Lean Operations and the Toyota Production System

Toyota is frequently associated with high quality as well as overall operational excellence, and, as we will discuss in this chapter, there are good reasons for this association - Toyota has enjoyed decades of economic success while changing the history of operations management.

- Various elements of the company's famous Toyota Production System (TPS) are covered throughout this book, but in this chapter we will review and summarize the components of TPS, as well as a few that have not been discussed in earlier chapters.

- We also will illustrate how the various elements of TPS are intertwined, thereby making it difficult to adapt some elements while not adapting others.

Readers who want to learn more about TPS are referred to excellent readings, such as Fujimoto (1999) or Ohno (1976), from which many of the following definitions are taken.

As we will discuss, one of the key objectives of TPS is the elimination of "waste" from processes, such as idle time, unnecessary inventory, defects, etc. As a result, people often refer to (parts of) TPS as "lean operations". The expression "lean operations" has been especially popular in service industries.

Section 8.1 The History of Toyota

To appreciate the elegance and success of the Toyota Production System, it his helpful to go back in time and compare the history of the Toyota Motor Company with the history of the Ford Motor Corporation.

Inspired by moving conveyor belts at slaughter houses, Ford pioneered the use of the assembly line in automobile production. The well known Model T was the first mass produced vehicle that was put together on an assembly line using interchangeable parts. Working with interchangeable parts allowed Ford to standardize assembly tasks, which had two important benefits. First, it dramatically reduced variability, and thereby increased quality. Second, it streamlined the production process, thereby making both manual and automated assembly tasks faster.

With the luxury of hind sight, it is fair to say that Ford's focus was on running his automotive production process with the goal of utilizing his expensive production equipment as much as possible, thereby allowing him to crunch out the maximum number of vehicles. Ford soon reached an unmatched production scale – in the early days of the Model T, 9 out of 10 automotive vehicles in the world were produced by Ford! Benefiting from his scale economies, Ford drove the price of a Model T down to a 2005-inflation- adjusted US$ 3,300. This made the automotive vehicle affordable to the American middle class, an enormous market that was well suited to be served by mass production.

The Toyota Motor Corporation grew out of Toyota Industries, a manufacturer of automated looms, just prior to World War II. Toyota supported the Japanese Army by supplying it with military trucks. Given the shortages of most supplies in Japan at that time, Toyota trucks were equipped with only one head-light and had an extremely simplistic design. As we will see, both the heritage as a loom maker as well as the simplicity of its first vehicle product had consequences for the future development of Toyota.

Following the war, shortages in Japan were even more severe. There existed virtually no domestic market for vehicles and little cash for the acquisition of expensive production equipment. The United States had an active role in the recovery process of Japan and so it is not surprising that the American production system had a strong influence on the young auto maker. Toyota's early vehicles were in part produced using second-hand U.S. equipment and also otherwise had significant resemblances with the U.S. brands of Dodge and Chevrolet.

As inspiring the Western industrial engineering must have been to Toyota, replicating it was out of the question. Mass production, with its emphasis on scale economies and large investments in machinery, did not fit Toyota's environment of a small domestic market and little cash.

Out of this challenging environment of scarcity, Toyota's management created the various elements of a system that we now refer to as the Toyota Production System (TPS). TPS was not invented over-night -it is the outcome of a long evolution that made Toyota the most successful auto maker in the world and the gold standard for operations management.

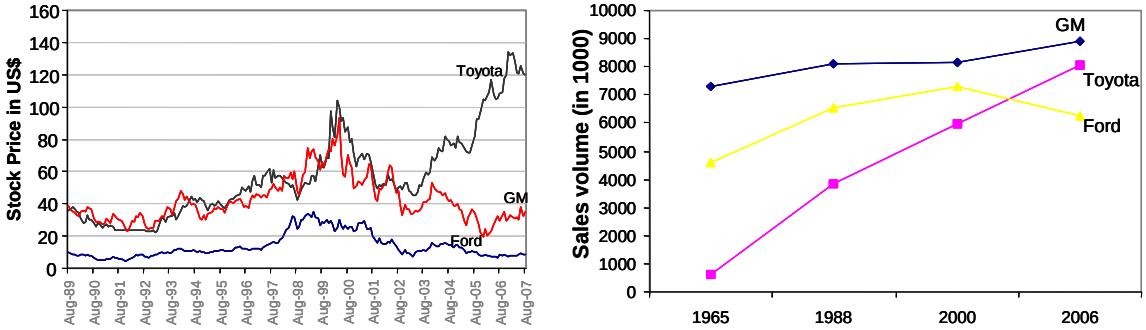


Exhibit STOCK-PRICE: Comparison of Toyota, General Motors, and Ford

As a measure of Toyota's success, consider the following business facts. In the first half of 2007, Toyota became the number one manufacturer of automobiles as measured by sales volume. This ended the over 70 year long leadership of GM. This leadership in sales volume is especially remarkable given the dominating position of Ford and GM just 20 years earlier (see Exhibit STOCK-PRICE, right).

However, the financial success of Toyota is even more impressive. Toyota had 2006 profits of $13.6 Billion and employed over 300,000 people worldwide. At the same time, the company created substantial economic value as measured by the returns to its shareholders. Exhibit STOCK-PRICE (left) compares the trajectory of Toyota's total return to shareholders with the ones of GM and Ford. Smart operations does pay-off!

Section 8.2 TPS Framework

While TPS is frequently associated with certain buzzwords, such as JIT, Kanban, and Kaizen, one should not assume that simply implementing any of these concepts would lead to the level of operational excellence at Toyota. TPS is not a set of off-the-shelf solutions for various operational problems, but instead a complex configuration of various routines ranging from human resource management to the management of production processes.

Exhibit FRAMEWORK summarizes the architecture of TPS. At the top, we have the principle of waste reduction. Below, we have a set of methods that help support the goal of waste reduction. These methods can be grouped into JIT methods (JIT stands for Just-in-Time) and quality improvement methods. There exist strong interdependencies among the various methods. We will discuss some of these interdependencies throughout this chapter, especially the interaction between JIT and quality.

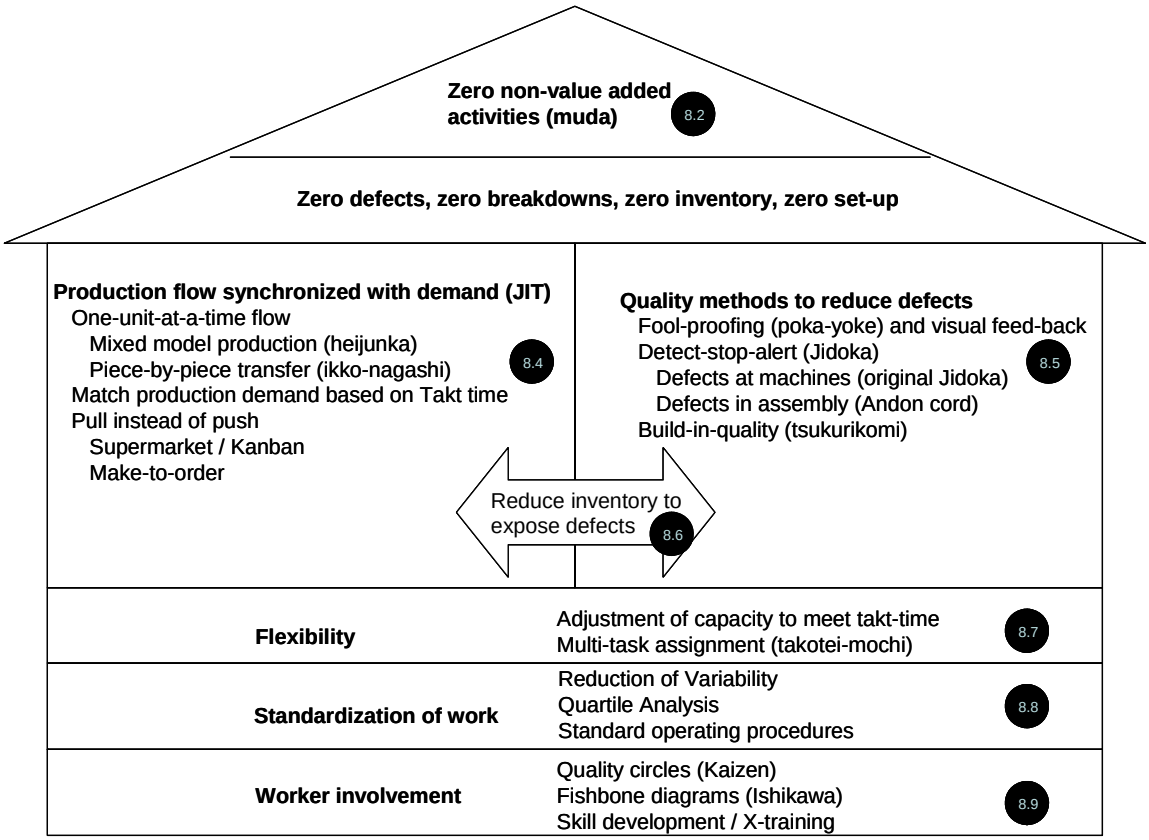


Exhibit TPS-FRAMEWORK: The basic architecture of TPS (the numbers in the black circles correspond to the related section numbers of this chapter)

Collectively, these methods help the organization to attack the various sources of waste that we will define in the next section. Among them, are overproduction, waiting, transport, over-processing, and inventory all of which reflect a mismatch between supply and demand. So the first set of methods that we will discuss (Section 8.4) relate to synchronizing the production flow with demand. Output should be produced exactly when the customer wants it and in the quantity demanded. In other words, it should be produced Just-in-Time.

If we want to obtain a flow rate of the process that reliably matches demand while also following the Just-in-time idea, we have to operate a process with no defects and no break-downs. This is a direct consequence of our discussion in Chapters 6 and 7 (buffer or suffer): defects create variability and the only way we can obtain our target flow rate in a process with variability is to use buffers.

Toyota's strong emphasis on quality lets the company overcome the buffer-or-suffer tension: by producing with zero defects and zero break-downs, the company neither has to suffer (sacrifice flow rate) nor to buffer (hold inventory). For this reason, and the fact that defects are associated with the waste of rework, quality management is the second pillar around which TPS is built.

Both JIT and quality management require some foundational methods, such as the standardization of work (which eliminates variability), the flexibility to scale up and down process capacity in response to fluctuations in demand, and a set of human resource management practices.

Section 8.3 The Seven Sources of Waste

In the late 1980s, a research consortia known as the International Motor Vehicle Program (IMVP) conducted a global benchmarking of automotive plants. The study compared quality and productivity data from plants in Asia, Europe, and North America. The results were a clear indication of how far Toyota already had journeyed in redesigning the historical concept of mass production.

General Motors Framingham Assembly Plant Versus Toyota Takaoka Assembly Plant, 1986

	GM Framingham	Toyota Takaoka
Gross Assembly Hours per Car	40.7	18
Assembly Defects per 100 Cars	130	45
Assembly Space per Car	8.1	4.8
Inventories of Parts (average)	2 weeks	2 hours

Gross assembly hours per car are calculated by dividing total hours of effort in the plant by the total number of cars produced
Defects per car were estimated from the JD Power Initial Quality Survey for 1987
Assembly Space per Car is square feet per vehicle per year, corrected for vehicle size
Inventories of Parts are a rough average for major parts

Exhibit WOMACK: General Motors Framingham Assembly Plant versus Toyota Takaoka Assembly Plant (based on 1986 benchmarking data from the IMVP Assembly Plant Survey, Source Womack et al)

Consider the data displayed in Exhibit WOMACK. The exhibit compares General Motors Framingham Assembly Plant with Toyota Takaoka Assembly Plant. The Toyota plant was about twice as productive and had three times less defects compared to the GM plant making a comparable vehicle. Moreover, it used its manufacturing space more efficiently and turned its components and parts inventory dramatically faster.

While the data underlying this exhibit is already 20 years old, it is still of high relevance today. First, the IMVP study in many ways was the first true proof of the superiority of TPS. For that reason, it constituted a milestone in the history of industrialization. Second, while all large automotive manufacturers made substantial improvements since the initial data collection, two more recent rounds of benchmarking (see Holweg and Pil) documented that the productivity of Japanese manufacturers has been a moving target. While US and European manufacturers could improve their productivity, the Japanese producers have

continued to improve theirs so that Toyota still enjoys a substantial competitive advantage today.

What accounts for the difference in productivity between the GM and the Toyota plant? Both processes end up with a very comparable car after all. The difference in productivity is accounted for by all the things that the GM did that did not contribute to the production of the vehicle: non-value added activities. TPS postulates the elimination of such non-value added activities, which are also referred to as *Muda*.

There are different types of Muda. According to T. Ohno, one of the thought leaders with respect to TPS, there are seven sources of waste:

1. Overproduction: producing too much, too soon, which leads to additional waste in the forms of material handling, storage, and transportation. The Toyota Production system seeks to produce only what the customer wants and when the customer wants it.

2. Waiting: in the spirit of "matching supply with demand" (see chapter PROCESS VIEW), there exist two types of waiting. In some cases, a resource waits for flow units, leading to idle time at the resource. Utilization measures the amount of waiting of this type – a low utilization indicates the resource is waiting for flow units to work on. In other cases, flow units wait for resources to become available. As a consequence, the flow time is longer than the value-add time. A good measure for this second type of waiting is the percentage of flow time that is value-add time (in the language of Chapter 6, this is the activity time, p, relative to the flow time, $T=T_q+p$). Both types of waiting reflect a poorly balanced process and can be reduced by using the tools outlined in chapter 4.

3. Transport: internal transports, be it carrying around half finished computers, wheeling patients through the hospital, or carrying around folders with insurance claims, correspond to the third source of waste. Processes should be laid out such that the physical lay-out reflects the process flow to minimize the distances flow units most travel through a process.

4. Over-processing: a close analysis of activity times reveals that workers often spend more time on a flow unit than necessary. A worker might excessively polish the surface of a piece of metal he just processed or a doctor might ask a patient the same questions that a nurse has asked five minutes earlier.

5. Inventory: in the spirit of matching supply with demand, any accumulation of inventory has the potential to be wasteful. Inventory is closely related to over-production and often indicates that the JIT methods have not (yet) been implemented correctly. Not only is inventory often non-value adding, it often hides other problems in the process as it leads to long information turnaround

times and eases the pressure to find and to eliminate underlying root causes (see Section 8.6 for more details).

6. Rework: A famous saying in the Toyota Production System and the associated quality movement has been "Do it right the first time". As we have discussed in the previous chapter, rework increases variability and consumes capacity from resources. Not only does rework exist in manufacturing plants, it is also (unfortunately) common in service operations. For example, hospitals all too frequently repeat x-rays because of poor image quality or readmit patients to the intensive care unit.

7. Motion: there are many ways to perform a particular task, such as the tightening of a screw on the assembly line or the movement of a patient from a wheelchair into a hospital bed. But, according to the early pioneers of the industrial revolution, including Frederick Taylor and Frank and Lillian Gilbreth, there is only one "right way". Every task should be carefully analyzed and should be optimized using a set of tools that today is known as ergonomics. To do otherwise is wasteful.

Just as we have seen in the context of line balancing, the objective of waste reduction is to maximize the percentage of time a resource is engaged in value adding activity by reducing the non-value added (wasteful) activities as much as possible.

At this point a clarification of wording is in order. TPS's objective is to achieve zero waste, including zero inventory and zero defects. However, this objective is more an ideological one than it is a numerical one. Consider the objective of zero inventory and recall from Little's Law: Inventory=Flow Rate * Flow time. Thus, unless we are able to produce at the speed of light (flow time equals to zero), the only way to achieve zero inventory is by operating at zero flow rate – arguably, not a desirable outcome. So, of course, Toyota's factories don't operate at zero inventory, but they operate at a low level of inventory and keep on decreasing this low level. The same holds for zero defects. Defects happen in each of Toyota's assembly plants many, many times a shift. But they happen less often than elsewhere and are always thought of as a potential for process improvement.

It is important to emphasize that the concept of waste is not unique to manufacturing. Consider, for example, the day of a nurse in a large hospital. In an ideal world, a nurse is there to care for patients. Independent of managed care, this is both the ambition of the nurse and the desire of the patient. However, if one carefully analyzes the work-day of most nurses, a rather different picture emerges. Most nurses spend less than half of their time helping patients, and waste the other time running around in the hospital, doing paper-work, searching for medical supplies, coordinating with doctors and the hospital administration, etc. (See Tucker 2004 for an excellent description of nursing work from and operations management perspective). This waste is frustrating for the nurse, leads to poor care for the patient, and is expensive for the healthcare provider.

Once we have reduced waste, we can perform the same work, yet at lower costs. In a process that is currently capacity constrained, waste reduction is also a way to increase output (flow rate) and hence revenues. As we have discussed in chapter FINANCE AND OPERATIONS the economic impact these improvements can be dramatic. .

A useful way to analyze and describe the effects of waste is the Overall Equipment Effectiveness (OEE) framework, used by McKinsey and other consulting firms. The objective of the framework is to identify what percentage of a resource's time is true, value-add time, and what percentage is wasted. This provides a good estimate for the potential for process improvement before engaging in waste reduction.

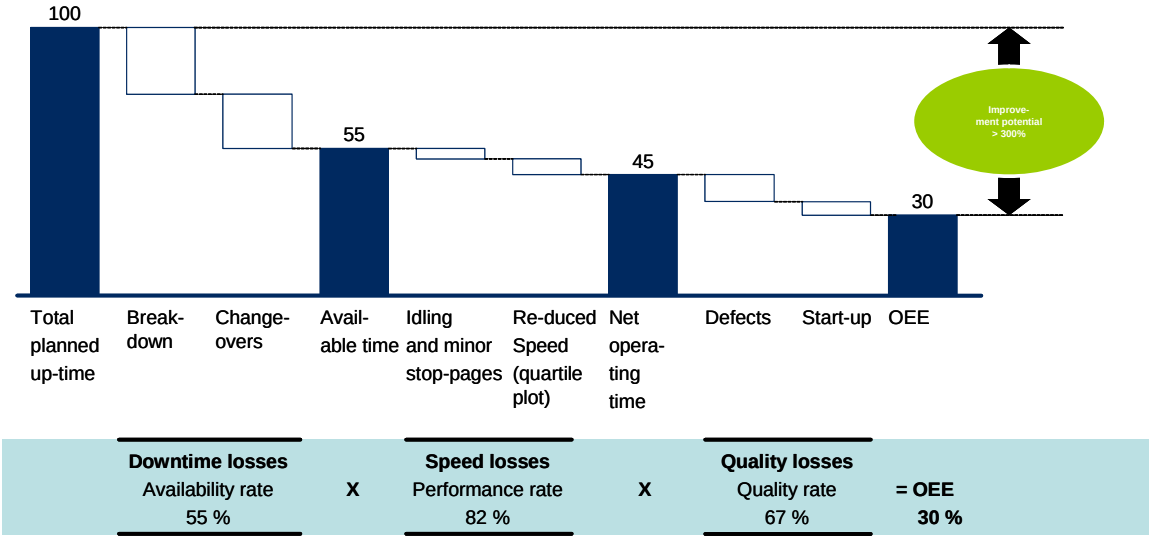


Exhibit OEE: The Overall Equipment Effectiveness Framework

As is illustrated by Exhibit OEE, we start the OEE analysis by documenting the total available time of the resource. From this total time (100%), some time is wasted on machine break-downs (or, in the case of human resources, absenteeism) and set-up times, leading to an available time that is substantially less than the total planned time (in this case, only 55% of the total planned time is available for production). However, not all of the remaining 55% are value add time. Because of poor process balance, the resource is likely to be occasionally idle. Also, the resource might not operate at an optimum speed as the activity time includes some waste and some incidental work that does not add direct customer value. In the case of Exhibit OEE, 82% of the available time is used for operation, which leaves a total of 45% (=55%*82%). If one then factors in a further waste of capacity resulting from defects, rework, and start-ups (67%), we see that only 30% (55%*82%*67%) of the available capacity is used to really add value!

The following two examples illustrate the usefulness of the OEE framework in non-manufacturing settings. They also illustrate that wasting as much as half of the capacity of an expensive resource is much more common than one might expect:

- In the loan underwriting process of a major consumer bank, a recent case study documented that a large fraction of the underwriting capacity is not used productively. Unproductive time included (a) working on loans that are unlikely to be accepted by customers because the bank has already taken too long to get a response back to the customer (b) idle time (c) processing loans that resources preceding underwriting already could have rejected because of an obviously low credit worthiness of the application (d) incidental activities of paper handling (e) attempting to reach customers on the phone but failing to do so. The study estimates that only 40% of the underwriting capacity is used in a value adding way.

- In the operating rooms of a major hospital, the capacity is left unused because of (a) gaps in the schedule (b) procedure cancellation (c) room cleaning time (d) patient preparation time and (e) procedure delays because of the doctor or the anesthesiologist arriving late. After completing waste identification, the hospital concluded that only 60% of its operating room time was used productively. One might argue that patient preparation is a rather necessary and hence value-adding step prior to surgery. Yet, it is not clear that this step has to happen in the operating room. In fact, some hospitals are now using the tools of set-up time reduction discussed in chapter BATCHING and preparing the patient for surgery outside of the operating room so that the change-over from one surgical procedure to another is reduced.

Section 8.4 JIT: Matching Supply with Demand

Just-in-time (JIT) is about matching supply with demand. The goal is to create a supply process that forms a smooth flow with its demand, thereby giving customers exactly what they need, when they need it.

In this section, we discuss three steps towards achieving a JIT process. The three steps build on each other and hence should be taken in the order they are presented. They presume that the process is already in-control (see chapter on SIX SIGMA) using standardized tasks and is able to achieve reliable quality:

1. Achieve a one-unit-at-a-time flow

2. Produce at the rate of customer demand

3. Implement a pull system using Kanban or make-to-order production

One-unit-at-a-time flow

Compare the following two technologies that move people from one level of a building to another: an escalator and an elevator. Most of us associate plenty of waiting with elevators – we wait for the elevator to arrive and we wait stuck between dozens of people as the elevator stops at seemingly every floor. Escalators, in contrast, keep people moving towards their destination, no waiting and no jamming of people.

People waiting for and standing in elevators are like batches in a production setting. Chapter BATCHING already has discussed the concepts of SMED, the reduction of set-up times that makes small production batches economically possible. In TPS production plans are designed to avoid large batches of the same variant. Instead, product variants are mixed together on the assembly line (mixed model production, which is also known as heijunka), as discussed in Chapter BATCHING.

In addition to reducing set-up times, we also should attempt to create a physical lay-out for our resources that closely mirrors the process flow. In other words, two resources that are close to each other in the process flow diagram should also be co-located in physical space. This avoids unnecessary transports and reduces the need to form transport batches. This way flow units can flow one-unit-at-a-time from one resource to the next (ikko-nagashi).

Produce at the rate of customer demand

Once we have created a one-unit-at-a-time flow, we should make sure that our flow rate is in line with demand. Historically, most large-scale operations have operated their processes based on forecasts. Using planning software (often referred to as MRP for materials resource planning and ERP for enterprise resource planning), work schedules were created for the various sub-processes required to create the final product.

Forecasting is a topic for itself (see Chapter NEWSVENDOR), but most forecasts have the negative property of not being right. So at the end of a planning period (e.g. one month), the ERP system would update its next production plan, taking the amount of inventory in the process into account. This way, in the long-run, production more or less matches demand. Yet, in the day-to-day operations, extensive periods of substantial inventories or customer back-orders exist.

TPS aims at reducing finished goods inventory by operating its production process in synchronization with customer orders. This is true for both the overall number of vehicles produced as well as with respect to the mix of vehicles across various models.

We translate customer demand into production rate (Flow rate) using the concept of takt time. Takt time is derived from the German word "Takt", which stands for "tact" or "clock". Just like an orchestra needs to follow a common tact imposed by the conductor, a JIT process should follow the tact imposed by demand. Takt time calculations are identical to what we have seen with demand rate and flow rate calculations in earlier chapters.

Pull systems

The synchronization with the aggregate level of demand through takt time is an important step towards the implementation of JIT. However, inventory not only exists at the finished good level, but also throughout the process (work in process inventory). Some parts of the process are likely to be worker paced with some (hopefully modest) amount of inventory between resources. We now have to design a coordination system that coordinates these resources controlling the amount of inventory in the process. We do this by implementing a pull system.

In a pull system, the resource furthest downstream (i.e. closest to the market) is paced by market demand. In addition to its own production, it also relays the demand information to the next station upstream, thus, ensuring that the upstream resource also is paced by demand. If the last resource assembles two electronics components into a computer, it relays the demand for two such components to the next resource upstream. This way, the external demand is transferred step-by-step through the process, leading to an information flow moving in the opposite direction relative to the physical flow of the flow units.

Such a demand driven pull system is in contrast to a *push system* where flow units are allowed to enter the process independent of the current amount of inventory in process. Especially if the first resources in the process have low levels of utilization – and are thereby likely to flood the downstream with inventory - push systems can lead to substantial inventory in the process.

To implement a pull system, TPS advocates two forms of process control:

- Kanban based pull (also known as fill-up or Super market pull): the upstream replenishes what demand has withdrawn from the downstream.

- Make-to-order refers to the release of work into a system only when a customer order has been received for that unit.

Consider the Kanban system first. *Kanban* refers to a production and inventory control system, in which production instructions and parts delivery instructions are triggered by the consumption of parts at the downstream step (Fujimoto 1999).

In a Kanban system, standardized returnable parts containers circulate between the upstream and the downstream resources. The upsteam resource is authorized to produce a unit when it receives an empty container. In other words, the arrival

of an empty container triggers a production order. The term "kanban" refers to the card which is attached to each container. Consequently, Kanban cards are frequently called work authorization forms.

A simplified description of a Kanban system is provided by Figure KANBAN. A downstream resource (right) consumes some input component that it receives from its upstream resource (left). The downstream resource empties containers of these input components – the downstream resource literally takes the part out of the container for its own use, thereby creating an empty container, which in turn, as already mentioned, triggers a production order for the upstream resource. Thus, the use of Kanban cards between all resources in the process provides an effective and easy to implement mechanism for tying the demand of the process (downstream) with the production of the resources (upstream). They therefore enforce a match between supply and demand.

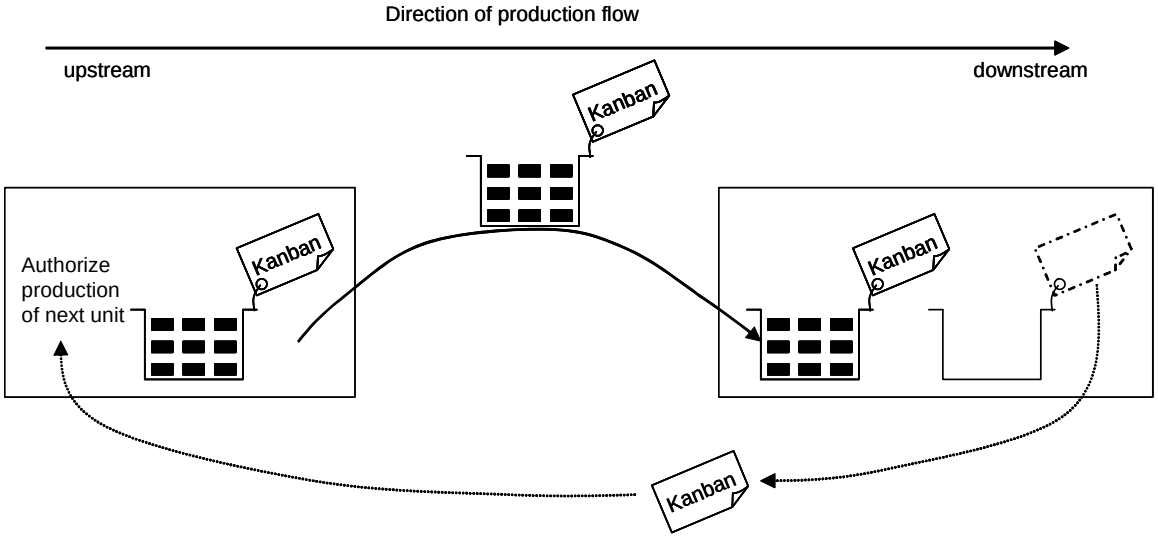


Exhibit KANBAN: The operation of a Kanban system

The main advantage of a Kanban system is that there can never be more inventory between two resources than what has been authorized by the Kanban cards – the upstream resource can only produce when it has an empty container, so production stops when all of the containers are full, thereby limiting the inventory to the number of containers. In contrast, with a push system, the upstream resource continues to produce as long as it has work. For example, suppose the upstream resource is a lathe that produces the legs for a wood chair. With a push system the lathe keeps producing legs as long as it has blocks of wood to work on. With a Kanban system the lathe produces a set of chair legs only if it has an empty Kanban. Hence, with a Kanban system the lathe stops working only when it runs out of Kanbans, whereas with a push system the lathe only stops working when it runs out of raw materials. The distinction can lead to very different

behavior. In a push system inventory can simply "happen" to management because there is theoretically no limit to the amount of inventory that can pile up after a resource (e.g. think of the plant manager walking through the process and saying "wow, we have a lot of inventory at this step today"). In contrast a Kanban system inventory becomes a managerial decision variable – the maximum inventory is controlled via the number of Kanban cards in the process.

As an alternative to a Kanban system, we can also implement a pull system using a make-to-order process. As is suggested by the word "make-to-order", resources in such a process only operate after having received an explicit customer order. Typically, these products corresponding to these orders then flow through the process on a first-in-first-out (FIFO) basis. Each flow unit in the make-to-order process is thereby explicitly assigned to one specific customer order. Consider the example of a rear view mirror production in an auto plant to see the difference between Kanban and make-to-order. When the operator in charge of producing the interior rear view mirror at the plant receives the work authorization through the Kanban card, it has not yet been determined which customer order will be filled with this mirror. All that is known is that there are – at the aggregate – a sufficient number of customer orders such that production of this mirror is warranted. Most likely, the final assembly line of the same auto plant (including the mounting of the rear-view mirror) will be operated in a make-to-order manner, i.e. the operator putting in the mirror can see that it will end up in the car of Mr Smith.

Many organizations use both forms of pull systems. Consider computer maker Dell. Dell's computers are configured in work cells. Processes supplying components are often operated using Kanban. Thus, rearview mirrors at Toyota and power supplies at Dell flow through the process in sufficient volume to meet customer demand, yet are produced in response to a Kanban card and have not yet assigned to a specific order.

When considering which form of a pull system one wants to implement, the following should be kept in mind:

- Kanban should be used for products or parts that are (a) processed in high volume and limited variety (b) are required with a short lead-time so that it makes economic sense to have a limited number of them (as many as we have Kanban cards) pre-produced and (c) the costs and efforts related to storing the components are low

- Make-to-order should be used for products or parts that are (a) processed in low volume and high variety (b) customers are willing to wait for their order (c) it is expensive or difficult to store the flow units

Section 8.5 Quality Management

If we operate with no buffers and want to avoid the waste of rework, operating at zero defects is a must. To achieve zero defects, TPS relies on defect prevention, rapid defect detection, and a strong worker responsibility with respect to quality.

Defects can be prevented by "fool-proofing" many assembly operations, i.e. by making mistakes in assembly operations physically impossible (poka-yoke). Components are designed in a way that there exists one single way of assembling them.

If, despite defect prevention, a problem occurs, TPS attempts to discover and isolate this problem as quickly as possible. This is achieved through the Jidoka concept. The idea of Jidoka is to stop the process immediately whenever a defect is detected and to alert the line supervisor. This idea goes back to the roots of Toyota as a maker of automated looms. Just like an automated loom should stop operating in the case of a broken thread, a defective machine should shut itself off automatically in the presence of a defect.

Shutting down the machine forces a human intervention in the process, which in turn triggers process improvement (Fujimoto 1999). The Jidoka concept has been generalized to include any mechanism that stops production in response to quality problems, not just for automated machines. The most well known form of Jidoka is the *Andon cord*, a cord running adjacent to assembly lines that enables workers to stop production if they detect a defect. Just like the Jidoka automatic shut-down of machines, this procedure dramatizes manufacturing problems and acts as a pressure for process improvements.

A worker pulling the Andon cord upon detecting a quality problem is in sharp contrast to Henry Ford's historical assembly line that would leave the detection of defects to a final inspection step. In TPS, "the next step is the customer" and every resource should only let those flow units move downstream that have been inspected and evaluated as good parts. Hence, quality inspection is "built-in" (tsukurikomi) and happens at every step in the line, as opposed to relying on a final inspection step alone.

The idea of detect-stop-alert that underlies the Jidoka principle is not just a necessity to make progress towards implementing the zero inventory principle. Jidoka also benefits from the zero inventory principle as large amounts of work in process inventory achieve the opposite of Jidoka: they delay the detection of a problem, thereby keeping a defective process running and hiding the defect from the eyes of management. This shows how the various TPS principles and methods are interrelated, mutually strengthening each other.

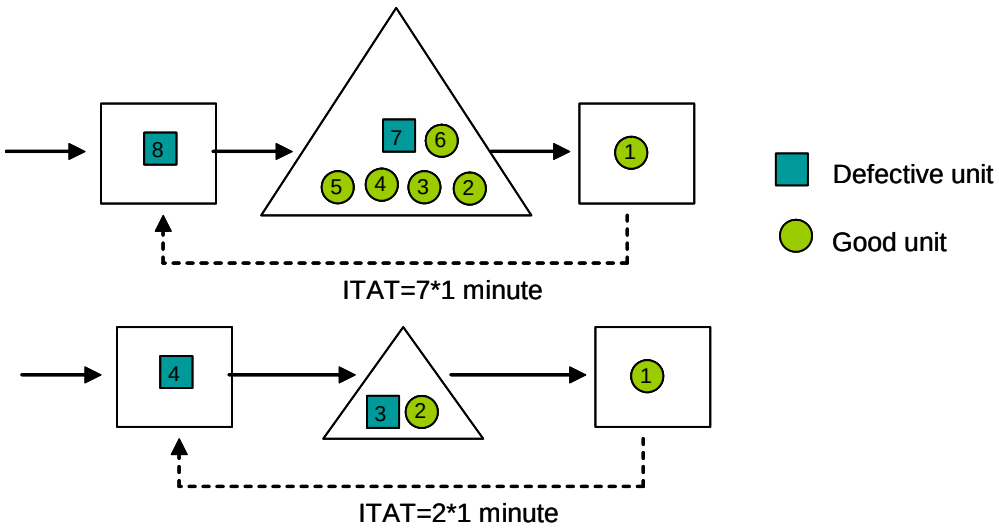


Exhibit ITAT: Information turnaround time and its relationship with buffer size

To see how work-in-process inventory is at odds with the idea of Jidoka, consider a sequence of two resources in a process as outlined in Figure ITAT. Assume the activity times at both resources are equal to one minute per unit. Assume further that the upstream resource (on the left) suffers quality problems, and - at some random point in time – starts producing bad output. In Figure ITAT, this is illustrated by the resource producing squares instead of circles. How long will it take until a quality problem is discovered? If there is a large buffer between the two resources (upper part of Figure ITAT), the downstream resource will continue to receive good units from the buffer. In this example it will take 7 minutes before the downstream resource detects the defective flow unit. This gives the upstream resource 7 minutes to continue producing defective parts that need to be either scrapped or reworked.

Thus, the time between when the problem occurred at the upstream resource and the time it is detected at the downstream resource depends on the size of the buffer between the two resources. This is a direct consequence of Little's Law. We refer to the time between creating a defect and receiving the feed-back about the defect as the *Information Turn-around Time (ITAT).* Note that we assume in this example that the defect is detected in the next resource downstream. The impact of inventory on quality is much worse if defects only get detected at the end of the process (e.g. at a final inspection step). In this case, the ITAT is driven by all inventory downstream from the resource producing the defect. This motivates the built-in inspection we mentioned above.

Section 8.6 Exposing Problems Through Inventory Reduction

Our discussion on quality reveals that inventory covers up problems. So to improve a process, we need to turn the "inventory hiding quality problems" effect

on its head: we want to reduce inventory to expose defects and then fix the underlying root cause of the defect.

Recall that in a Kanban system, the number of Kanban cards – and hence the amount of inventory in the process – is under managerial control. So we can use the Kanban system to gradually reduce inventory and thereby expose quality problems. The Kanban system and its approach to buffers can be illustrated with the following metaphor. Consider a boat sailing on a canal, which has numerous rocks in it. The freight of the boat is very valuable, so the company operating the canal wants to make sure that the boat never hits a rock.

One approach to this situation is to increase the water level in the canal. This way, there is plenty of water over the rocks and the likelihood of an accident is low. In a production setting, the rocks correspond to quality problems (defects), set-up times, blocking or starving, break-downs, or other problems in the process and the ship hitting a rock corresponds to lost throughput. The amount of water corresponds to the amount of inventory in the process (i.e., the number of Kanban cards), which brings us back to our previous "buffer or suffer" discussion.

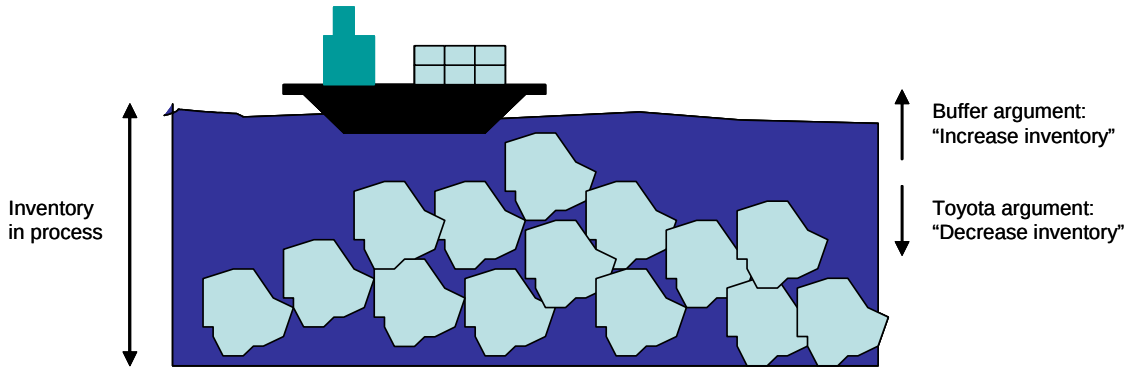


Exhibit BOAT: More or less inventory? A simple metaphor

An alternative way of approaching the problem is this: instead of covering the rocks with water, we could also consider reducing the water level in the canal (reduce the number of Kanban cards). This way, the highest rocks are exposed (i.e., we observe a process problem), which provides us with the opportunity of removing them from the canal. Once this has been accomplished, the water level is lowered again, until – step-by-step – all rocks are removed from the canal. Despite potential short-term losses in throughput, the advantage of this approach is that it moves the process to a better frontier (i.e., it is better along multiple dimensions).

This approach to inventory reduction is outlined in Figure FRONTIER. We observe that we first need to accept a short-term loss in throughput reflecting the

reduction of inventory (we stay on the efficient frontier, as we now have less inventory). Once the inventory level is lowered, we are able to identify the most prominent problems in the process (rocks in the water). Once identified, these problems are solved and thereby the process moves to a more desirable frontier.

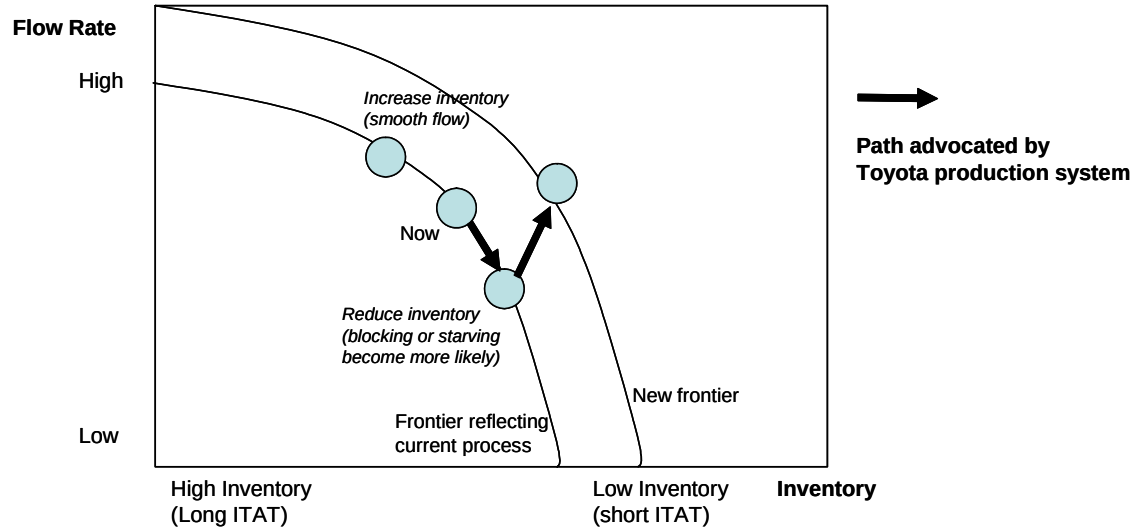


Exhibit FRONTIER: Tension between flow rate and inventory levels / ITAT

Both, in the metaphor and our ITAT discussion above, inventory is the key impediment to learning and process improvement. Since with Kanban cards, management is in control of the inventory level, it can proactively manage the tension between the short-term need of a high throughput and the long term objective of improving the process.

Section 8.7 Flexibility

Given that there typically exist fluctuations in demand from the end market, TPS attempts to create processes with sufficient flexibility to meet such fluctuations. Since forecasts are more reliable at the aggregate level (across models or components, see discussion of pooling in Chapter VARIABILITY and again in Chapter RISK POOLING), TPS requests workers to be skilled in handling multiple machines.

- When production volume has to be decreased for a product because of low demand, TPS attempts to assign some workers to processes creating other products, and to have the remaining workers handle multiple machines simultaneously for the process with the low demand product.

- When production volume has to be increased for a product because of high demand, TPS often uses a second pool of workers (temporary workers) to help out with production. Unlike the first pool of full-time employees (typically with life time employment guarantee and a broad skill set), these workers are less skilled and can only handle very specific tasks.

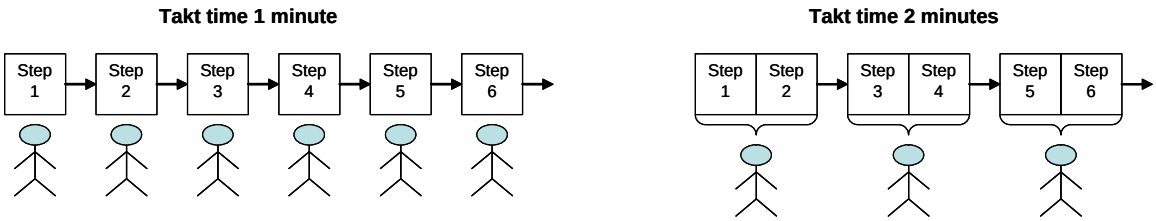


Exhibit TAKT-FLEX (Note: the Exhibit assumes a 1 min/unit activity time at each station)

Consider the six step operation shown in Exhibit TAKT-FLEX. Assume all activities have an activity time of 1 minute per unit. If demand is low (right), we avoid idle time (low average labor utilization) by running the process with only three operators (typically, full time employees). In this case, each operator is in charge of 2 minutes of work, so we would achieve a flow rate of 0.5 units per minute. If demand is high (left in the exhibit), we assign one worker to each step, i.e. we bring in additional (most likely temporary) workers. Now, the flow rate can be increased to 1 unit per minute.

This requires that the operators are skilled in multiple assembly tasks. Good training, job rotation, skill based payment and well documented standard operating procedures are essential requirements for this. This flexibility also requires that we have a multi-tiered workforce consisting of highly skilled full-time employees and a pool of temporary workers (who do not need such a broad skill based) that can be called upon when demand is high.

Such multi-task flexibility of workers can also help decrease idle time in cases of activities that require some worker involvement, but are otherwise largely automated. In these cases, a worker can load one machine and while this machine operates, the worker – instead of being idle – operates another machine along the process flow (takotei-mochi). This is facilitated if the process flow is arranged in a U-shaped manner, in which case a worker can not only share tasks with the upstream and the downstream resource, but also with another set of tasks in the process.

Section 8.8 Standardization of Work and Reduction of Variability

As we have seen in chapters VARIABILITY1 and VARIABILITY2, variability is a key inhibitor in our attempt to create a smooth flow. In the presence of variability, we either need to buffer (which would violate the zero inventory

philosophy) or we suffer occasional losses in throughput (which would violate the principle of providing the customer with the requested product when demanded). For this reason, the Toyota Production System explicitly embraces the concepts of variability measurement, control, and reduction discussed in the previous chapter.

The need for stability in a JIT process and the vulnerability of unbuffered processed was visible in the computer industry following the 1999 Taiwanese earthquake. Several of the Taiwanese fabs that were producing key components for computer manufacturers around the world were forced to shut down their production due to the earthquake. Such an unpredicted shut-down was more disruptive for computer manufacturers with JIT supply chains than those with substantial buffers (e.g. in the form of warehouses) in their supply chains (Papadakis 2002).

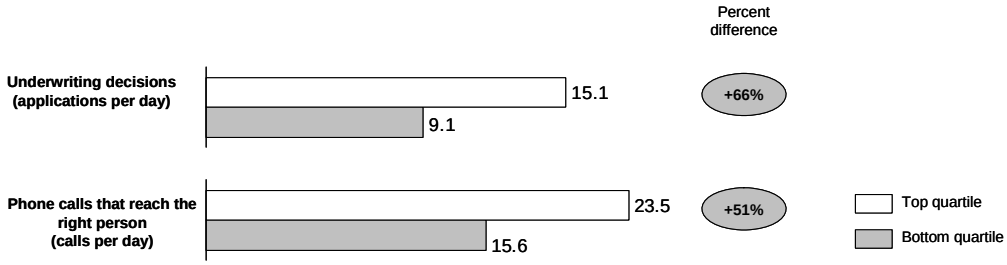


Exhibit QUARTILE: Productivity comparison across underwriters

Besides earthquakes, variability occurs because of quality defects (see above) or because of differences in activity times for the same or for different operators. Exhibit QUARTILE shows performance data from a large consumer loan processing organization. The exhibit compares the performance of the top quartile operator (i.e. the operator that has 25% of other operators achieving a higher performance and 75% of the operators achieving a lower performance) with the bottom quartile operator (the one who has 75% of the operators achieving a higher performance). As we can see, there can exist dramatic differences in the productivity across employees.

A quartile analysis is a good way to identify the presence of large differences across operators and to estimate the improvement potential. For example, we could estimate what would happen to process capacity if all operators would be trained so that they achieve a performance in line with the current top quartile performance.

Section 8.9 Human Resource Practices

We have seen seven sources of waste, but the Toyota Production System also refers to an eighth source – the waste of the human intellect. For this reason, a visitor to an operation that follows the Toyota Production System philosophy often encounters signs saying expressions like "In our company, we all have two jobs: (1) to do our job and (b) to improve it".

To illustrate different philosophies towards workers, consider the following two quotes. The first one comes from the legendary book "Principles of Scientific Management" written by Frederick Taylor, which still makes an interesting read almost a century after its first appearance (once you will have read the quote below, you will at least enjoy Taylor's candid writing style). The second quote comes from Konosuka Matsushita, the former Chairman of Panasonic.

Let us look at Taylor's opinion fist and consider his description of pig iron shoveling, an activity that Taylor studied extensively in his research. Taylor writes: "This work is so crude and elementary that the writer firmly believes that it would be possible to train an intelligent gorilla so as to become a more efficient pig-iron handler than any man can be".

Now, consider Matsushita, whose quote almost reads like a response to Taylor: "We are going to win and you are going to lose. There is nothing you can do about it, because the reasons for failure are within yourself. With you, the bosses do the thinking while the workers wield the screw drivers. You are convinced that this is the way to run a business. For you, the essence of management is getting the ideas out of the heads of the bosses and in to the hands of the labour. […] Only by drawing on the combined brainpower of all its employees can a firm face up to the turbulence and constraints of today's environment".

TPS, not surprisingly, embraces Matsushita's perspective of the "Combined brainpowers". We have already seen the importance of training workers as a source of flexibility.

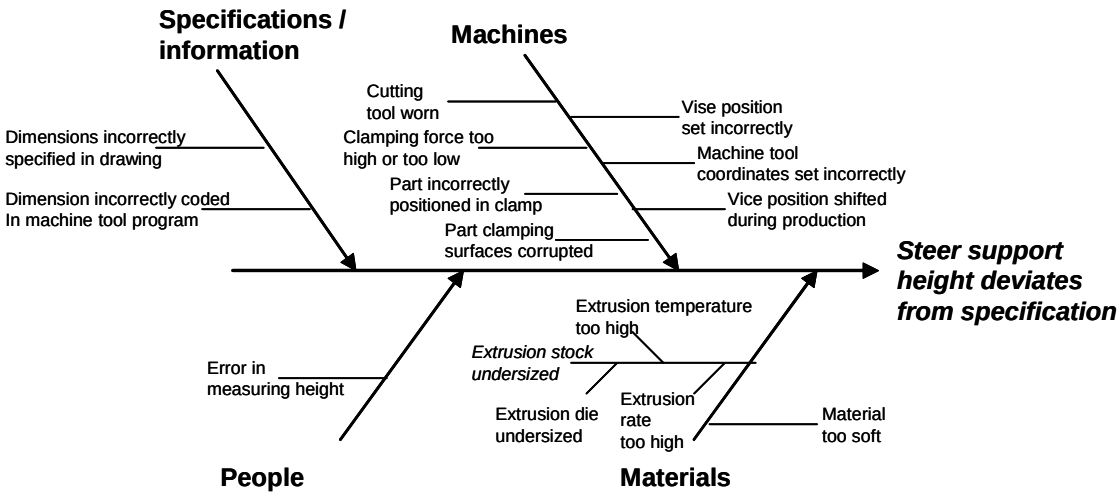


Exhibit ISHIKAWA: Example of an Ishikawa diagram

Another important aspect of the human resource practices of Toyota relates to process improvement. Quality circles bring workers together to jointly solve production problems and to continuously improve the process (kaizen). Problem solving is very data driven and follows a standardized process, including control charts, fish-bone (Ishikawa) diagrams, the "Five Whys" and other problem solving tools. Thus, not only do we standardize the production process, we also standardize the process of improvement.

Ishikawa diagrams (also known as fishbone diagrams or cause-effect diagrams) graphically represent variables that are causally related to a specific outcome, such as an increase in variation or a shift in the mean. When drawing a fishbone diagram, we typically start with a horizontal arrow that points at the name of the outcome variable we want to analyze. Diagonal lines than lead to this arrow representing main causes. Smaller arrows then lead to these causality lines creating a fishbone-like shape. An example of this is given by Figure ISHIKAWA. Ishikawa diagrams are simple yet powerful problem-solving tools that can be used to structure brainstorming sessions and to visualize the causal structure of a complex system.

A related tool that also helps in developing causal models is known as the "5 Whys". The tool is prominently used in Toyota's organization when workers search for the root cause of a quality problem. The basic idea of the "5 Whys" is to continually question ("Why did this happen") whether a potential cause is truly the root cause, or is merely a symptom of a deeper problem.

In addition to these operational principles, TPS includes a range of human resource management practices, including stable employment ("lifetime employment") for the core workers combined with the recruitment of temporary workers, a strong emphasis on skill development, which is rewarded financially

through skill-based salaries, and various other aspects relating to leadership and people management.

Section 8.10 Lean Transformation

How do you turn around an existing operation to achieve operational excellence as we have discussed it above? Clearly, even an operations management text-book has to acknowledge that there is more to a successful operational turnaround than the application of a set of tools.

McKinsey, as a consulting firm with a substantial part of its revenues resulting from operations work, refers to the set of activities required to improve the operations of a client as a Lean Transformation. There exist three aspects to such a lean transformation: the operating system, a management infrastructure, and the mindsets and behaviors of the employees involved.

With operating system, the firm refers to various aspect of process management as we have discussed in this chapter and throughout this book: an emphasis on flow, matching supply with demand, and a close eye on the variability of the process.

But technical solutions alone are not enough. So the operating system needs to be complemented by a management infrastructure. A central piece of this infrastructure is performance measurement. Just as we discussed in chapter FINANCE, defining finance-level performance measures and then cascading them into the operations is a key struggle for many companies. Moreover, the performance measures should be tracked over time and be made transparent throughout the organization. The operator needs to understand which performance measures she is supposed to achieve and how these measures contribute to the bigger picture. Management infrastructure also includes the development of operator skills and the establishment of formal problem solving processes.

Finally, the mindsets of those involved in working in the process is central to the success of a lean transformation. A nurse might get frustrated from operating in an environment of waste that is keeping him from spending time with patients. Yet, the nurse will in all likelihood also be frustrated by the implementation of a new care process that an outsider imposes on her ward. Change management is a topic well beyond the scope of this book: open communication with everyone involved in the process, collecting and discuss process data, and using some of the tools discussed in chapter SIX SIGMA as well as with respect to Kaizen can help make the transformation a success.

Section 8.11 Further Reading

Fujimoto (1999) describes the evolution of the Toyota Production System. While not a primary focus of the book, it also provides excellent descriptions of the main elements of the Toyota Production System. The results of the benchmarking studies are reported in Womack et al 1990 and Holweg and Pil 2004.

Bohn and Jaikumar (1992) is a classical reading that challenges the traditional, optimization-focused paradigm of operations management. Their work stipulates that companies should not focus on optimizing decisions for their existing business processes, but rather should create new processes that can operate at higher levels of performance.

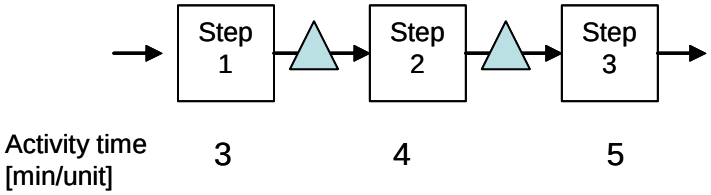
Drew et al 2004 describe the "Journey to Lean", a description of the steps constituting a lean transformation as described by a group of McKinsey consultants.

Tucker 2004 provides a study of TPS like activities from the perspective of nurses who encounter quality problems in their daily work.

The Wikipedia entries for Toyota, Ford, Industrial Revolution, Gilbreth, and Taylor are also interesting summaries.

Practice Problems

(Three Step) Consider a worker paced line with three process steps, each of which is staffed with one worker. The sequence of the three steps does not matter for the completion of the product. Currently, the three steps are operated in the following sequence.

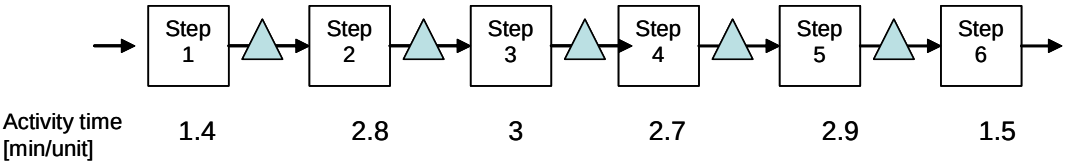


a. What would happen to the inventory in the process if the process were operated as a push system?

b. Assuming you would have to operate as a push system, how would you resequence the three activities?

c. How would you implement a pull system?

(Six Step) Consider the following six step worker paced process. Each resource is currently staffed by one operator. Demand is 20 units per hour. Over the past years, management has attempted to rebalance the process, but given that workers can only complete tasks that are adjacent to each other, no further improvement has been found.



a. What would you suggest to improve this process? (Hint: think “out of the box”)