TO PULL OR NOT TO PULL:
WHAT IS THE QUESTION?

Wallace J. Hopp†and Mark L. Spearman‡

†Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, Illinois 60208
hopp@northwestern.edu

‡Factory Physics, Inc.
5107 Laurel Valley Court, Suite 200
College Station, Texas 77845
www.factoryphysics.com

August 15, 2002
Revised July 23, 2003

Abstract

   The terms *pull* and *lean* production have become cornerstones of modern manufacturing practice. But, although they are widely used, they are less widely understood. In this paper, we argue that while the academic literature has steadily revealed the richness of the pull/lean concepts, the practitioner literature has progressively simplified these terms to the point where serious misunderstandings now exist. In hopes of reducing confusion, we offer general but precise definitions of pull and lean. Specifically, we argue that pull is essentially a mechanism for limiting WIP and lean is fundamentally about minimizing the cost of buffering variability.

1   Introduction

We have been teaching executive courses on manufacturing regularly since the mid-1980's. Starting at the height of the just-in-time craze and continuing through the time-based manufacturing, business process reengineering, enterprise resource management and lean manufacturing movements, we have been doing our best to explain pull production to people who really want to use it. We even wrote a book on the subject [18]. With all of our experience and all of the attention pull has received in the practitioner literature, you would think our task would be getting easier.

   But it's not. In recent years, both of us have been finding increasing confusion among our students about how pull is defined and decreasing willingness to accept our definition. The main controversy has usually stemmed from some participants (who have almost always received some sort of lean training prior to attending one of our classes) insisting pull means making products to order, as opposed to making them to stock or forecast. Interestingly, this is a fairly recent phenomenon. Ten years ago, our students equated pull with kanban, a make-to-stock system, but not with make-to-order. Nevertheless, it has been a disturbing experience for us, since we do not feel that pull is properly defined as either kanban or make-to-order.

   Had we been wrong all along? The sheer number of confrontations in our classes forced us to consider that possibility. To decide, we revisited fundamental issues we thought were long resolved, such as what constitutes a pull system?, what makes pull work?, what is lean and how does it relate to pull? In addition, we reviewed carefully the practitioner and academic literatures related to pull. This paper is the result of our soul-searching process. In it, we:

1. Provide a history of pull, from its antecedents up through recent history when trends have caused the distinction between push and pull to become confused.

2. Describe the essence of pull and how "strategic pull" differs from "tactical pull."

3. Investigate the relationship of pull and push to the concepts of make-to-order and make-to-stock.

Although we cite a significant amount of literature that is important to understanding pull and its history, this paper is not intended as a comprehensive review of the pull literature. For that, we refer the reader to excellent reviews by Uzsoy and Martin-Vega [51] and Huang and Kusiak [23].

2 A Brief History of Pull

2.1 First There Was MRP

To understand why and how pull came about, it is necessary to first appreciate the environment that preceded it, namely the world of MRP. Prior to the dominance of the computer in manufacturing, inventory was controlled using reorder-point/reorder-quantity (ROP/ROQ) type methods. During the 1960's, Joseph Orlicky, Oliver Wight, and George Plossl along with others developed a new system, which they termed *Material Requirements Planning (MRP)*. Orlicky obviously believed that they were on to something big; he subtitled his book on the subject "The New Way of Life in Production and Inventory Management" [37]. After a slow start, MRP began to gather steam during the 1970's fueled by the "MRP Crusade" of the American Production and Inventory Control Society (APICS). Orlicky reported 150 implementations in 1971 [37]. By 1981, the number had grown to around 8,000 [56]. As it grew in popularity, MRP also grew in scope, and evolved in the 1980's into *Manufacturing Resources Planning (MRP II)*, which combined MRP with Master Scheduling, Rough-Cut Capacity Planning, Capacity Requirements Planning, Input/Output Control and other modules. In 1984 alone, 16 companies sold $400 million in MRP II software [60]. By 1989, over $1.2 billion was sold to American industry, constituting just under one-third of the entire software industry [24].

While MRP was steadily dominating the American production control scene, history was taking a different course in Japan. There, perhaps because it lacked a strong indigenous computer industry, the computer was far less pervasive in production and inventory control. Instead, several Japanese companies, most notably Toyota, developed the older ROP/ROQ methods to a high level. Starting in the 1940's, Taiichi Ohno began evolving a system that would enable Toyota to compete with American automaker but would not depend on efficiencies resulting from long production runs that Toyota did not have the volumes to support. This approach, now known as the "Toyota Production System," was designed to "make goods, as much as possible, in a continuous flow" [36].

According to Ohno, the Toyota Production System rests on two "pillars":

1. "Autonomation"

2. Just-in-time production

Autonomation, or "automation with a human touch," is the practice of determining the optimal way to perform a given task and then making this the "best practice" standard method. Autonomation also involved "fool proofing" or "*poke yoke*," which involved using devices to quickly check dimensions and other quality attributes to allow workers to be responsible for their own quality. If problems were found, the production line stopped until the problems were corrected. This eliminated the need for rework lines and, eventually, eliminated most scrap. The Toyota Production System also promoted "5S," *Seiri, Seiton, Seiso, Seiketsu, and Shitsuke* which are organization and housekeeping techniques aimed at achieving Autonomation and Visual Control.

Just-in-time production according to Ohno involved two components: kanban and level production. Kanban or "pull production" became the hallmark of the Toyota Production System (which was also frequently referred to as Just-In-Time) to the point where many thought they were synonymous. But kanban was just a means to an end. Ohno famously described his

inspiration for kanban as coming from a visit to the U.S. during the 1950's in which he was more impressed with American supermarkets than with American manufacturing. The idea of having all goods available at all times was, to Ohno, novel and revolutionary. He said:

> *From the supermarket we got the idea of viewing the earlier process in a production line as a kind of store. The later process (customer) goes to the earlier process (supermarket) to aquire the needed parts (commodities) at the time and in the quantity needed. The earlier process immediately produces the quantity just taken (re-stocking the shelves).* [36, p. 26]

To do this, Ohno had to make some major system changes. If the supermarket was to replenish only what was just taken in a timely manner, lot sizes had to be drastically reduced. To achieve the efficiencies needed, Ohno and his Toyota colleagues found many creative ways to reduce change-over times. Although change did not happen overnight (the JIT revolution was really an evolution), the results were substantial. In 1945 setups on large presses took 2-3 hours. By 1962 they had been reduced to 15 minutes and by 1971 some were down to 3 minutes [36]. With such short change-overs, Ohno could achieve "one piece flow" and just-in-time production.

By the early 1980's, American manufacturers had become acutely aware that they had fallen behind in manufacturing innovation and even in manufacturing efficiency (especially in the automotive sector). Even while MRP sales continued to climb, many were thinking that MRP had been a mistake. A 1980 survey showed that less than 10% of the firms interviewed had recouped their investment within two years [19]. JIT began being hailed as the next great thing.

American managers quickly became enamored with everything Japanese. American professors went to Japan to learn first-hand what was going on, and of course they wrote books. The first JIT book, published in 1981, was Hall's *Driving the Productivity Machine: Production and Control in Japan* [15]. This was followed by Schonberger's, *Japanese Manufacturing Techniques: Nine Lessons in Simplicity* in 1982 [38]. By 1983 Yasuhiro Monden, a Japanese professor, got on the bandwagon with *The Toyota Production System* [35]. Shigeo Shingo, who worked with Ohno, published a book on setup reduction in 1985, *The SMED System* [39]. Ohno's book finally appeared in English in 1988 [36].

Despite the frenzy of interest in JIT throughout American industry during the 1980's, results were mixed. Firms implementing JIT were faced with a deceptively simple philosophy and a complicated array of techniques (see [61] for an elegant discussion of the disconnect between "romantic" JIT (philosophy) and "pragmatic" JIT (techniques)). So, managers had to contribute considerable site-specific innovation to produce workable systems. Depending on how creative and insightful they were, JIT sometimes worked and sometimes did not.[1]

By the end of the 1980's, JIT began being eclipsed by the next great thing—*Enterprise Resources Planning (ERP)*. With the development of the client/server information technology architecture, it became feasible to integrate virtually all of a corporation's business applications with a common data base. ERP offered both near-total integration and "best of breed" software

---

[1]The early confusion about JIT had many philosophical, cultural and technical explanations. But it may also have been partly intentional. In a 1990 interview, Ohno claimed that Toyota considered the system so powerful that they deliberately coined misleading terms and words to describe it. "If in the beginning the U.S. had understood what Toyota was doing, it would have been no good for us." Eventually, however, Toyota became very open and invited the whole world to see their factories in the 1980's and 1990's. [34].

in the specific applications. Of course, ERP was much more complex than MRP II, containing modules for every business function imaginable, from accounting and financial functions to human resources. And it was correspondingly more expensive, with implementation costs at some companies soaring as high as $250 million [9]. In spite of the price tag and the growing number of implementation horror stories, ERP continued to grow in popularity. As the 1990's drew to a close and fear of the Millennium Bug intensified, ERP was being installed at a feverish rate.

As ERP began its rise, it appeared that the JIT movement had run its course. Even though Toyota continued to rank at the top of the automotive industry in quality and efficiency metrics, interest in the Toyota Production System was on the wane. But in 1990, a landmark case study conducted by MIT was published in *The Machine That Changed the World* by Womack, Jones and Roos [57]. This study compared American, European, and Japanese automobile manufacturing techniques and concluded in no uncertain terms that the Japanese methods, particularly those of Toyota, were vastly superior. In addition, the authors freshened JIT by recasting it as "Lean Manufacturing." With a new name and a new set of stories to rekindle interest, the system created by Taiichi Ohno again became a hot topic in the world of manufacturing.

The JIT movement also spawned a separate movement that ultimately became larger than JIT itself—Total Quality Management (TQM). Originally cast as a means for facilitating smooth production flow, TQM grew into a popular management doctrine institutionalized in the ISO 9000 Certification process. The focus on TQM in the 1980's also spurred Motorola to establish an ambitious quality goal and to develop a set of statistical techniques for measuring and achieving it. This approach became known as "Six Sigma" and was eventually adopted by companies such as Allied Signal and then General Electric. Six Sigma entered the mainstream management lexicon when Jack Welch, the charismatic CEO of GE, declared that it played a major role in his company's financial success. Today, Six Sigma carries on the legacy of TQM just like lean carries on the legacy of JIT.

Despite traveling separate paths since emerging from JIT, it now appears that the Lean Manufacturing and Six Sigma movements are about to merge, as suggested by the recent best selling management book, *Lean Six Sigma* [20]. At the same time, the computer approach characterized by MRP/ERP is also undergoing consolidation. After Y2K proved to be a non-event, ERP became passe as a term but was quickly replaced by SCM (Supply Chain Management). Remarkably, SAP, the world's largest provider of ERP software, changed its entire product line in a matter of months (or so it would appear when a search of their web-site revealed that all references to ERP were gone and had been replaced with the new SCM acronym). Since both ERP and SCM trace their roots back to quantitative production and inventory control, it is perhaps not surprising that they would wind up embodied in the same software products. However, all this leads one to wonder when the Lean Six Sigma Supply Chain Management movement will begin.

3 Pull Research

While corporate America tried to navigate this sea of buzz words and to comprehend the Toyota Production System along with the mysterious concept of pull, those in the academy were busy as well. As with any new management trend, academic researchers were quick to get on the JIT bandwagon. Hall and Schonberger were both academics who went to Japan to study just-

in-time first-hand. Monden, of the University of Tsukuba, did not have to travel as far. It appears the first academic paper describing kanban was published by Japanese researchers in 1977. The title of the paper by Sugimori, et al. [47] is telling: "Toyota Production System and Kanban System: Materialization of Just-In-Time and Respect-For-Human System." Like Orlicky's "New Way of Life," the Toyota Production System is something big that goes beyond production control. The title of this first paper may also have contributed to the mistaken equivalence between the Toyota Production System and kanban. It was followed four years later by another paper by two other Japanese researchers, Kimura and Terada [30]. Both of these papers described the mechanics of kanban and the requirements for its implementation. As such, they set the stage for the books of Schonberger, Hall and Monden. Hence, by the early 1980's the *mechanics* of kanban had been widely described. Because it represented the first system to be termed a "pull" system, and is hence central to the subject of this paper, we briefly describe these mechanics below.

3.1 Kanban Mechanics

The classic version of kanban as pioneered by Toyota is called a "two card kanban" system, as shown schematically in Figure 1. Production begins when a material handler with a **move card** removes a **standard container** of parts from the **outbound stock point**. The move card authorizes that handler to take these parts and tells him where the parts are needed. Before the container is removed, the **production card** is removed from the standard container and placed on the **production card board**.[2] Production cannot begin without a production card, a container of the appropriate incoming parts, and an idle work station. When all three are available, the worker removes a standard container of parts from the **inbound stock point**, removes the move card from the container, places it in the hopper for move cards at the work station and begins to process the parts. Periodically, a material handler collects the move card(s), locates the needed parts, transports them to the work station, and the process repeats at the next work station upstream.

It is easy to see that the two card kanban system is the result of an artificial distinction between parts processing and material movement. If we include material movement as a separate process, we see the two card kanban system becomes a one card system with the move card becoming a "production" card for the move process. This is illustrated in Figure 2, which also shows that a kanban system is essentially a serial production system with blocking.

3.2 What Is So Special About Pull?

But if kanban is nothing more than a serial production system with blocking, then why all the fuss? Systems with blocking occur naturally (e.g., in automotive systems where limitations on space and the number of pallets can cause downstream congestion to shut down upstream production). Why would Toyota deliberately constrain the flow in their system? Researchers and practitioners have had to address this fundamental question since the beginnings of JIT.

The first, and easiest, step in understanding kanban in specific, and pull in general, is to characterize its benefits. These have been widely cited as (see, e.g., [17, 11]):

---

[2]Some authors indicate that the term "kanban" refers to the card while others indicate that it is the board holding the cards.
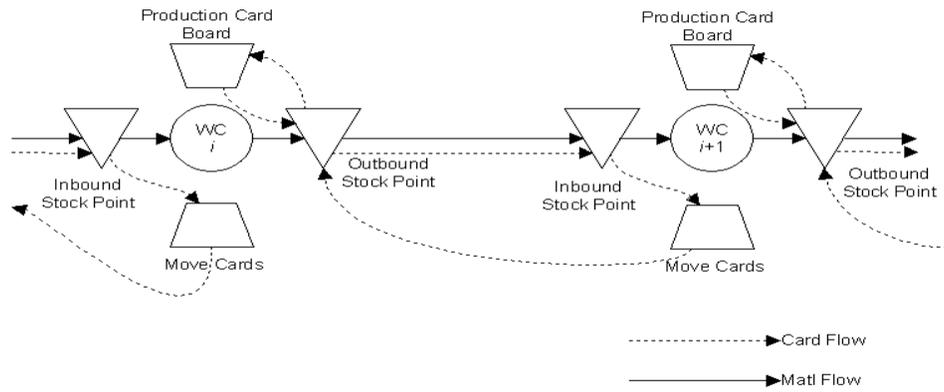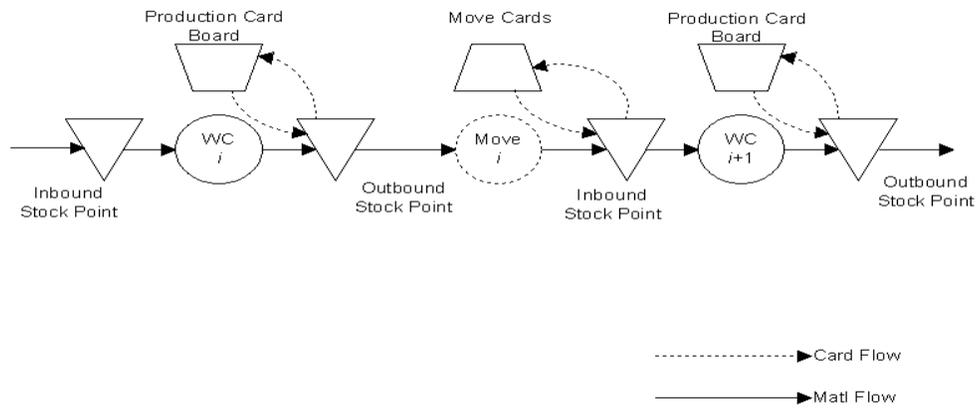
Figure 1: A two card kanban system.

Figure 2: Equivalence of the one and two card kanban systems.

1. *Reduced WIP and Cycle Time:* By limiting releases into the system, kanban regulates WIP and hence results in a lower average WIP level. By Little's Law, this also translates into shorter manufacturing cycle times.

2. *Smoother Production Flow:* By dampening fluctuations in WIP level, kanban achieves a steadier, more predictable output stream.

3. *Improved Quality:* A system with short queues cannot tolerate high levels of yield loss and rework because these will quickly shut down the line. Additionally, short queues reduce the time between creation and detection of a defect. As a result, kanban both applies pressure for better quality and provides an environment in which to achieve it.

4. *Reduced Cost:* By switching the control from release rate to WIP level (card count), kanban provides an explicit means to "stress" the system. Each reduction in WIP will cause problems (e.g., a setup is too long, assemblies have fit problems, worker breaks are uncoordinated, etc.) to show up in the form of blocking and starving in the line. Only by solving these can the progression toward "leaner" (lower inventory) production proceed. This process was widely described via the analogy of lowering the water (inventory) in a river to find the rocks (problems). The end result is a more efficient system with lower costs.

The second, and subtler, problem is to identify what it is about kanban that yields these benefits. This is important if we are to understand how much kanban can be simplified or generalized and still be effective. In essence, this is the question of what really constitutes pull.

The water-and-rocks analogy suggests that the benefits of kanban stem from environmental improvements (removing "rocks" or problems). Some research has supported this view. For instance, Huang et al. [22] performed a simulation of kanban and concluded that without significant changes in the manufacturing environment, kanban would not be successful in the U.S. Krajewski et al. [31] performed an even more comprehensive simulation study and concluded that the benefit of kanban is due more to an improved environment than to any fundamental change in logistics.

But ascribing kanban's benefits to environmental improvements does not explain exactly what about kanban leads to the improvements. Furthermore, most researchers were not willing to credit environmental improvements as the sole explanation of all of JIT's benefits. Ohno clearly regarded kanban as essential to JIT and many companies had used versions of it to great success. Moreover, if the flow control method didn't matter, then successful JIT companies could have reverted to MRP after improving their environments. But this did not happen. Hence, to understand more fully what makes kanban work, researchers turned to mathematical models.

Karmarkar appears to have been first to note the similarities of kanban to earlier base stock systems in two simple and insightful papers [26, 27]. The system he discusses is very similar to a traditional base stock system as described by Simpson (1958). Simpson attributes this system to George Kimball in an unpublished Arthur D. Little report, which Karmarkar had published in 1988 [29].

Simpson describes the base stock system as follows:

When an order is placed, it is filled from inventory if the inventory is not zero. If the inventory is zero, the order is placed in a backorder file, to be filled when an item arrives. In any event, a manufacturing order is immediately placed with the preceding manufacturing operation to produce an item to replace the item that has been consumed. The manufacturing operator, in turn, immediately places an order for the required raw materials against the preceding inventory, and as soon as this order is filled [i.e., he has the needed inventory], proceeds to 'operate' on it to produce the required item. In this way an order against the last inventory for a finished item is immediately transmitted all the way back along the line to all the manufacturing operations, each of which is galvanized into production.

But, while kanban and the base stock systems are similar, they are not identical. The difference was described by Spearman [45] as follows:

Kanban would not place an order for more parts if a demand (in the form of a move card) had arrived when there was no stock in the outbound stock point. Instead, there would be one or more production cards already in process and whenever one of these was completed, the waiting move card would be immediately attached to the recently completed container of parts and the production card would be sent back into production. In this way a kanban system bounds the amount of WIP there can be in the system while the base stock system does not.

Spearman et al. [42] and Spearman and Zazanis [46] found that while specific environmental improvements are certainly influential (e.g., setup reduction, production smoothing), there are three primary logistical reasons for the improved performance of pull systems:

1. *Less Congestion:* Comparison of a open queueing network with an "equivalent" closed one shows that the average WIP is lower in the closed network than the open network given the same throughput. The effect is relatively minor and is due to the fact that queue lengths have no correlation in an open system but are negatively correlated in a closed queueing network, an observation made earlier by Whitt [54].

2. *Easier Control:* This is a fundamental benefit, which results from several observations:

   a. WIP is easier to control than throughput since it can be observed directly.
   b. Throughput is typically controlled with respect to *capacity.* Because it cannot be observed directly, capacity must be estimated by considering process time, setup time, random outages, worker efficiency, rework and other factors that affect the potential rate of production.
   c. Throughput is controlled by specifying an *input* rate. If the input rate is less than the capacity of the line then throughput is equal to input. If not, throughput is equal to capacity and WIP builds without bound. By incorrectly estimating capacity, input can easily exceed the true capacity. This is particularly true when seeking high utilization rates. As a result, systems that control WIP are substantially more *robust* to control errors than are systems that control throughput (see Chapter 10 of [18] for a mathematical illustration of this robustness effect).

8

3. *WIP Cap:* The benefits of a pull environment owe more to the fact that WIP is bounded than to the practice of "pulling" everywhere. This was argued by observing that a simple overall bound on the WIP (i.e., as in a closed queueing network without blocking) will promote the same benefits as those cited for kanban and by showing that the throughput of a closed queueing network without blocking is greater than that of a closed queueing network with blocking (i.e., a kanban system).

Based on these findings, Spearman and Zazanis proposed a hybrid push/pull system known as CONWIP that possesses the benefits of kanban but can be applied to more general manufacturing settings.[3] A number of properties of CONWIP systems were demonstrated using simulation in Spearman, et al. [42], which also described several practical implementation considerations. The book *Factory Physics* [18] also discussed several ways in which CONWIP can be implemented, including using CONWIP in make-to-order environments [18]. Independent work by Veatch and Wein [52] also concluded that the WIP constraint is a major source of the benefit of kanban.

3.3   Comparisons and Generalizations

During the late 1980's and early 1990's a number of researchers compared different versions of kanban, CONWIP and other systems. A good overview of the work before 1996 can be found in [23].

As more variants of kanban and related systems were proposed, researchers began to seek ways to unify them. For example, Buzacott described a "generalized kanban policy" in [7] and introduced the concept of "production authorization cards." This concept was developed further in [8] where it was shown to represent MRP, kanban, CONWIP, as well as many hybrid systems. Using this framework and sample path arguments, Buzacott and Shanthikumar showed that CONWIP is the optimal structure for maximizing throughput for any WIP limiting system. Tayur [49] also considered the question of the optimal pull structure, viewing systems as a continuum from kanban to CONWIP depending on how they are partitioned into WIP controlled segments and considered how to optimize the allocation of cards in such systems so as to maximize throughput.

In another unification paper, Axsäter and Rosling [2] showed that kanban is a special type of "installation stock" policy (i.e., one in which the reorder point is an integer multiple of the reorder quantity and where backlogs are not subtracted from the inventory position). They also defined an "echelon stock policy" in which the inventory position is the sum of the installation inventory positions for the installation and all downstream installations. They then demonstrated that echelon stock policies will always dominate installation stock policies and so kanban will be dominated by a echelon stock policy. In later work, Axsäter and Rosling [3, 4] showed that a kanban system can be reproduced within MRP with a suitable selection of parameters (i.e., zero lead time, and the safety stock less one is a multiple of the order quantity). The result is completely general except that MRP is a periodic review system while the kanban system is a continuous review system (although there is no reason why the review in

---

[3]They termed this a hybrid system because the first station in the line requires a pull signal (kanban card) but the other stations in the line do not. So, all operators except the one at the first station behave the same as they would in a conventional push system; they just process jobs when they have them.

9

MRP could not be done continuously). This matters since if kanban is a subset of MRP, then it is clearly dominated by MRP. However, this dominance will be limited to the experimental conditions since the Axsäter and Rosling model does not consider robustness issues, such as those arising from mistakenly releasing too much work into the line.

Other generalizations of kanban include those of Glasserman and Yao [21] and Yao and Cheng [12]. These papers introduce a "block-and-hold-k" card that determines how much work in process should be kept before moving. However, because justification of this approach relies on the existence of separate move cards, the benefits disappear when we consider the equivalent one-card kanban system.

Suri [48] introduced a generalization of kanban called "Paired-cell Overlapping Loops of Cards with Authorization" or POLCA. POLCA is different from kanban in that cards are assigned to pairs of cells rather than particular parts within a cell. The result is a more general construct than kanban that can be applied to make-to-order situations. Because the initial authorization is from an MRP system, but cards are used to limit WIP within pairs of cells, POLCA possesses both push and pull characteristics. One drawback, however, is that one must set a card count for every set of consecutive cells. Nonetheless, Suri describes several examples where POLCA has been used to greatly reduce overall cycle times and WIP.

More recently, Liberopoulos and Dallery [33] (building on [13, 10]) developed another generalized pull system in the form of a make-to-stock system with separate limits on finished goods (S) and WIP (K) that we will refer to as a (K, S) system. This model incorporates classical base stock ($K = \infty$, $S < \infty$), CONWIP ($K = S$), single-stage reserved-stock kanban as in [7] ($K < S$), single-stage backordered kanban policy as in [7] and [13] ($K > S$) and several other formulations. They define a "critical WIP," K_c, as a minimum WIP level that supplies enough throughput from the production system so that for any $K \geq K_c$, the optimal base stock level is equal to the base stock level that would occur with $K = \infty$, noted as S_∞. They then conjecture that the policy parameters that minimize total WIP and finished goods (with identical carrying costs) are K_c and S_∞. If this conjecture is true, then one can easily improve on a CONWIP make-to-stock policy by providing separate limits on WIP and finished goods. A trivial case is that of a single server production-inventory system, where the optimal values are clearly $K_c = 1$ and S set to whatever value is required to maintain the fill rate.

4 Meanwhile, In the Real World . . .

The research literature gave a clear definition of kanban and a detailed summary of its benefits. It also showed that kanban could be generalized in a wide variety of practical ways. However, all these variants of kanban suggested that pull was certainly a more general idea than kanban. So, what exactly is it?

Industry did not have a clear answer. One reason for this was that Ohno and the other early practitioners of JIT discussed pull only in very general, high-level terms.

> *Manufacturers and workplaces can no longer base production on desktop planning alone and then distribute, or push, them onto the market. It has become a matter of course for customers, or users, each with a different value system, to stand in the front line of the marketplace and, so to speak, pull the goods they need, in the amount and at the time they need them.* (Ohno, 1988, p. xiv)

In other words, one should not simply make a large amount of stock and then try to go and sell it. One needs to be aware of the market and pay attention to the customer.

While this might sound like common sense today, it was revolutionary for the mid 20th Century. But it only described the *strategy* of pull, not the *tactics* of pull. Although Ohno did outline the elements needed to make pull to work, namely (1) standard work methods (autonomation) and (2) level production, his writings fell short of providing a working description of pull.

Because of the lack of a more precise definition of pull at the shop floor level, pull quickly became equated with its first manifestation, kanban. Symmetrically, push became nearly synonymous with MRP (see, e.g., [16]). Most magazine articles and trade press books provided vague and self referential definitions, such as "pull is the opposite of push" (see below).

A rare exception was the 1993 book *Just-in-Time Manufacturing: An Introduction* which stated, "... the pull mode of manufacturing only allows parts to be moved from the previous operation to the next when the subsequent operation is ready to process." [11, p. 42]. They identified one of the advantages of pull as follows:

> Pull systems by far outreach the responsiveness of a push system. The responsiveness of the system to changes and problems which arise in upstream processes allows the downstream processes to be shut down. This prevents the accumulation of inventory on the plant floor. [11, p. 43]

In short, a WIP cap prevents a WIP explosion. Sadly, this book was not typical.

In the mid-1990's "pull" shifted in popular usage from being synonymous with kanban to shorthand for make-to-order. A key catalyst for this change was the 1996 book by Womack and Jones, *Lean Thinking*, which was a follow-on to their highly successful *The Machine That Changed the World*. Unfortunately, while this book was widely read and provided many details and case histories related to lean techniques, including pull, it did not provide clean definitions of basic concepts. (In a sense, this book was the opposite of Ohno's, which was short on details, but clear on basic philosophy.) For instance, in the chapter titled, "Pull" Womack and Jones begin,

> Pull *in simplest terms means that no one upstream should produce a good or service until the customer downstream asks for it, but actually following this rule in practice is a bit more complicated.*

While this may look like a reformulation of Ohno's description of pull, it is not. While Ohno was speaking at the strategic level about the basic connection between production and demand, Womack and Jones were talking about the tactics of implementing lean. At the tactical level, waiting for customers to ask for goods or services does indeed cause things to get "a bit more complicated." For instance, consider Ohno's supermarket example. The grocer would have to wait for the customer to ask for his groceries before stocking them!

Lean Thinking was just one of many books and articles that confused the concept of pull with the simpler idea of make-to-order. But because of its popularity, it has served to muddle the understanding of pull significantly. It has become common for managers to state that pull is make-to-order while push is make-to-stock. For example, when Alcoa announced its heavily

publicized Alcoa Production System, one of the basic tenants was: "Produce for use, not for inventory" [1] even though its principle method of production control was kanban.

Unfortunately, such "lean thinking" has become pervasive in industry. Boeing, now a champion of lean and a provider of lean training to its suppliers, gives the following definition for pull on their web site:

> "Pull Production is the opposite of push. It means products are made only when the customer has requested or 'pulled' it, and not before." [5]

Using this definition, a make-to-order MRP system would be an ideal pull system, which clearly contradicts the historical intent of pull.

Despite the confusion over tactical pull, companies implementing lean did appreciate Ohno's insight that leveling production was key to "strategic" pull and that the way to accomplish this was to use a "takt time" or "takt-paced production." For example, Boeing defined "takt-paced production" as [6]:

> Takt-Paced Production describes the rate of assembly in a factory. Lean does not mean doing things faster; it means doing things at the right pace. Essentially, the customer's rate of demand establishes the pace, or takt time. So, rather than simply maximizing the rate of work, lean sets the pace in the factory, ensuring that the customer's needs are met on time.

Taken at face value, it would appear that demand must be extremely regular because otherwise following customer demands would be hugely inefficient. In reality, however, setting a pace (instead of chasing demand) is exactly what Boeing (and Toyota) do to *smooth* the demand that is seen by the plant. This means that they set the takt time based on a current backlog of orders and then adjust it from time to time. Since most releases can be connected to a customer order, these systems are, in an overall sense, make-to-order.

In an analytical sense, production smoothing is really a simple matter of buffering the production line from demand variability. This buffering is done with either a time backlog or inventory. If demand temporarily increases, orders are backlogged. If orders are needed later, the line will build up some inventory. Furthermore, while the takt time drives final assembly, component parts must be available in order for them to be "pulled" from internal fabrication centers or outside suppliers. Clearly, any component with a lead time longer than the time between start of the final assembly and when the unit is needed must be made-to-stock.

If actual demand varies enough (e.g., due to seasonality or random surges) the order backlog may occasionally run dry. When this occurs the firm must adjust the takt time and/or release some jobs without explicit customer orders. Hence, it is possible for a takt time based system to pre-release jobs in make-to-forecast mode.

The bottom line is that although the practitioner literature has commonly defined pull to be make-to-order, this only applies at the strategic level. At the tactical level, the systems actually used to implement pull can be make-to-stock, or even make-to-forecast. Given this, it is no wonder that many practitioners have found the literature on lean confusing and difficult to implement.

5 The Essence of Pull

To remove this confusion, we need a definition of pull that captures what is essential for obtaining its benefits, but does not overly restrict implementation (e.g., by equating it with a specific system such as kanban).

As we noted above, the term "pull" can be applied at the both strategic and the tactical levels. Establishing a takt time to set the output of the plant to be equal to demand is a way to establish strategic (or market) pull. This was an important part of Ohno's original vision (around 1950) and was adopted (or, more likely, independently discovered) by those who developed computer based push systems. For instance, in 1970 Wight described very similar logic (i.e., "Input/Output Control") to be used with MRP [55]. Hence, strategic pull can be implemented even in, what have been traditionally called, push systems.

Therefore, the key question we must answer to fully define pull is what constitutes tactical pull, as characterized by kanban. To be meaningful, we must define pull so that it includes those systems that people tend to think of as pull (e.g., kanban) but not those that have been considered as push (e.g., MRP). Moreover, if pull is somehow intrinsically better, our definition should provide a means to determine if a system belongs to a class of systems that dominate in performance. Finally, although admittedly less important, such a definition should settle long standing debates among academics as to whether various systems (e.g., basestock models and other make-to-stock variations) are fundamentally pull or push systems.

5.1 The Fundamental Difference Between Push and Pull

Given common usage, we begin with the assumption that kanban is a pull system, while MRP is a push system. To define (tactical) pull, we need to characterize what is fundamentally different about these systems.

Axsäter and Rosling [3] show that MRP is more general than an installation stock (Q, r)-policy so that any (Q, r)-policy can be replaced by an MRP system. This implies that MRP should also dominate a kanban policy since it is simply an installation stock policy with an additional limit on the number of outstanding orders [2]. However, the additional constraint appears to be the key to the effectiveness of the pull system. The fact that kanban does not continue adding orders to the system beyond a certain point puts a natural limit on WIP. Thus, no matter how wrong the forecast or how great the demand, the system cannot be overwhelmed beyond its capacity. Since the output of a production line is an increasing but bounded function of WIP (bounded by the bottleneck rate), while flow time begins to grow almost linearly beyond a certain point, it is pointless to add more work to a system that is already "saturated" [44]. Findings by Veatch and Wein [52] also point to the beneficial effects of having a WIP limit and we discuss these at length in *Factory Physics* (Chapter 10).

Based on this observation we propose the following definition:

Definition (Pull and Push): A *pull production system* is one that explicitly limits the amount of work in process that can be in the system. By default, this implies that a *push production system* is one that has no explicit limit on the amount of work in process that can be in the system.

If we are interested only in the *essence* of push or pull, it is quite easy to construct a mathematical model that is either *purely push* or *purely pull*. For instance, a closed queueing

	Make-to-Forecast	Make-to-Order	Make-to-Stock
Push	MRP with forecast	MRP with firm orders	(Q, r) with pull from FGI
Pull	kanban with takt time & forecast	kanban with takt time & orders	kanban with pull from FGI

Table 1: Examples of Push and Pull.

network, with a rigidly set limit on the number of entities, is a pure pull system, while an open queueing system (e.g., a GI/G/1 queue) is an pure push system. The former has a clear limit on WIP while the latter does not.

But in the real world there are no pure push or pure pull systems. For example, while a kanban system establishes a clear limit on WIP via the production cards, there are almost always circumstances under which this limit will be overridden (e.g., a downstream non-bottleneck machine goes down). Conversely, while MRP does not establish a limit on WIP, there is almost certainly some level of WIP that will cause management to ignore the planned order releases or revise the master production schedule in order to prevent further WIP growth. Indeed, there is presumably some limit on WIP for every system (e.g., an amount equivalent to 1000 years of demand). The distinction, however, is that the WIP limit in practical pull systems is explicitly stated and is generally small. Any WIP limit in a practical push system is implicit, large and usually comes into play too late (i.e., after WIP is out of control).

Hence, our definition gives a black and white distinction of push and pull among mathematical models. But the real world, as is generally the case, is a matter of shades of gray. The extent to which a system will obtain the benefits of pull depends on how sharply the up-front WIP limit is imposed.

5.2 Make-to-Stock and Make-to-Order

Note that our definition of pull does not involve the concepts of make-to-order (MTO), make-to-stock (MTS) or make-to-forecast (MTF). Indeed, as we illustrate in Table 1, the push/pull distinction is orthogonal to the MTO/MTS/MTF decision. These examples show that all combinations are possible. Hence, while MTO may indeed be superior in many environments to MTS or MTF, this distinction does not capture the essence of why tactical pull systems improve production efficiency. In the end, it is the explicit limit on WIP that defines a tactical pull system and makes it work.

5.3 Push or Pull

The good news inherent in our definition of pull is that it implies that pull can be implemented in a variety of ways. Kanban is certainly one way to limit WIP. But there are others. To illustrate this we consider some of the most common systems found in industry and the literature and classify them as either push or pull.

1. **MRP** is a **push** system because releases are made according to a master production schedule without regard to system status. Hence, no a priori WIP limit exists.

14

2. **Classic kanban** is a **pull** system. The number of kanban cards establishes a fixed limit on WIP.

3. **Classic Base Stock System** is, somewhat surprisingly, a **push** system because there is no limit on the amount of work in process in the system. This is because backorders can increase beyond the basestock level.

4. **Installation stock** (Q, r) is also a **push** system as are echelon stock (Q, r) systems because neither imposes a limit on the number of orders in the system.

5. **CONWIP** is a **pull** system because it limits WIP via cards similar to kanban. An important difference from kanban from an implementation standpoint is that the cards are line specific rather than part number specific. However, from a push/pull perspective, CONWIP cards limit WIP in the same manner as kanban cards.

6. (K, S) **systems** (proposed by Liberopoulos and Dallery) are **pull** systems if $K < \infty$ and are push systems otherwise.

7. **POLKA systems** proposed by Suri is a **pull** system because, like kanban and CONWIP, WIP is limited by cards.

8. **PAC systems** proposed by Buzacott and Shanthikumar is a **pull** system when the number of process tags (which serve to limit WIP) is less than infinity.

9. **MRP with a WIP constraint** (as suggested by Axsäter and Rosling) is a **pull** system.

6   Other Issues

The confusion surrounding pull has led to confusion around other issues. These include the "push/pull interface" and the essence of what defines lean production.

6.1   The Inventory/Order Interface

If we accept the above definition of push and pull systems, the "push/pull interface" described by Lee and Billington [32] and by ourselves in [18] is actually a misnomer. What we and Lee and Billington were pointing out was the fact that virtually all production systems include make-to-stock and make-to-order segments. For instance, consider the stylized McDonald's system shown in Figure 3. During rush hour, when there are specified targets for the warming table, production up to the warming table is make-to-stock; production beyond this point (i.e., bagging and checkout) is make-to-order.

Because the warming table generally establishes a WIP cap on the front end of the line but there is no firm limit on WIP in the back end of the line, we defined these segments as pull and push and called the interface (warming table) the push/pull interface. But, since theoretically one could envision systems like this with no WIP limit on the front end and/or a finite limit on WIP in the back end, this is not a proper definition. A more accurate term would be the *inventory/order (I/O) interface*, since it represents the point in the production process where the stimulus for work movement shifts from make-to-stock to make-to-order.
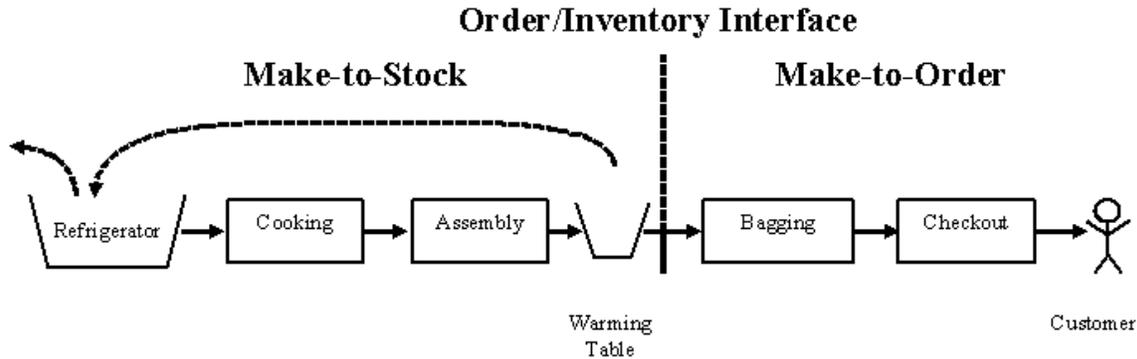
Figure 3: The inventory/order interface.

This correction in terminology does not change the fundamental point that we and Lee and Billington were making, which is that these interfaces exist in all production systems and that positioning them is a management decision. Locating the inventory/order interface at raw materials produces a conventional make-to-order system, while locating it at finished goods produces a classic make-to-stock system. Locating it at an intermediate point (e.g., by stocking components or subassemblies) leads to an assemble-to-order system, which can be used if the firm wants to be able to quote customer lead times that are shorter than the overall manufacturing cycle time.

6.2   Lean Production

Just as kanban was part of the larger construct of JIT, pull is part of the larger construct of lean production. And just as there is considerable confusion about pull, there is confusion about lean. Most sources describe the essence of lean production as waste reduction. But since this doesn't sound very deep, lean books and articles are fond of speaking in terms of elimination of *muda*. But whether one speaks in English or Japanese, the idea of reducing waste in search of efficiency is hardly new. Andrew Carnegie, Fredrick Taylor and Henry Ford were aggressive champions of waste elimination a hundred years ago. So, either lean production is simply a new name for an old concept or it means something more than waste reduction.

Our opinion is that while lean is certainly concerned with driving out waste, it represents a more fundamental framework for enhancing efficiency. Using the language of factory physics we can define lean as:

Definition (Lean): Production of goods or services is *lean* if it is accomplished with minimal buffering costs.

The first source of excess buffering is **obvious waste**. Such waste includes operations that are not needed, excessive setup times, unreliable machines that can be made more reliable, rework that can be eliminated, etc. Indeed, to the dismay of some lean practitioners, Shingo once said, "Eliminate waste is a non-sensical slogan" for this very reason. We take it as a given

16

that the elimination of "obvious waste" is obvious. Unfortunately, this tends to be the focus and limit of most lean implementations that never get to the real "meat" of the problem.

The less obvious source of buffering costs is **variability**. This can take on many forms, including variability in process times, delivery times, yield rates, staffing levels, demand rates, etc. Anything in the system that is not absolutely regular and predictable exhibits variability. The causes of variability can be classified into *internal* factors, such as setups, downtime (scheduled and unscheduled), operator induced fluctuations in production rates, yield loss, rework, engineering change orders and many others, and *external* factors, including irregular demand, product variety (to meet market needs), customer change orders, etc. External variability is often the consequence of a firm's business strategy, such as offering high levels of product variety to achieve a competitive advantage. In such instances, a certain amount of variability is simply a cost of doing business. But external variability is not always justified by strategic considerations. For example, offering discounts, pushing to make end-of-quarter sales targets, interrupting operations to satisfy a specific customer, and so on, may not generate sufficient revenue to justify their cost.

Regardless of its source, *all variability in a production system will be buffered.* A fundamental principle of factory physics is that there are three types of **variability buffer**: **inventory**, **capacity** and **time**. For example, safety stocks represent inventory buffers against variability in demand and/or production. Excess capacity can also provide protection (i.e., a capacity buffer) against fluctuations in demand and/or production. Finally, safety lead times provide a time buffer against production variability. While the exact mix of buffers is a management decision, the decision of whether or not to buffer variability is not. If variability exists, it will be buffered somehow.

This framework suggests a few simple steps for lean implementation:

1. **Eliminate Obvious Waste:** Unnecessary moves (e.g., into and out of a warehouse), mistakes that require an operation to be repeated, poor layout that leads to excessive material handling, and thousands of other examples of poor manufacturing practice represent obvious waste. Most lean sources focus on these and they are clearly important. But, the goal of eliminating direct waste is as old as the factory system itself. While this needs to be done more thoroughly than ever to be competitive in a global economy, the principle is the same as it has always been.

2. **Swap Buffers:** Inventory buffers are "evil" because they hide problems. However, if we simply reduce the inventory buffer without deliberately increasing another buffer, the *default* buffer will be first, time (i.e., late orders will cause poor customer service), and then capacity (i.e., customers cancel orders which reduces equipment utilization). Since both of these are bad, it may make sense to deliberately increase the capacity buffer (by adding capacity, not reducing demand). Indeed, one of the most revolutionary, and overlooked, steps taken by Toyota was a conscious shift from inventory buffering to capacity buffering. At a time when automotive plants generally ran three shifts a day, Toyota went to a two shift schedule, with ten hour shifts separated by two-hour preventive maintenance (PM) periods. These PM periods served as capacity buffers to allow shifts to make up any shortfalls on their production quotas. With these capacity buffers as backup, Toyota could afford to run much leaner with respect to inventory. Furthermore, there has been much made of the use of "problem solvers" on the line at Toyota [41] that are available

17

whenever an operator has a problem. This additional staff represents an even larger capacity buffer that was not available in the original system. Increasing capacity first and then reducing inventory using a pull system allows one to reduce cycle times without losing throughput or hurting customer service. Because cycle times are reduced, we can more easily determine the root cause of variability problems. At this point we can begin making real improvements by reducing variability.

3. **Reduce Variability:** Because variability necessitates buffering it is a fundamental source of waste. But it is an indirect source of waste and is therefore often overlooked. Managers may be very aware of the obvious waste associated with machine failures (i.e., the lost capacity), but they may not appreciate the extent to which the variability caused by these failures results in high WIP, lost throughput or long cycle times (i.e., increased inventory, capacity or time buffers). Because of this, we feel that variability reduction is close to the core of lean. Indeed, with its emphasis on production smoothing, quality improvement, setup time reduction, total preventive maintenance and many other practices, it is clear that Toyota appreciated the key role of variability reduction in JIT right from the start. As Inman [25] elegantly put it, "inventory is the *flower* of all evil" and *variability* is its root. Because it is a variability reduction method, Six Sigma has a natural connection to lean in the same way TQM had a central role in JIT. However, it should be clear that Six Sigma is a methodology for variability reduction, not a general strategy for improvement (e.g., Six Sigma does not address obvious waste). So once the reduction in WIP by using the pull system has reduced cycle times, the plant can begin to identify and eliminate many sources of variability.

4. **Continual Improvement:** As variability is reduced, we can reduce the capacity buffer and keep the inventory buffer low. As Toyota did this, they were able to reduce capacity buffers to the point of running their plants close to capacity while keeping cycle time and inventory low. The result was improved productivity that could be sustained. Nonetheless, regardless of how diligently management pursued variability reduction, variability will always be part of production systems. So, the decision of how to buffer the variability will require continual attention. As the system changes (new products are introduced, processes are updated, etc.) variability will increase. If approached passively, buffers will simply arise. For instance, process variability in a production system will cause inventory to increase, throughput to fall, customer service to decline, or a combination of these and other consequences. But buffering need not be done passively; management can, and should, choose the mix of buffers it wishes to use.

There are other paths to becoming lean other than the one Toyota took. For instance, one of us has been working with Moog, Inc., a producer of precision servo valves. Their approach was quite different from Toyota's but consistent with the basic principles outlined above. Moog's basic problem was that lead times were too long for the changing market, costs were up, and customer service was down. There was a real danger of losing customers. To address these issues we developed and implemented the following strategy:

1. **Eliminate obvious waste.** Using lean methods such as Value Stream Mapping and 5S look for and eliminate easy problems. We save the hard problems for later.

2. **Increase the inventory buffer to insulate problems in fabrication.** Essentially we isolated fabrication from the sub-assembly and final assembly areas by establishing kanban stores for most of the components. To set these inventory levels we used more sophisticated inventory models than those proposed by lean texts. In this way, the lead time to the customer was cut from 16-20 weeks to less than 4 weeks.

3. **Reduce variability in sub-assembly and final assembly.** This was done by streamlining the flow and establishing a CONWIP system. Lead time came down from 23 days to 6 days while service went from less than 50% to over 95%.

4. **Reduce inventory buffers.** Because of the smoother flow, inventory could be reduced.

5. **Address problems in fabrication.** Now that customer service has been restored, the focus has shifted to the more difficult problems in fabrication. Issues such as setup reduction and machine maintenance are now being addressed with no disruption of plant output.

The result has been much greater responsiveness to the customer with improved service. The improved flow also resulted in an unexpected (for management) benefit—a greater than 5% improvement in productivity.

In this section we have defined lean in terms of the *cost* of buffering. This is critical, since the fundamental objective is not to reduce inventory, increase utilization, shorten lead time, or even the "perfect value stream" [59]. The objective is to make money. Since the cost of the various options for reducing or buffering variability will vary between environments, no one solution is right for all systems. The real challenge of lean is to find the mix of policies that is best for each particular environment.

7 Conclusions

Pull systems have been a part of the manufacturing lexicon for a quarter of a century. During this time, academics have incrementally described the mechanics of pull systems, characterized their benefits, extended kanban into an array of variants and offered unifying frameworks for integrating and comparing the range of possible pull systems. This work has steadily enhanced our understanding of how and why pull works. Industry practitioners, in spite of their many successes at implementing pull, seem to have focused on describing pull in ever simpler terms. Starting with a tendency to confuse the general concept of pull with the specific practice of kanban, the popular literature has evolved into a simplistic view of pull as nothing more than making products to customer orders. We think this is a serious mistake that compromises the ability of firms to construct effective pull systems for their environments.

The magic of pull is the maintenance of a WIP cap. While pull systems can take on many forms to suit different sets of circumstances, all of them have in common that releases are regulated according to internal system status in a manner that prevents inventory from growing beyond a specified limit. Because this is fundamental to achieving the benefits of pull, it is essential that pull be defined in terms of the WIP cap. Decisions of whether to make-to-order or make-to-stock and how to rely on forecasting are certainly important, but are

orthogonal to the push versus pull decision. The WIP cap definition of pull given in this paper will prevent confusion of these important, but separate, issues.

The sin of oversimplification has been carried over from pull to lean. Practitioner literature that describes lean production only in terms of reducing waste does the profession a disservice by overlooking important but indirect sources of waste. Moreover, there are many causes of variablity that result in a buffer that are not "waste." For instance, adding variety to a product mix to accommodate customer's demands is hardly considered waste but it will reduce efficiency and increase buffers. Lean is better defined as "best buffer" production than "low waste" or even "low buffer" production. Thinking in terms of the more fundamental concepts of variability and buffering encourages more comprehensive consideration of efficiency improvement alternatives.

While there have been plenty of cases where academics have confused or missed the point of a practical problem (see e.g., [14] for a criticism of scheduling research), we feel the pull arena is one where the academic community has been a voice of reason. The question for us now: *is anyone listening?*

8   Acknowledgments

Bibliography

[1] "A Framework for Success: The Alcoa Business System," *Alcoa Update,* at www.alcoa.com, July 2001.

[2] Axsäter, S., and K. Rosling, "Notes: Installation vs. Echelon Stock Policies for Multilevel Inventory Control," *Management Science* **39**:10, 1274–1280, 1993.

[3] Axsäter, S., and K. Rosling, "Multi-level Production-Inventory Control: Material Requirements Planning or Reorder Point Policies," *European Journal of Operational Research* **75**, 405–412, 1994.

[4] Axsäter, S., and K. Rosling, "Ranking of Generalised Multi-Stage Kanban Policies," *European Journal of Operational Research* **113** 560–567, 1999.

[5] The Boeing Company, http://www.boeing.com/commercial/initiatives/lean/key.html, 2002.

[6] The Boeing Company, http://www.boeing.com/commercial/initiatives/lean/movingline.html, 2002.

[7] Buzacott, J.A., "Queueing Models of Kanban and MRP Controlled Production Systems," *Engineering Cost and Production Economics* **17**, 3–20, 1989.

[8] Buzacott, J.A. and J.G. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[9] Boudette, N., "Europe's SAP Scrambles to Stem Big Glitches—Software Giant to Tighten Its Watch After Snafus at Whirlpool, Hershey," *The Wall Street Journal*, Nov 4, 1999.

[10] Chaouiya, C., G. Liberopoulos and Y. Dallery, "The Extended Kanban Control System for Production Coordination of Assembly Manufacturing Systems," *IIE Transactions* **32**, 999–1012, 2000.

[11] Cheng, T.C.E., and S. Podolsky, *Just-in-Time Manufacturing: An Introduction*, Chapman & Hall, London, 1993.

[12] Cheng, D.W., and D.D. Yao, "Tandem Queues With General Blocking: A Unified Model and Comparison Results," *Discrete Event Dynamic Systems: Theory and Applications*, **2**, 207–234, 1993.

[13] Dallery, Y., and G. Liberopoulos, "Extended Kanban Control System: Combining Kanban and base stock," *IIE Transactions* **32**, 369–386, 2000.

[14] Dudek, R.A., S.S. Panwalkar, and M.L. Smith. 1992. "The Lessons of Flowshop Scheduling Research," *Operations Research* **40**(1): 7-13.

[15] Hall, R.W., *Driving the Productivity Machine: Production and Control in Japan*, American Production and Inventory Control Society, Falls Church, VA, 1981.

[16] Hall, R.W., *Zero Inventories*, Dow Jones-Irwin, Homewood, IL, 1983.

[17] *JIT Factory Revolution: A Pictorial Guide to Factory Design of the Future*, H. Hirano, Editor-in-Chief, JIT Management Laboratory Company, Ltd. (Tokyo), J.T. Black, Editor, English Edition, Productivity Press, Cambridge, Mass., 1987 (Japanese) 1988 (English).

[18] Hopp, W.J. and M.L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, McGraw-Hill, New York, 2001.

[19] Fox, R.E., "Keys to Successful Materials Management Systems: A Contrast Between Japan, Europe and the U.S.," *23rd Annual Conference Proceedings*, APICS, 440–444, 1980.

[20] George, M.L., *Lean Six Sigma: Combining Six Sigma Quality with Lean Speed*, McGraw-Hill, New York, 2002.

[21] Glasserman, P. and D.D. Yao, *Monotone Structure in Discrete-event Systems*, John Wiley, New York, 1994.

[22] Huang, P.Y., L.P. Rees, and B.W. Taylor III, "A Simulation Analysis of the Japanese JIT Technique (with kanban) for Multiline, Multistage Production Systems, *Decision Systems* **14**, 326–344, 1983.

[23] Huang, C.C., and A. Kusiak, "Overview Of Kanban Systems," *Int. J. Computer Integrated Manufacturing* **9**:3 169–189, 1996.

[24] "Competition in Manufacturing leads to MRP II," *IE* **23** July, 10–13, 1991.

[25] Inman, R.R., "Inventory is the <u>Flower</u> of All Evil," *Production and Inventory Management Journal.*, 1993, 4th Qtr., 41–45.

[26] Karmarkar, U.S., "Kanban Systems," Working Paper Series No. QM8612: Center for Manuf. and Oper. Mngmt., The Graduate School of Mngmt., The Univ. of Rochester, Jun. (1986).

[27] Karmarkar, U.S., "Push, Pull and Hybrid Control Schemes," *Tijdschrift voor Economie en Management* **26**, 345–363, 1991.

[28] Karmarkar, U.S. and S. Kekre, "Batching policy in kanban systems." *Journal of Manufacturing Systems*, **8**:4, 317–328, 1989.

[29] Kimball, G.E., "General Principles of Inventory Control," *J. Manufacturing and Opns. Mgmt.* 1 (1988):1, 119-130.

[30] Kimura, O. and H. Terada, "Design and Analysis of Pull System, A Method of Multi-Stage Production Control," *Intl. J. of Prod. Res.* **19**:3, 241–253, 1981.

[31] Krajewski, L.J., B.E. King, L.P. Ritzman, and D.S. Wong, "Kanban, MRP, and Shaping the Manufacturing Environment," *Mgmt. Sci.* **33**:1 (1987), 39-57.

[32] Lee, H.L., and C. Billington, "The Evolution of Supply-Chain-Management Models and Practice at Hewlett-Packard," *Interfaces* **25**(5), 42–63, 1995.

[33] Liberopoulos, G., and Y. Dallery, "Base Stock Versus WIP Cap in Single-Stage Make-to-Stock Production-Inventory Systems," *IIE Transactions* **34**, 627–636, 2002.

[34] Meyers, F.S., "Japan's Henry Ford," *Scientific American* 262 (1990):5, 98.

[35] Monden, Yasuhiro., *Toyota Production System: Practical Approach to Management*, Industrial Engineering and Management Press, Norcross, GA, 1983.

[36] Ohno, T. *Toyota Production System: Beyond Large Scale Production*, Productivity Press, Cambridge, Mass, 1988.

[37] Orlicky, J., "Material Requirements Planning: The New Way of Life in Production and Inventory Management", McGraw Hill, New York, 1975.

[38] Schonberger, R.J. 1982. *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity*, New York: The Free Press.

[39] Shingo, Shigeo, *A Revolution in Manufacturing: The SMED System*, Productivity Press, Cambridge, MA (1985)

[40] Simpson, K.F., "In-Process Inventories," *Ops. Res.* **6**:6 (1958), 863-873.

[41] Spear, S. and H.K. Bowen, "Decoding the DNA of the Toyota Production System," *Harvard Business Review*, September–October, 1999.

[42] Spearman, M.L., D.L. Woodruff and W.J. Hopp, "CONWIP: A Pull Alternative to Kanban," *International Journal of Production Research* **28**:5 (1990), 879–894.

[43] Spearman, M.L. and M.A. Zazanis, "Push and Pull Production Systems: Issues and Comparisons," *Operations Research*, **40** (1992), 521–532.

[44] Spearman, M.L., "An Analytic Congestion Model for Closed Production Systems," *Management Science*, 37:8 (1991), 1015-1029.

[45] Spearman, M.L., "Customer Service in Pull Production Systems," *Operations Research*, **40** (1992), 948–958.

[46] Spearman, M.L. and M.A. Zazanis, "Push and Pull Production Systems: Issues and Comparisons," *Operations Research*, **40** (1992), 521–532.

[47] Sugimori, Y., K. Kusunoki, F. Cho and S. Uchikawa, "Toyota Production System and Kanban System: Materialization of Just-In-Time and Respect-For-Human System," *Internat. J. Production Res.*, 15 (1977):6, 553-564.

[48] Suri, R. 1998. *Quick Response Manufacturing: A Companywide Approach to Reducing Leadtimes.* Portland, OR: Productivity Press.

[49] Tayur, S.R., "Structural Properties and a Heuristic for Kanban-controlled Serial Lines," *Management Science* **39**, 1347–1368, 1993.

[50] Turbide, D.A. "Flow Manufacturing—A Strategy White Paper," at www.powercerv.com/downloads/flow.pdf, 1999.

[51] Uzsoy, R., and L.A. Martin-Vega, "Modeling Kanban Based and Demand-pull Systems: A Survey and Critique," *Manufacturing Review*, **3**, 155–160, 1990.

[52] Veatch, M.H., and L.M. Wein, "Optimal Control of a Two-station Tandem Production/Inventory System," *Operations Research*, **42**, 337–350, 1994.

[53] White, R.E., and V. Prybutok, "The relationship between JIT practices and type of production system," *Omega* **29**, 113–124, 2001.

[54] Whitt, W., "Open and Closed Models for Networks of Queues," *AT&T Bell Lab. Tech. J.* **63**(1984), 1911-1978.

[55] Wight, O., "Input/Output Control a Real Handle on Lead Time," *Prod. & Inv. Mgmt. J.* **11**:3, 9–31, 1970.

[56] Wight, O., *MRP II: Unlocking America's Productivity Potential.* Boston: CBI Publishing, 1981.

[57] Womack, J.P., D.T. Jones and D. Roos, *The Machine That Changed the World: The Story of Lean Production*, HarperCollins Publishers, New York, 1990.

[58] Womack, J.P. and D.T. Jones, *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*, Simon & Schuster, New York, 1996.

[59] Womack, J.P. "Jim Womack on how lean compares with Six Sigma, Re-engineering, TOC, TPM, etc., etc.," *Lean Enterprise Institute Newsletter*, July 14, 2003.

[60] Zais, A., "IBM Reigns in Dynamic MRP II Marketplace," *Computerworld* Jan. 27 (1986), 37.

[61] Zipkin, P.H. 1991. "Does Manufacturing Need a JIT Revolution?" *Harvard Business Review,* January-February, 40-50.