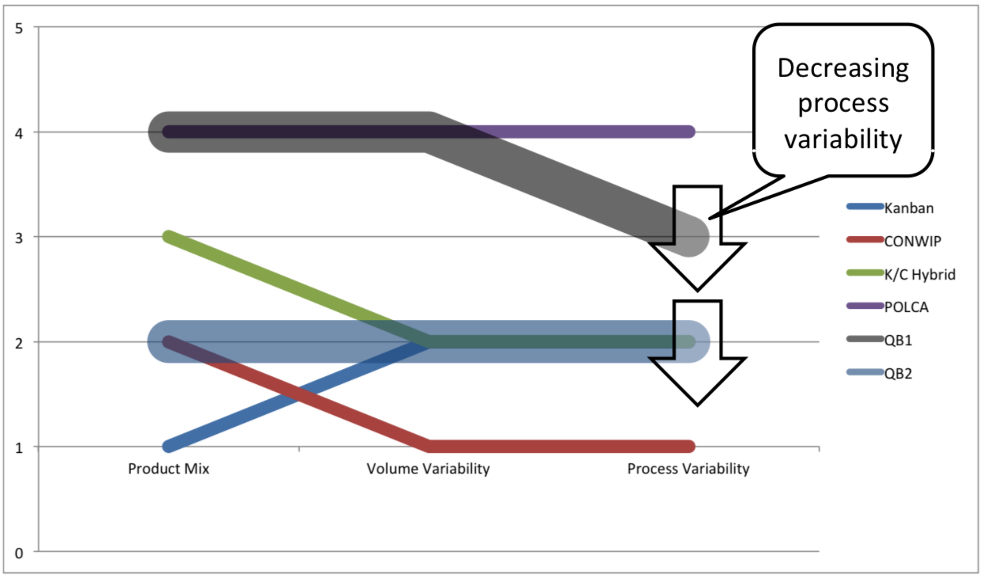

Figure 28. By decreasing the process variability the white-collar settings become more receptive to pull-based control systems

Generally, there are several ways to improve internal processes in terms of capabilities and reduced variability within an organization. The most relevant methods for reducing process variability factors in relation to pull-based control is the focus of the discussion below. The different methods suggested below are not linked to a single issue raised above, rather the total benefit from the combination of the improvement methods should reduce the variability.

5.6.1 Task Discretion vs. Standardized tasks

A big issue with white-collar work is the often incorporated discretion in execution of tasks. The discretion implies a more personalized solution to a task, and consequently higher uncertainty in terms of quality and processing time. Discretion also leads to a variable output that obstruct measurements and effective interfaces between workers. The standardization of tasks in terms of execution would thus be a leap towards incorporating a pull-based control system. Since there are many variants of white-collar settings there are varying suitability for standardization, given the nature of the task. Looking at the setting represented by QB1, the features of knowledge (codifiability, completeness, and diversity) that are representing the nature of tasks indicate a more complex situation. Since QB1 has influences of problem solving and variability in the outcome for every product, standardization will be difficult to achieve. Furthermore, standardizing the task too much would reduce the solution space for the problem which is not desirable in this context. In QB2 on the other hand, the tasks are more clearly defined and can be broken down into parts and standardized. Also, the completeness is bigger for QB2 since a bigger proportion of the knowledge needed can be held by one worker as variability is lower. Thus, the standardization of task can be done in a wider extent than for QB1. Given this reasoning, relaxing the assumption of order is important, meaning that for times when standardization is expensive and might limit the solution space, it is not always desirable.

5.6.2 Visual Control

In a white-collar setting there is a problem of controlling the work for a number of reasons. The lack of visibility makes it hard to clearly understand the load on workers and also to control the flow since it is not visual by nature. Another reason for a more complex control is the higher degree of interdependency between workers, leading to a big amount of possible interactions. Incorporating visual control is a vital tool for easing the control of the flow. It can give immediate response on WIP levels throughout the flow and clearly indicate where items belong, which is beneficial for effective routing. It can also be a support for workers to follow standard procedures by providing immediate feedback if errors occur. Furthermore, it can give immediate feedback to the whole flow as errors occur along a production line, thus increasing the speed of problem solving. The visual control should ideally make the handovers more effective as well, thanks to simple and transparent interfaces between workstations.

5.6.3 Standardized Routing

Due to a higher degree of interdependencies and variation in the kind of work, the routing and sequence of doing things may differ a lot from time to time, increasing the process variability. Standardizing the sequences of execution would thus create a more predictable and controllable flow, reducing the process variability. The standardization is more or less difficult depending on the white-collar setting, it would for example be more complex in QB1 compared to QB2. In complex settings, it might not be possible to standardize all possible routes since they cannot be known beforehand. However, there are always core routings taking part most of the times which should be identified and standardized. For the variation causing unique routing, there could be capacity slots dedicated for demand that is out of the regular scope. Furthermore, when standardization of routing is difficult it is possible to create teams with a broader level of knowledge, thereby making the routing more similar from time to time. Many of these possibilities are utilized in a POLCA control system.

5.6.4 Leveling and Takt

Having worked with the above mentioned methods to reduce process variability, another step for further improvement is leveling of the workload and implementation of a takt time for each specific case. Leveling the workload means that a work stream should be run on approximately the same speed all the time, referred to as the takt time. In order to level out the workload on the capabilities there must be a defined takt time that is a reference for the specific work stream. The leveling and takt will thus make the process more predictable and controllable since deviations can be identified immediately and therefore the actual improvements will be made possible thanks to leveling, not by leveling and takt time in itself. The leveling can be considered to be more or less effective depending on the white-collar setting. For example in QB1 there is relatively big variability in demand, which aggravates the leveling. It is important to be aware of the fact that leveling can be done effectively only after reasonable levels of standardization has been achieved. It would for example not be beneficial to level the workload if there are different sequences of execution for every single task.

To foster effective leveling, it is important to have extensive knowledge sharing and learning throughout. As knowledge is being passed on to more workers, the sequence of execution and need for different routing will be reduced and therefore predictability is increased.

5.6.5 Defined Deadlines and Agreements

As was experienced by the AL team leader, having clearly defined deadlines and agreements regarding what is to be delivered will ease the planning. The strictly defined deadlines allows a team to plan resources and can trust the system to deliver on time. However, the trust can only be achieved after having standardized work and cycle times, to foster accurate planning. The defined deadlines and agreement is much a behavioral issue and common practice within an organization that must be changed. It is about building a culture of trust to the system.

Concluding the analysis of the different methods for reducing process variability, it is clear that no method can be incorporated alone in order to achieve the desirable effects. Rather, the benefits can only be achieved if all the methods are mutually incorporated as they are much related to one another.

5.7 Limiting Product Mix and Volume Variability in White-Collar Processes

Two other main elements of variability that influence the implementation of a pull-based control system are the product mix and the volume variability. However, these two elements cannot be reduced internally, one can only make the processes as robust as possible in order to increase the adaptability of product mix and demand variability into the system. As these two elements are external to the system, the only influence on the nature of variation is strategic decisions. Taking strategic decisions on what kind of products to produce and on what markets to serve would limit the variation. In the case of the QB process there is no possibility to make strategic decisions on the mix of cases to enter the process since each case has risen from a unique quality issue from the market. The method for dealing with product mix and volume variability in this environment is the prioritization of cases, which is the only limiting factor on the natural variation. Thus, the prioritization should be effective enough to sort out all scrap while giving higher priority to the most critical cases. The prioritization should preferably be based from a customer perspective and maximize the customer satisfaction primarily, secondly it should be based on the cost implication for the organization. Methods for prioritization is not a topic of focus for this thesis why it will not be further discussed.

5.8 Adapting Pull Systems to Better Suit White-Collar Settings

From the analysis model it is evident that CONWIP is the least suitable system for both QB1 and QB2. As for Kanban and Kanban-CONWIP Hybrid, Kanban-CONWIP Hybrid is deemed more suitable for QB2 since it is more capable of dealing with high product mix. Furthermore, Kanban requires the pulling of existing products, but as the solutions are always unique in the QB process pulling cannot happen. Both these systems have certain characteristics, resulting in their specific pros and cons, which can be combined and integrated to form a system that is not pure POLCA, nor pure Kanban-CONWIP Hybrid.

Although these two systems have their commonalities, there are some important differences:

- The lowest controlled unit is the cell for POLCA and the operator for Kanban-CONWIP Hybrid. Less detailed control allows for more flexibility.

- Kanban-CONWIP Hybrid systems are organized as sequential lines while POLCA systems utilize a more flexible layout using parallel resources. This enables POLCA to have higher flexibility regarding routing.
- Kanban-CONWIP Hybrid systems require more accurate estimates of time consumption. This will result in more accurate planning but requires a higher degree of standardization and lower product mix.

These factors can be varied to have a more suitable control system with regards to a specific situation. For QB1, POLCA is deemed to allow more process variation than what is actually needed. In this case, there might be room to introduce a more detailed control by reducing the smallest unit of control to operator level. Similarly, Kanban-CONWIP Hybrid systems might “raise” the smallest unit of control in order to be more flexible. The typical layout of the two systems may also be changed. In other words, a POLCA system might be arranged like a sequential line and a Kanban-CONWIP Hybrid system might be organized with parallel, flexible resources and have the routing determined in the central planning unit. These changes will help in designing a system with the required routing flexibility. Kanban-CONWIP Hybrid requires more accurate time estimates than POLCA, but there is nothing saying that POLCA cannot have detailed plans when accuracy is available. When utilizing a Kanban-CONWIP Hybrid system there is also the possibility to increase or decrease the number of cards being circulated. Increasing the number of cards will result in increased flexibility but also increased WIP (i.e. increased throughput time).

All things considered, by understanding the fundamental workings of the different pull systems, different characteristics can be combined and utilized in order to create a system that suits the complexity of a specific context.

5.9 Discussion

The very basis of the research is a theoretical framework consisting of well-known and accepted theories. With this as the starting-point, the application of Lean literature to white-collar settings should have increased validity. Furthermore, experts within the field have supported logical reasoning when moving forward into new areas. Having VCCS as sample context have confirmed much of the literature on white-collar settings and provided two specific white-collar contexts where analysis could be made. The model for analyzing different pull-based control systems and white-collar settings is of the general kind why it should be usable for most white-collar settings. However, since the thesis has focused on high variability settings, if the model were to be applied to low variability settings the ranking of the different pull-based control systems may be totally different. The findings and conclusions are therefore only generalizable to similar white-collar settings where similar evaluation would take place.

Given the purpose of the thesis and the high variability and complexity environment, a push-based control system would not have been beneficial as was concluded from the theory chapter. However, it could have been interesting to complement the model of analysis with this kind of system to make the model more applicable for other kind of settings. This comparison would have been a good way to contrast a push system to a pull system.

Even though the thesis is based on well-known and accepted theories, there is a wide spectra of literature available for Lean and pull-based control topics. The literature is less comprehensive for Lean philosophy related to complex environments, especially for pull-based control systems originating from traditional blue-collar production system. The experts have thus been of great importance throughout the thesis in supporting the way forward. Nevertheless, much of the analysis is based on a combination of theory and independent judgments, why there is a risk of misinterpretations. A trustworthy way of validating the rationale and conclusions would be to perform a pilot case. The wish from VCCS was to also do this pilot, given time restrictions it was though not feasible. It is, however, a suitable proposal for further investigation in the subject.

One could argue there is a lack of empirical research and details concerning the different value streams at VCCS. Since the purpose of the research has been to investigate the possibilities for pull-based control on a principle level, there was no need for detailed value stream mapping. Furthermore, the vision of VCCS (as for any company applying a Lean approach) is to have a flow-based way of working as compared to the functional and over-the-wall approach used today. Therefore, digging into detail on the existing value streams would not add value, it is enough to understand the principle interactions in order to find a conceptual pull-based control system suitable for the existing situation. Ideally, the value streams should be extensively investigated, and waste should be reduced before pull-based control can be successfully implemented. Since much of the understanding of the QB process, especially QB1, is based on interviews with managers, there is a possibility of bias in the answers which is a factor that must be taken into consideration. It is obvious that there is a great deal of complexity given the investigative and developing tasks, but it is not known to the authors to what exact extent. Thoroughly working with observation and value stream mapping could have given the detailed insight.

Even though the thesis is based on well-known and accepted theories, there is a risk for missing factors when performing evaluations and suggesting improvement efforts. The risk is even bigger for this thesis as there was no successful pull-based control system in complex white-collar settings found by the authors. It seems the subject is relatively unexplored why uncertainty is expected to be higher and assumptions must be applied to a greater extent. Once again, an appropriate way of finding possible discrepancies would be to perform a pilot case and learn from the potential failures.

6. Conclusions

In this chapter, the research questions are formally answered.

Q1: What are the enabling preconditions for achieving a pull-based flow in blue-collar settings?

There are three basic preconditions of pull for conventional blue-collar work. Although different authors express these preconditions in slightly different terms, the basic preconditions can be summarized as:

- Defined agreements: There must be a clearly defined agreement between the two operations sharing a refill/withdraw-action that they are connected to one another. Also, the agreement must cover what product mix that is to be shared, the sequence of the model mix and volume limits for the products.

- Dedicated items: All items and resources needed for the interaction between two parties must be explicitly dedicated to them. This means that personnel and storage locations should only be used by the two interacting parties, for example, personnel should not be working on other stations or putting items on different storage locations. Furthermore, there should be a common takt time as reference determining how many items to produce per time period.

- Controlled: There must be a built in control that is easy to see and use between all interacting parties. Having the control mechanism visual and physical, it supports in maintaining the definition and dedication between parties.

These three general preconditions set the stage for what needs to be present and improved. Going slightly deeper into details, the critical principles that are most clearly connected to these preconditions are:

- Leveling out the workload: The smoothing of production is essentially important in leveling out the product mix, the volume, and the sequence of producing the product mix to make the demand more predictable for the production.

- Standardization: Standardization mainly refers to three key elements. First, there needs to be a standard takt time in order to balance the production line in the best way. Second, there must be a standardized operations routine, defining in what order the different operations are to take place and also in what way the tasks should be performed. Third, there must be a standard level of WIP specified for each station.

- Visual control: The visual control encourages people to take corrective actions as well as it provides a quick built-in feedback loop. Being able to incorporate the smoothening and standardization in an organization, together with a visual control system, will enable the successful implementation of a pull system.

Q2: What are the main differences between conventional blue-collar and white-collar settings concerning these preconditions?

In general terms, white-collar work differs from blue-collar work in several ways. Some authors have chosen to distinguish along the dimensions of “routine vs. creativity” others along the dimensions “elusiveness vs. invisibility”.

Summarizing these and other views, the most prominent differences include:

- Higher routing variability
- Higher demand variability
- Higher output variability
- Less repetitive tasks
- Non-physical and low visibility of products
- Higher task discretion
- Stronger interdependencies (need for trust)
- More difficult to measure output (including latent value)

These differences are the primary reasons for complexity in white-collar settings and present challenges when it comes to achieving the preconditions of pull:

Defined: Since white-collar workers generally have a higher degree of discretion, agreements tend to be less defined. This, combined with the fact that white-collar workers usually have stronger interdependencies, makes trust an important issue. Introducing a higher degree of standardization might reduce discretion and interdependencies. The risk of doing so, however, is a less flexible process. Since output measurability is generally lower in white-collar processes, deliveries between process steps might be less defined.

Dedicated: High variability will generally result in varying routing and resource requirements, therefore, it could be difficult to have fully dedicated resources. In other words, large variations in routing, product mix and volume make it difficult to have standardized, dedicated resources.

Controlled: Most white-collar products are non-physical. This will require more from control systems in order to make it visible and easy to use. However, compared to the other preconditions of pull, making white-collar processes visible should not be the biggest challenge.

Q3: How suitable are conventional pull-based control systems in white-collar settings?

By choosing pull system sensibly and/or by making appropriate changes to a specific white-collar context, conventional pull systems can be as suitable for white-collar settings as for blue-collar settings. White-collar settings might have varying degrees of complexity. Hence, the suitability depends on the selection of pull system and the characteristics of the context. Complexity associated with white-collar work may be broken down into the three kinds of variability: process variability, volume variability and product mix. Four conventional pull systems were compared for relative suitability with regards to QB1 and QB2, see figure 29. QB1 represents a context with a high degree of complexity and QB2 represents a context with relatively low complexity. In white-collar settings with high variability, a POLCA system is more suitable and in settings with lower variability a Kanban-CONWIP Hybrid is more suitable. However, these systems may be modified and combined to create systems that suit specific contexts.

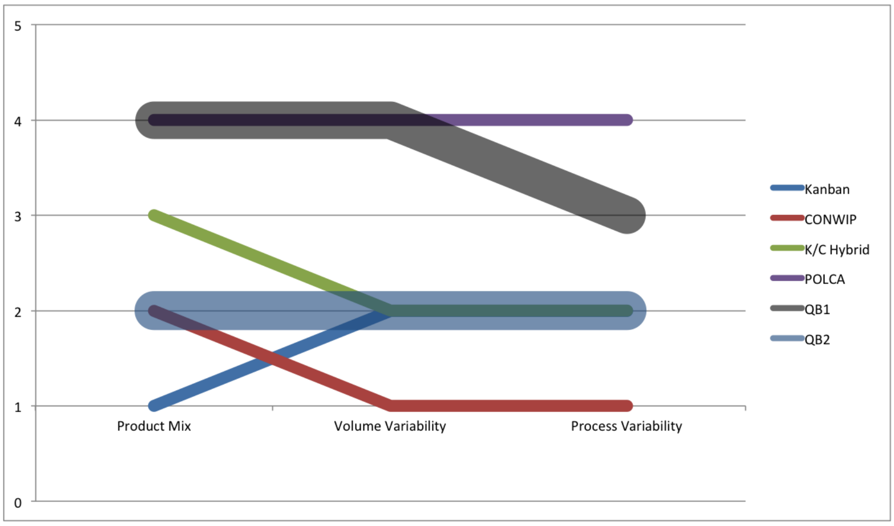

Figure 29. Two-dimensional representation of overlaps between pull systems and two contexts of varying complexity (QB1 and QB2)

Q4: What can be done to increase the suitability of conventional pull-based control systems in white-collar settings?

There are three basic ways of increasing the suitability of pull systems to work in white-collar settings: modify the system to suit the specific needs of the context, change aspects of the context so that it fits the system better, or both.

The aspects of the context most likely to be changed is the internal capability, e.g. to reduce process variability. Regarding the possibility to create a pull-based flow, suitable ways of lowering process variability is to:

- Lower the amount of discretion, i.e. to increase task standardization
- Introduce visual control
- Standardize routes
- Introduce takt time and leveling
- Have well defined, standardized deadlines and agreements

When lowering internal variability, it is important to realize how to balance the trade-off between discretion and standardization. Some contexts require high flexibility in order to work optimally. In other words, there should be no assumption that more order always is better. Other ways of modifying a context is to reduce volume variability and product mix. These are strategic decisions that will lead to prioritizations.

Certain characteristics of different pull systems can be combined and modified to create a customized system. These characteristics include:

- Level of control, e.g. to control at team or operator level.
- Resource layout within the system, e.g. sequential or flexible parallel resources.
- Rules for governing the flow, e.g. how and when interactions should take place and what should happen when certain inventory levels are reached.
- The use of a backlog
- What is pulled, e.g. capacity or actual components
- Amount of signaling cards within the system

- Level of visibility

Suitability of systems might be improved by modifying both the pull system and the context simultaneously. In this way, work is both proactive and reactive as illustrated in figure 30. When process stability is increased, the pull system is modified accordingly. In other words, the system is tuned to work effectively in the improved context. This may be compared to continuously decreasing the number of Kanban cards in a traditional pull system as the process becomes more stable. If, instead, the system were to be designed for a future state right away, the system might be ineffective until that state is reached.

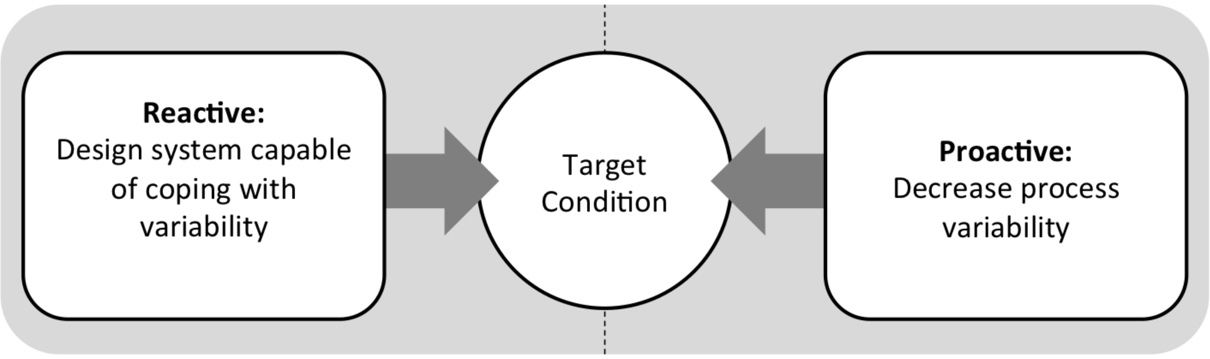

Figure 30. Working both reactively and proactively to reach the target condition

7. Recommendations

In this chapter, managerial implications, with specific recommendations for VCCS, and recommendations for future research are presented. First, possible ways of reducing process variability within VCCS are suggested. Then, a short discussion regarding the reduction of external variability follows. After that, a suggested pull-based control system and how this system might be evolved is presented. Finally, some interesting areas for future research are discussed.

7.1 Managerial Implications

Given the findings in the thesis there are a number of recommendations on the way forward for VCCS. Specific recommendations on the design and requirements of a pull-based control system are directed towards the current situation at VCCS. Also, recommendations concerning how the suggested system might evolve as variability decreases over time are given.

7.1.1 Reducing Process Variability

As mentioned in chapter 6, it is suggested that the target condition is approached from two directions: reduced process variability (see figure 31) and designing a system to cope with variability. Below are a number of suggestions of how process variability might be reduced at VCCS.

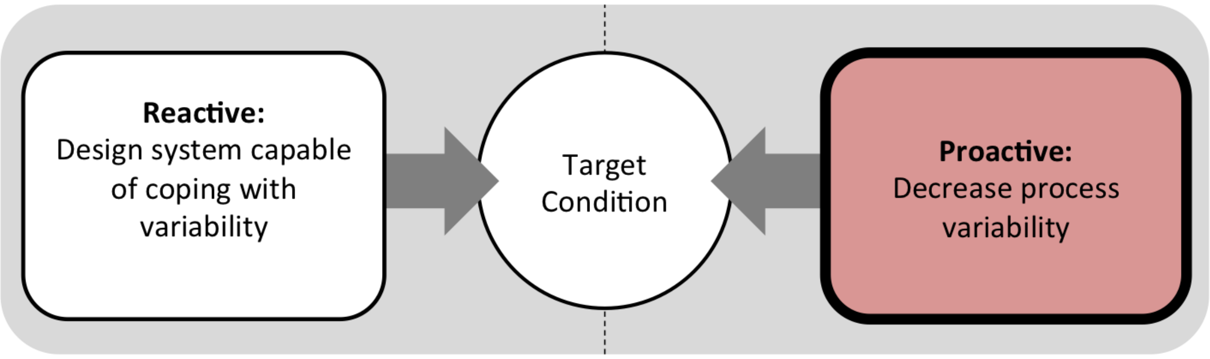

Figure 31. Reducing process variability in order to reach target condition

7.1.1.1 Visual Control

Visualization is important to effectively coordinate the flow and to offer everyone an idea of how the flow works and their place within it. Information may, as supposed to physical entities, be at different places at any time. To spread information regarding the flow, immediate feedback can be achieved. Although VCCS is used to physical visual planning, i.e. to actually write on physical post-its, a digital visualization serves a purpose in this case. Physical planning is very effective when involved parties are in close proximity and it should be utilized at team level. However, to create foresight and understanding, the value stream could be summarized in visualizations on a higher scale. Preferably, digital visualizations should appear on large screens placed so that the entire team can see it rather than in data bases accessed through individual computers. To summarize: digital visualization might be used to connect the different parts of the value stream to create understanding and foresight, and physical planning might be used locally to plan daily/weekly activities.

Aspects that should be visual on a higher level include: cell workload, common backlog (planned activities), planned routes, potentially critical problems, local and total WIP levels, WIP level targets and local and total status against time plan. Color codes should be utilized in order to make the status visualizations as clear as possible. If it is not practical to visualize all of these aspects at once, the digital visualization might be interactive so that it shows the desired level of detail. In other words, all teams should be able to switch between different detail levels and different scopes, including the status of other teams. Also, upstream functions such as R&D and PCO should be able to view the same information in order to create understanding. By being able to see what is happening up ahead, upstream functions might be more effective in prioritizing their activities and what is sent forward to VCCS for processing instead of "throwing items over the wall".

7.1.1.2 Standardization

Standardization is a crucial part in making the work at VCCS more predictable. Until a higher degree of standardization is achieved, predictions need to be done using experience. In other words, operators are left with high discretion were it is not really needed. Some processes within VCCS may be standardized to a very high degree, like the information and product flow. However, VCCS need to be cautious when standardizing not to cause processes demanding discretion by nature to become ineffective.

One of the most important areas for VCCS to standardize is routing and time requirements (for both total and specific operations) for items flowing through the information and products sections. This will be crucial in enabling pull-based control when output is both non-physical and fairly non-repetitive. By constantly do follow-ups regarding deviations from standard, over time, VCCS will learn to do more accurate predictions. This, in turn, will reduce their dependence on experience. Consequently, control will be able to take place efficiently on a more detailed level, ideally for each operator. Other aspects that might be important to standardize in order to increase throughput time and quality include: maximum allowed WIP level and maximum local inventory levels (derived from the point of flow choke).

7.1.1.3 Leveling

One effective and obvious way of leveling the workload is to broaden the skill-set of each operator, making one team capable of dealing with an increased variation of issues. VCCS could, for instance, try to integrate various specialized sections of the car into one competence. Instead of having one operator being able to create a method for, lets say, changing a generator but not for changing an exhaust system, the same operator could learn to create both methods. In this way, the workload might be shared effectively at team or group level. Another way of leveling the workload is to integrate up- and downstream activities into one. An example of this is to integrate the method and time stages of the information flow. In other words, teach the method-operators how to determine a time for their methods and to teach the time-setting-operators how to create a method. This will decrease the number of handovers and increase the possibility to effectively level the capacity. There is, however, a natural limit when individual operators cannot learn more without losing quality on other tasks.

To deal with volume variability it is common to dimension capacity for 80 % of peak demand. This is also recommended to VCCS. It may be difficult to assess what might be considered the right capacity initially. As work is standardized and measured, the capacity of the process and individual teams and operators will become less ambiguous. Since there is still substantial volume variation (compared to traditional mass production) in the information and product flow of VCCS, there will be times of both under- and overcapacity. When there is overcapacity it is important to fill these slots with various improvement work. Since the work performed by VCCS might be both event driven and planned, there is also a need for having excess capacity for urgent, prioritized issues. The spare 20 % capacity might be used for these issues. This is essentially a method for coping with external variability, it will still create a baseline for VCCS to work from, and identified deviations will drive further reduction in process variability.

Many of the activities performed by VCCS are dependent on input from other VCC functions such as R&D. Using deadlines set by the upstream functions to make a rough long-term capacity plan, and to use items that actually have been received to make a detailed short-term plan, VCCS can level both their capacity and product mix thanks to better foresight.

Takt time is usually an important part of leveling. In the case of VCCS, however, there are currently relatively high volume variability and product mix resulting in difficulties utilizing a takt based on mean throughput times. Instead of a common takt for all issues, VCCS can define required time and timing for each product and map their status against plan. In other words, VCCS can utilize deadlines, broken down into the greatest possible detail, to get an idea whether they are on track or not. Time is referred to as expected throughput time, while timing is referred to when a case is to be released into the information flow for example. The important thing is to know what is expected for each product in order to get indications of progress, not to have a common takt time. If a takt time is determined, it may be used to calculate appropriate total WIP levels.

7.1.2 Reducing External Variability

External variability may be reduced through prioritization. VCCS should have a clear view of what is most important to the customer, e.g. quality or quick response time, and make that their first priority. The second prioritization could be internal cost.

7.1.3 A Pull-Based Control System at VCCS

Another important way of approaching target condition is to design a robust system capable of coping with the variability at hand (see figure 32).

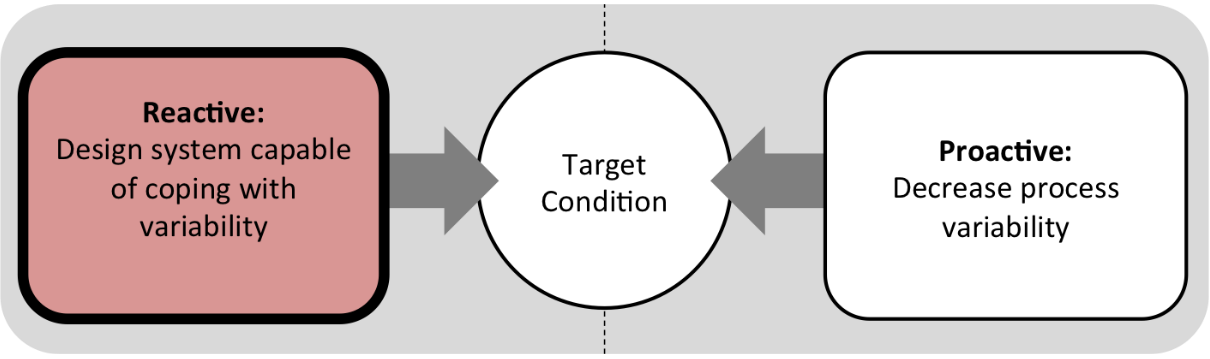

Figure 32. Designing a robust system capable of coping with variability

Although the analysis in chapter 5 indicates that Kanban-CONWIP Hybrid should be the most suitable option for white-collar settings with limited complexity, the recommended system is more POLCA-like. There are three main reasons for this:

- A POLCA system offers greater flexibility and is therefore deemed a better starting point
- A POLCA system can be modified and trimmed easily to allow for less flexibility and greater control as processes within VCCS are improved and more predictable
- A Kanban-CONWIP Hybrid requires relatively accurate predictions, which is currently implausible

As improvements are made, the system can be continuously modified to behave more and more like a Kanban-CONWIP Hybrid incorporating detailed control.

The suggested POLCA system receives input mainly from the market, through AL, and R&D. The conceptual dynamic of the POLCA system as applied to VCCS's meta map is visualized in figure 33. The input items are pushed until they reach a VCCS central planning office, basically acting as the HL/MRP as described in chapter 3.2.4. By letting, or forcing, the upstream activities to take part in prioritizing the items that enter the planning office even this stage could be called pull since items are pulled into the planning office by prioritization rules. By creating more commonality between upstream functions and VCCS, mainly through shared visual information, the deadlines and predictions made at the upstream functions can be integrated into the planning office, creating a forecast (see figure 34). When upstream deadlines draw near, the predictability of when the item will actually enter the backlog becomes higher. The planning office should initially consist of experienced personnel from both VCCS and concerned upstream functions in order to make accurate predictions about time, timing and routing requirements. The planning office should be responsible for the overall plan, authorizing items for specific times and priorities in order to make the whole system flow better.

The information flow is used to exemplify the workings of the suggested pull system (see figure 34) but the system would work in the same way for the product flow. All the teams within the information section should be able to view the forecast, backlog and authorized items (including time, timing and routing) through visual digital boards that are updated continuously. The first team in the route, in this case P2, will pull an item from the list of authorized items. However, P2 can only pull an item once they have received a common POLCA card from the next team in the route, in this case a P2/PE2 card. The item is then pulled through the information flow as described in chapter 3.2.4.

Depending on geographical closeness, the cards might be either physical cards or digital signals at the teams' visual boards. The amount of cards in the system will determine the total WIP. VCCS needs to determine a suitable planning cycle that offers a suitable updating frequency, balancing the benefits of high resolution and efforts of recalculate the number of cards. However, if the system is digitalized and automated the planning can basically be updated constantly.

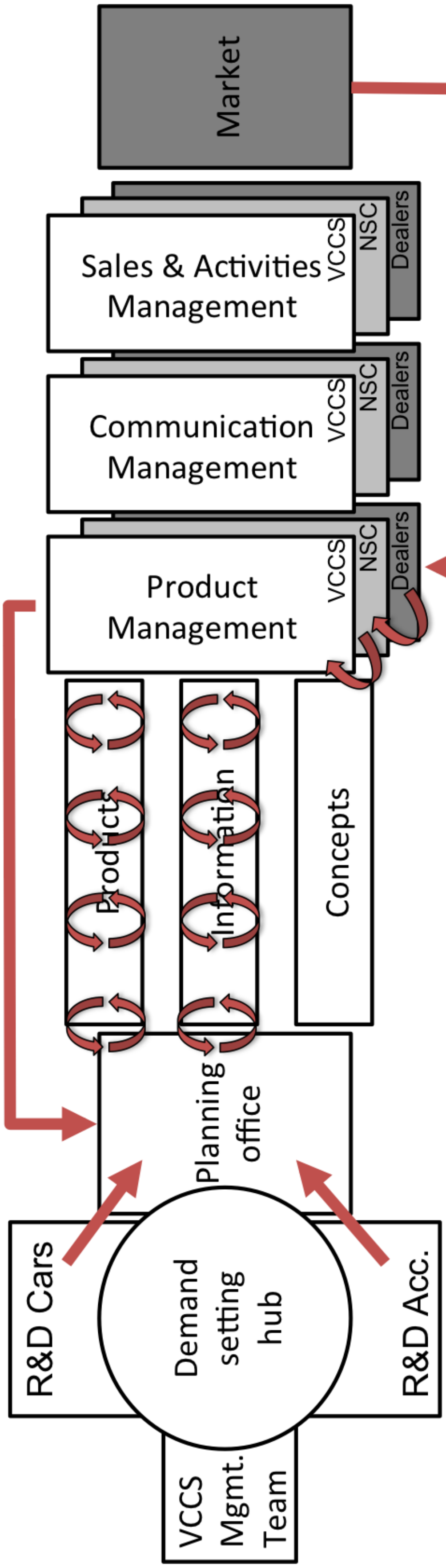

Figure 33. The VCCS meta flow with inputs to the information and product sections from the market and R&D

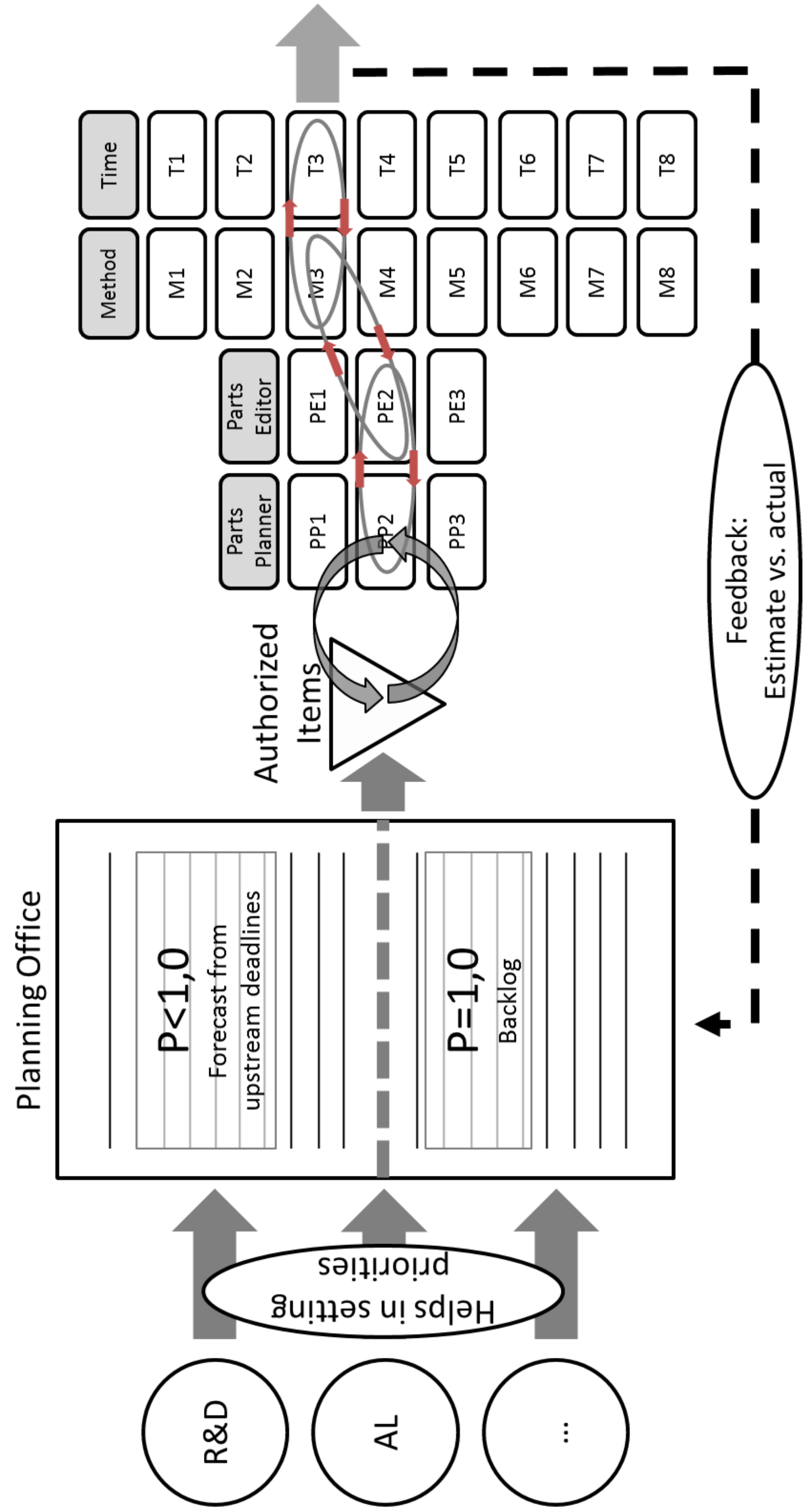

Figure 34. The suggested VCCS pull system with input channels, planning office and information flow. Routing example: P2, PE2, M3 and T3

The suggested pull system has a feedback loop (see figure 34) that is vital for improving predictability. All items should be measured for deviations from the plan at every step of the process and at the end of the process. This is important in order to continuously improve the planning offices' ability to set accurate estimates for time, timing and route. As mentioned, VCCS is initially most likely better off using specific deadlines rather than a common takt because of high variability. In order to learn and to get accurate status updates, these deadlines and plans needs to be broken down and constantly measured and evaluated. This is fundamental in order for VCCS to continuously improve.

7.1.4 A Glimpse into the Future

As VCCS improves their internal capabilities the pull system should also be evolved. By systematically measuring and providing feedback, VCCS, and especially the planning office, will be able to rely less on gut feeling and more on statistics and facts. As time goes by, VCCS will be able to identify main routes and can therefore focus resources where they are needed the most. Routes or teams that are identified as infrequently used might be integrated into the main routes by cross training. As fewer, but more substantial routes are arranged the system will be able to look more like a Kanban-CONWIP Hybrid system. Also, since VCCS will improve their ability to estimate time requirements, it is possible to have the level of confidence required by a Kanban-CONWIP Hybrid system. With improved ability to make estimations it will be possible to collect tasks into groups according to time requirements and required competence in order to reduce effects from the variable product mix.

As processes become more stable, VCCS will be able to control the flow in greater detail, i.e. on operator level. This will possibly make the system more sensitive to fluctuations but will offer greater controllability and less total WIP. By initially controlling at team level, and letting the teams handle the local balancing, the system will promote learning and sharing within the teams, making them more capable of helping each other balance the burden. This will result in more flexible resources. Also, some stages of the process, like Method and Time, should be horizontally integrated so that fewer handovers are required. In other words, VCCS should try to both integrate vertically and horizontally to facilitate a better flow. The pull-based control system facilitates and promotes effective handovers, however, the ultimate objective should be to reduce the number of handovers. Another positive aspect of having flexible resources is that if a problem occurs, e.g. a deadline is missed, more resources can be allocated in order fix the problem quickly. With only specialized resources, this is not an option.

Having more flexible resources and an improved ability to do predictions opens up for the possibility of increased frontloading, i.e. to focus resources early on in order to decrease process time and effort towards the end, where generally cost are at its highest. Frontloading can cause complexity to decrease rapidly and by having high confidence in forecasts. Issues with tight deadlines can even be started before the upstream function is 100 % ready. Frontloading is thus a method for the investigation and problem solving before a case is sent to VCCS, and therefore a more reliable forecast can be delivered to VCCS faster.

All in all, VCCS will hopefully enjoy a system with clearer routing, more manageable product mix, decreased WIP and throughput time. In general, the system will be more predictable and controllable.

7.2 Future Research

The most logical next step is to test the findings from this thesis in reality. It would be interesting to test the results in a pilot, evaluating a context before and after implementing a pull system like the one recommended for VCCS. In doing so, it is important to define metrics significant to both internal and external customer satisfaction such as throughput time, accuracy, quality and operator stress level. Then evaluation can take place to see whether the metrics improve or worsen after implementing such a pull system in a white-collar setting.

Since there are clear difficulties controlling white-collar organizations that have highly customized output, it would be interesting to look at how modularization could be a possible remedy. After all, similar situations in manufacturing are usually solved by using modules. If modularization strategies were to be utilized in white-collar settings it would reduce the unpredictability of having a high product mix while still being able to offer a highly customized output. This, in turn, would mean that pure pull strategies could be approached.

An interesting topic would be to further use the model of analysis, used for the different pull-based control systems and white-collar settings, to see how well it can be utilized for different white-collar settings. Using the findings to do a pilot case on that context would provide insight on the generalizability of the model.

The recommendations of this thesis expect the effects from variability to be lowered as standardization and continuous measurement and follow-up will strengthen predictability of VCCS’s information flow. This makes much sense, however it would be interesting with detailed research into how much better the predictability can get. There is a chance that the variations are too big and random that better predictions cannot be made. If that would be the case, it is an important insight when striving for pull-based control in white-collar settings.

Bibliography

Adler, P. S., Mandelbaum, A., Nguyen, V., & Schwerer, E. (March 1995). From Project to Process Management: An Empirically-based Framework for Analyzing Product Development Time. *Management Science, 41*(3), 458-484.

Blackburn, J. D. (July-August 1992). Time-based Competition: White-Collar Activities. *Business Horizons,* 96-101.

Bonvik, A. M., Couch, C. E., & Gershwin, S. B. (1997). A comparison of production-line control mechanisms. *INT. J. PROD. RES., vol. 35*, 789-804.

Bryman, A., & Bell, E. (2011). *Business Research Methods* (Third edition uppl.). Oxford: Oxford University Press.

Chen, J. C., & Cox, R. A. (2012). Value Stream Management for Lean Office - A Case Study. *American Journal of Industrial and Business Management, 2*, 17-29.

Conant, R. G. (September 1988). JIT In A Mail Order Operation Reduces Processing Time From Four Days to Four Hours. *Industrial Engineering*, 34-37.

Cuatrecasas-Arbos, L., Fortuny-Santos, J., & Vintro-Sanchez, C. (2011). The Operations-Time Chart: A graphical tool to evaluate the performance of production systems - From batch-and-queue to Lean manufacturing. *Computers & Industrial Engineering*, 663-675.

Drucker, P. F. (1999). Knowledge-Worker Productivity: THE BIGGEST CHALLENGE. *California Managment Review, 41*(2).

Esparrago, R. A. (1988). KANBAN. *Production and Intentory Management Journal*, 6-10.

Feather, J. J., & Cross, K. F. (1988). Workflow analysis, just in time techniques simplify administration. *Industrial Engineering, 20*(1), 32-39.

Gonzales-Rivas, G., & Larsson, L. (2011). *Far from the Factory: Lean for the Information Age.* New York: Productivity Press.

Hamm, S. (2005). Taking A Page From Toyota's Playbook. *Business Week*(3948).

Holmdahl, L. (2010). *Lean Product Development På Svenska* (1.6 uppl.). Göteborg: DigitaltryckNu.

Hopp, W. J., & Spearman, M. L. (2004). To Pull or Not to Pull: What Is the Question? *Manufacturing & Service Operations Managment, 6*(2), 133-148.

Hopp, W. J., Iravani, S. M., & Liu, F. (2009). Managing White-Collar Work: An Operations-Oriented Survey. *Production and Operations Management, 18*(1), 1-32.

Huberman, B. A., & Hogg, T. (1994). Communities of Practice: Performance and Evolution. *Dynamics of Computation, Xerox Palo Alto Research Center*.

Kurtz, C. F., & Snowden, D. J. (2003). The new dynamics of strategy: Sense-making in a complex and complicated world. *IBM Systems Journal, 42*, 462-483.

Liberopoulos, G., & Dallery, Y. (2000). A unified framework for pull control mechanisms in multi-stage manufacturing systems. *Annals of Operations Research*, 325-355.

Liker, J. K. (2004). *The Toyota Way.* McGraw-Hill.

Liker, J. K., & Convis, L. G. (2012). *The Toyota Way to Lean Leadership: Achieving and Sustaining Excellence through Leadership Development.* McGraw-Hill.

Liker, J. K., & Meier, D. (2006). *The Toyota Way Fieldbook: A Practical Guide for Implementing Toyota's 4Ps.* McGraw-Hill.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Sage Publications.

Little, J. D. (1961). A proof for the queueing formula: L = kW. *Operations Research, 9*(3), 383-387.

Locher, D. (2011). *Lean Office and Service Simplified: The Definitive How-to Guide.* New York: Productivity Press.

Marsh, R. F., & Conard, M. A. (2008). A pull system for delegating knowledge work. *Operations Management Research,* 61-68.

McNamar, T. (Nov/Dec 1973). White Collar Job Enrichment: The Pay Board Experience. *Public Administration Review, 33*(6), 563-568.

Monden, Y. (1998). *Toyota Production System: An Integrated Approach to Just-In-Time.* Chapman & Hall.

Nidumolu, S. R., & Subramani, M. R. (2003). The Matrix of Control: Combining Process and Structure Approaches to Managing Software Development. *Journal of Management Information Systems, 20*(3), 159-196.

Palvannan, K. R., & Teow, K. L. (2012). Queueing for Healthcare. *J Med Syst, 36*, 541-547.

Ramírez, Y. W., & Nembhard, D. A. (2004). Measuring knowledge worker productivity. *Journal of Intellectual Capital, 5*(4), 602-628.

Ramírez, Y. W., & Nembhard, D. A. (2004). Measuring knowledge worker productivity: A taxanomy. *Journal of Intellectual Capital, 5*(4), 602-628.

Sendil Kumar, C., & Panneerselvam, R. (2007). Literature review of JIT-KANBAN system. *International Journal of Advanced Manufacturing Technology, 32*, 393-408.

Spearman, M. L., Woodruff, D. L., & Hopp, W. J. (1990). CONWIP: a pull alternative to kanban. *Int. J. PROD. Res., vol. 28*, 879-894.

Spender, J.-C., & Kijne, H. J. (1996). *Scientific Management: Frederick Winslow Taylor's Gift to the World?* Kluwer Academic Publishers.

Stevenson, M., Hendry, L. C., & Kingsman, B. G. (March 2005). A review of production planning and control: the applicability of key concepts to the make-to-order industry. *International Journal of Production Research, 43*(5), 869-898.

Suri, R., & Krishnamurthy, A. (2009). Planning and implementing POLCA: a card-based control system for high variety or custom engineered products. *Production Planning & Control, 20*(7), 596-610.

Swank, C. K. (2003). The Lean service machine. *Harvard Business Review, 81*(10), 123-129.

Tanaka, T. (2011). TMS Toyota's Management System.

Turner, K. L., & Makhija, M. V. (2006). The Role of Organizational Controls in Managing Knowledge. *Academy of Management Review, 31*(1), 197-217.

Womack, J. P., & Jones, D. T. (2003). *Lean Thinking: banish waste and create wealth in your corporation.* London: Free Press Business.