

# **Standing on the Shoulders of Giants**

## **Production concepts versus production applications**

### **The Hitachi Tool Engineering example**

© Eliyahu M. Goldratt, 2008

#### ***Introduction***

It is easy to trace the popularity of Lean production to Toyota's success. Toyota's success is undeniable. Toyota now manufactures as many cars as the traditional leader – GM – and does it while making profits. Over the last five years, Toyota's average net profit over sales was 70% higher than the industry average, while GM is losing money.<sup>1</sup> The success of Toyota is fully attributed to the Toyota Production System (TPS).<sup>2</sup> At least this is the conviction of Toyota's management – the stated number one challenge of Toyota is to pass TPS on as the company's DNA to the next generation.

Given that Toyota is the flagship of Japan's industry, one should expect that Lean would be widely implemented in Japan. Surprisingly, this is not the case. It is commonly known in Japan that less than 20% of the manufacturers have implemented Lean. How come?

It is not because they did not try to implement it. Many companies in Japan put serious efforts into trying to implement Lean but failed. One such company is Hitachi Tool Engineering. Their inability to implement Lean cannot be explained by a lack of serious efforts. This company repeatedly tried to implement Lean but the deterioration in production performance forced them to go back to the more traditional ways of managing production.

Likewise, the fact that most of Japanese industry did not implement Lean cannot be attributed to a lack of sufficient knowledge. Toyota was more than generous in sharing their knowledge. This company put all the TPS knowledge in the public domain and even went as far as inviting their direct competitors to visit their plants. Hitachi, like so many other companies, was

---

<sup>1</sup> <http://moneycentral.msn.com/investor/invsb/results/compare.asp?Page=ProfitMargins&Symbol=TM>

<sup>2</sup> The Toyota Production System became known worldwide first under the name Just-In-Time (JIT) and later as Lean production. Toyota itself claims that Lean production does not fully capture its TPS spirit due to distortions in communications and implementations.

using the available knowledge and was not shy about hiring the help of the best experts available.

There is an explanation to these companies' failure to implement Lean; an explanation that is apparent to any objective observer of a company like Hitachi Tool Engineering. The failure is due to the fundamental difference in the production environments. When Taiichi Ohno developed TPS, he didn't do it in the abstract; he developed it for his company. It is no wonder that the powerful application that Ohno developed might not work in fundamentally different production environments. But, that doesn't mean that Ohno's work cannot be extremely valuable for other environments. The genius of Ohno is fully revealed when we realize that he faced the exact same situation. At that time, the production system that revolutionized production was the flow line method that Henry Ford developed. Ford's method was already used not only for almost all vehicle assembly, but also in very different industries like beverages and ammunition. Also, at that time, it was already accepted that flow lines can and must be implemented only in environments where the required quantities justify dedication of equipment to a single product. Whenever production quantities were not big enough, no one contemplated the possibility of using lines – no one except for Ohno.

Ohno realized that the concepts that underlie Ford's system are generic; that his application is restricted to certain types of environments, but the concepts are universal. Ohno had the clear vision to start from the concepts, the genius to design an application that is suitable for Toyota's environment, where it is not feasible to dedicate equipment to the production of a component, and the tenacity to overcome the huge obstacles standing in the way of implementing such an application. The result is TPS.

Rather than refraining from using the right concepts or, even worse, trying to force the application in environments that are apparently too different, we should follow in Ohno's footsteps.

In this paper, we will present

- The fundamental concepts of supply chains – the concepts that Lean is based upon,
- A generic application of these concepts that can be used in a much wider spectrum of environments, and
- The impressive results Hitachi Tool Engineering achieved with this broader application.

### *Historical perspective*

The manufacturing industry has been shaped by two great thinkers, Henry Ford and Taiichi Ohno. Ford revolutionized mass production by introducing the flow lines. Ohno took Ford's ideas to the next level in his TPS, a system that forced the entire industry to change its grasp of inventory from an asset to a liability.

Ford's starting point was that the key for effective production is to concentrate on improving the overall flow of products through the operations. His efforts to improve flow were so successful that, by 1926, the lead time from mining the iron ore to having a completed car, composed of more than 5,000 parts on the train ready for delivery, was 81 hours!<sup>3</sup> Eighty years later, no car manufacturer in the world has been able to achieve, or even come close, to such a short lead time.

Flow means that inventories in the operation are moving. When inventory is not moving, inventory accumulates. Accumulation of inventory takes up space. Therefore, an intuitive way to achieve better flow is to limit the space allowed for inventory to accumulate. To achieve better flow, Ford limited the space allotted for work-in-process between each two work centers. That is the essence of the flow lines, as can be verified by the fact that the first flow lines didn't have any mechanical means, like conveyers, to move inventory from one work center to another.

The daring nature of Ford's method is revealed when one realizes that a direct consequence of limiting the space is that when the allotted space is full, the workers feeding it must stop producing. Therefore, in order to achieve flow, Ford had to abolish local efficiencies. In other words, flow lines are flying in the face of conventional wisdom; the convention that, to be effective, every worker and every work center have to be busy 100% of the time.

One might think that preventing resources from working continuously will decrease throughput (output) of the operation. That undesirable effect might have been the result if Ford would have been satisfied with just limiting the space. But, there is another effect that stems from restricting the accumulation of inventory. It makes it very visible to spot the real problems that jeopardize

---

<sup>3</sup> Ford, Henry, *Today and Tomorrow*, Productivity Press, 1988 (originally published in 1926).

the flow – when one work center in a line stops producing for more than a short while, soon the whole line stops. Ford took advantage of the resulting clear visibility to better balance the flow by addressing and eliminating the apparent stoppages.<sup>4</sup> The end result of abolishing local efficiencies and balancing the flow is a substantial increase in throughput. Henry Ford achieved the highest throughput per worker of any car manufacturing company of his time.

In summary, Ford's flow lines are based on the following four concepts:

1. Improving flow (or equivalently lead time) is a primary objective of operations.
2. This primary objective should be translated into a practical mechanism that guides the operation when not to produce (prevents overproduction).
3. Local efficiencies must be abolished.
4. A focusing process to balance flow must be in place.

Like Ford, Ohno's primary objective was improving flow – decreasing lead time – as indicated in his response to the question about what Toyota is doing:

*All we are doing is looking at the time line from the moment the customer gives us an order to the point when we collect the cash. And we are reducing that time line...<sup>5</sup>*

Ohno faced an almost insurmountable obstacle when he came to apply the second concept. When the demand for a single product is high, dedicating a line to producing each component, as Ford did, is justified. However, at that time in Japan, the market demand was for small quantities of a variety of cars. Therefore, Ohno could not dedicate lines at Toyota. As we already said, all other industries that faced this situation simply did not contemplate using lines. Ohno, however, was toying with the idea of using lines when the equipment is not dedicated, when each work center is producing a variety of components. The problem was that in this case using the mechanism of limited space would lead to gridlocks – not all components are available for assembly (assembly cannot work) while the allotted space is already full (feeding lines are prevented from working).

---

<sup>4</sup> Balancing the flow is not equal to balancing the capacity – having the capacity of each work center match its load – a common mistake made when balancing flow lines.

<sup>5</sup> Ohno, Taiichi, *Toyota Production System*, Productivity, Inc. 1988, page ix (in Publisher's forward). It is also worth noting that in this and his other books Ohno gives full credit to Ford for the underlying concepts.

Ohno writes that he realized the solution when he heard about supermarkets (much before he actually saw a supermarket during his visit to the US in 1956). He realized that both supermarkets and the feeding lines at Toyota needed to manage a large variety of products. In the supermarkets, products were not jam packing the aisles, rather most merchandise was held in the backroom storage. In the store itself, each product was allocated a limited shelf space. Only when a product was taken by a client, replenishment from the backroom storage was triggered to refill that product's allotted shelf space. What Ohno envisioned is the mechanism that would enable him to guide Toyota's operation when not to produce. Rather than using a single limited space between work centers to restrict work-in-process production, he had to limit the amount allowed to accumulate of each component specifically. Based on that realization Ohno designed the Kanban system.

The Kanban system has been described in numerous articles and books. In this article we'll describe just the essence, to show how true Ohno was to the fundamental concepts. Between each two work centers,<sup>6</sup> and for each component separately, the accumulation of inventory is limited by setting a certain number of containers and the number of units per container. These containers, like every container in every industry, contain also the relevant paperwork. But, one page of the paperwork – usually a card (kanban in Japanese) – a page that specifies only the component code name and the number of units per container, is treated in an unconventional way. When the succeeding work center withdraws a container for further processing that card is not moved with the container, rather it is passed back to the preceding work center. This is the notification to that work center that a container was withdrawn, that the allotted inventory is not full. Only in that case is the preceding work center allowed to produce (one container of parts specified by the card). In essence the Kanban system directs each work center when and what to produce but, more importantly, it directs when not to produce. No card – no production. The Kanban system is the practical mechanism that guides the operation when not to produce (prevents overproduction). Ohno succeeded to expand Ford's concepts by changing the base of the mechanism from space to inventory.

Adhering to the flow concept mandates the abolishment of local efficiencies. Ohno addressed this issue again and again in his books, stressing that there is

---

<sup>6</sup> To reduce the number of places containers must be held, Ohno extensively used U-cells rather than using work centers that are composed of a single type of machines.

no point in encouraging people to produce if the products are not needed in the very short-term. This emphasis is probably the reason that outside Toyota TPS first became known as Just-in-time production.<sup>7</sup>

Once the Kanban system – the system that guides the operation when not to produce – was implemented on the shop floor the immediate reduction in throughput mandated the mammoth effort to balance the flow. The challenge that Ohno faced was orders of magnitude bigger than the one Ford faced. To realize how big the challenge was, it is enough to highlight just one aspect out of many. Not like in dedicated line environments, Ohno's system forced a work center to frequently switch from producing one component to another. For most work centers every such switch necessitates spending time to do the required setup. Since the containers, by design, called for a relatively small number of parts the production batches that they dictated were, many times, ridiculously small relative to the setup required. Initially for many work centers the time required for setups was more than the time required for production, resulting in a significant drop in throughput. It is no wonder that Ohno faced enormous resistance – so much so that Ohno wrote that his system was referred to as the 'abominable Ohno system' from the late 1940's to the early 1960's.<sup>8</sup> Ohno (and his superiors) certainly had an extraordinary determination and vision to continue to push for the implementation of a system, that for any person who looked at it from a local perspective, as most shop personnel must have, simply didn't make sense.

Ohno had to pave a new way to overcome the setup obstacle. At the time, and until TPS became famous worldwide, the traditional way to deal with setups was to increase the batch size – 'economical batch quantity' was the popular name on which thousands of articles were written.<sup>9</sup> Ohno ignored all that body of knowledge since yielding to using 'economical' quantities would have doomed his quest to reduce the lead times. Rather, he insisted that the setups required are not cast in stone, that the processes can be modified to drastically reduce the setup time required. He led the efforts to develop and implement setup reduction techniques that eventually reduced all setup times

---

<sup>7</sup> Nevertheless in the Lean literature there is no explicit stress on the fact that TPS mandates the abolishing of local efficiencies.

<sup>8</sup> Ohno, Taiichi and Setsuo Mito, *Just-In-Time For Today and Tomorrow*, Productivity Press, 1988.

<sup>9</sup> The first article was published by Ford W. Harris in *Factory, The Magazine of Management*, Volume 10, Number 2, February 1913, pp. 135-136, 152. Since then more articles on that subject are published almost every month.

in Toyota to be, at most, just a few minutes.<sup>10</sup> It is no wonder that Lean is now strongly associated with small batches and setup reduction techniques.

But, the need to balance the flow necessitated much more than just dealing with the setup obstacle. The fact that most work centers were not dedicated to a single component made it almost impossible to spot by direct observation the real problems which jeopardize the flow. Ohno was fully aware that there were too many things that can be improved, that without a way to focus the process improvement efforts it would take too long to balance the flow.

The Kanban system provided him such a way. The rocks and water analogy of Lean is useful for understanding how this is done. The water level corresponds to the inventory level, while the rocks are the problems disturbing the flow. There are many rocks at the bottom of the river and it takes time and effort to remove them. The question is which rocks are important to remove. The answer is given by reducing the water level; those rocks which emerge above the water are the ones that should be removed. At the initiation of the Kanban system, to achieve reasonable throughput, Ohno had to start with many containers each holding a non-negligible quantity of a particular part. Gradually, Ohno reduced the number of containers and then the quantities in each container. If the flow was not noticeably disturbed, then the reduction of the number of containers and quantities per container continued. When the flow was disturbed the Five Why's method was used to pinpoint the root cause. It had to be fixed before the quantities could be further reduced. It took time but the end result was a remarkable improvement in productivity.

It should be noted that even though, in the last twenty years, every other car company has implemented one version or another of the Toyota system and reaped major benefits, the productivity of Toyota is unmatched by any other car company. This fact points to the importance of choosing correctly the process that focuses the local improvement efforts. Unfortunately, the improvement efforts of other companies are misguided since they are aimed at achieving cost savings rather than being totally focused on improving the flow.

Ohno did not invest so much effort in reducing the setup times in order to gain some cost savings. If saving cost would have been his target he would

---

<sup>10</sup> For example, Toyota's die changes went from two to three hours in the 1940's to less than one hour and as low as 15 minutes in the 1950's to 3 minutes in the 1960's (Ohno wrote this in his book, *Toyota Production System*).

not have 'wasted' the time saved by further reducing the batches and therefore doing much more setups. Ohno did not try to reduce the number of defective parts in order to save some (trivial) costs; he did it to eliminate the major disruptions to flow that result from having a defective part. Ohno did not even try to squeeze better prices from Toyota suppliers or to cut the payroll of Toyota (the two main elements of cost); rather he put all his energy into improving the flow.

What is obliterating the picture is that the end result of focusing on flow and ignoring local cost considerations is a much lower cost per unit. Exactly like the end result of abolishing local efficiencies is much higher efficiency of the labor force. If it looks strange, it is because managers have not yet internalized the conceptual difference in guiding operations to concentrate on improving throughput rather than concentrating on reducing costs. One of the ramifications of concentrating on cost reduction is that almost all initiatives to foster a process of on-going improvement quickly reach a point of diminishing returns and as a result many deteriorate to lip service. But, that issue is too broad and too important to be squeezed into this article.

In summary, both Ford and Ohno followed four concepts (from now on we'll refer to them as the concepts of supply chain):

1. Improving flow (or equivalently lead time) is a primary objective of operations.
2. This primary objective should be translated into a practical mechanism that guides the operation when not to produce (prevents overproduction). Ford used space; Ohno used inventory.
3. Local efficiencies must be abolished.
4. A focusing process to balance flow must be in place. Ford used direct observation. Ohno used the gradual reduction of the number of containers and then gradual reduction of parts per container.

### ***The boundaries of TPS***

Ohno's approach in developing Lean demonstrates an important idea: there is a difference between an application and the fundamental concepts on which the application is based. The fundamental concepts are generic; the application is the translation of the concepts for a specific environment. As we have already seen, the translation is not trivial and necessitates a number of solution elements. What we have to bear in mind is that the application

makes assumptions (sometimes hidden assumptions) about the environment. We should not expect an application to work in environments for which its assumptions are not valid. We can save a lot of effort and frustration if we bother to explicitly verbalize these assumptions.

The most demanding assumption that TPS is making about the production environment is that it is a stable environment. And it demands stability in three different aspects.

The first aspect is revealed once we pay attention to the fact that, even when an appropriate environment is chosen and the best experts are supervising the implementation, it takes considerable time to implement Lean. Liker points out in *The Toyota Way* that Lean implementations led by the Toyota Supplier Support Center (TSSC, the organization Toyota created to teach U.S. companies TPS) take a minimum of six to nine months per production line.<sup>11</sup> This is not a surprise to anybody who is aware of the number of disruptions to flow that exist in almost any production environment and the sensitivity of the Kanban system once it starts to reach its target of low inventory. Since the Kanban system takes time to implement, its assumption is that the environment is relatively stable – that the processes and the products do not change significantly for a considerable length of time.

Toyota enjoys a relatively stable environment. The car industry allows changes only once a year (a model year change) and usually, from one year to another, the vast majority of the components are the same. That is not the case for many other industries. For example, in major sections of the electronics industry, the life span of most products is shorter than six months. To some extent, instability of products and processes exists in most other industries. For example, Hitachi Tool Engineering is producing cutting tools, a relatively stable type of product, but fierce competition forces this company to launch new cutting tools, that require new technology, every six months. It is a Sisyphean task to implement Lean in such an environment.

A second aspect of the stability required by TPS is stability in demand over time per product. Suppose that the lead time to produce a certain product is two weeks but the demand for that product is sporadic; on average there is just one order per quarter for that product. Currently, this product contributes to the work-in-process only during two weeks in a quarter; the rest of the time it is not present on the shop floor. But, that will not be the case under Lean,

---

<sup>11</sup> Liker, Jeffrey K., *The Toyota Way*, McGraw-Hill, 2004.

which mandates permanently holding containers for each product between each two work centers.

Hitachi Tool Engineering is producing over twenty thousand different SKUs. For most SKUs the demand is sporadic. The necessity to permanently hold, for each SKU, inventory between each two work centers would lead, in the Hitachi case, to holding considerably more work-in-process inventory than what they hold today. This is apparently not a suitable environment for Ohno's application.

But, the most demanding aspect of the stability required by TPS is stability in total load placed by the orders on the various types of resources. Suppose that, like in most companies, the orders are not uniform throughout. It is very likely that the load placed this week on a particular work center is considerably lower than its capacity while next week the load is higher than its capacity. In this very common case, the Kanban system, that is preventing build ahead – preventing producing ahead of time – will lead to missed due dates in the second week. Toyota's orders are relatively stable and nevertheless, Toyota had to establish a mode of receiving orders (and promising deliveries) that restricts the mix change from one month to another. Most companies are not able to enforce on their clients such favorable conditions.

It is important to note that the required stability is outside the power of production to improve. All three aspects of stability have to do with the way the company designs and sells its products and not with the way it produces them. Unfortunately, the majority of companies suffer from at least one aspect of instability, if not from all three.

The above doesn't mean that, for environments in which the assumptions of Lean are not valid, fragments of Lean cannot be used (e.g. U-cells can be helpful in many environments and setup reduction techniques can be used in almost every environment). But, it does mean that, in such environments, one should not expect to get the same magnitude of results that Toyota achieves – results that elevate that company into what it currently is. Using some specific techniques of Lean, being satisfied with some cost saving programs, shouldn't be considered as implementing Lean.

### *The importance of flow in relatively unstable environments*

Ford and Ohno opened our eyes to the fact that better flow – reducing lead time – leads to much more effective operations. They have demonstrated it on stable environments but what is the impact of improved flow on relatively unstable environments?

The first aspect of instability is instability due to short product life. When the products' life is short, overproduction can become obsolescence. Moreover, since the lifetime is short, long production lead-times lead to missing the market demand. For example, suppose that the lifetime of a product is about 6 months and the production lead-time of that product is two months. The long production lead time results in lost sales, not because the demand is not there, but because, for a significant portion of the market life, production cannot satisfy the demand.

The second aspect of instability is instability in demand over time per product. The common practice in environments that have a large number of SKUs that are subject to sporadic demand is to reduce the hassle by trying to satisfy this demand from stock. The disadvantage of this practice is high finished goods inventories that turn extremely slowly coupled with high levels of shortages. A production system that is capable of organizing the shop floor to the extent that much better flow is achieved has a drastic impact on these environments.

Environments that suffer from the third aspect of instability – instability in the overall load – are the ones that can gain the most from much better flow. The temporary overloads on the various resources cause these companies to usually have relatively poor due date performance (< 90%) and, as a result, they are inclined to add more capacity. Experience shows that when such companies succeed to drastically improve flow, not only do their due-dates reach the high nineties but excess capacity, as high as 50%, is revealed.<sup>12</sup>

Ohno demonstrated that the concepts Ford introduced are not restricted to mass production of a single type of product. Even though the obstacles to apply these concepts to a less restrictive environment looked insurmountable,

---

<sup>12</sup> Mabin, Victoria J. and Balderstone, Steven J., *The World of the Theory of Constraints*, CRC Press LLC, 2000. A review of the international literature on TOC analyzed the average results achieved: 70% reduction in lead time, 44% improvement in due-date performance and a 76% increase in revenue/throughput/profit.

Ohno's genius and tenacity proved to us, not only that it can be done but how to do it.

We now realize that:

- TPS is restricted to relatively stable environments,
- Most environments suffer from instability, and
- Relatively unstable environments have much more to gain from better flow than even stable environments.

Now that we realize the above shouldn't we follow in the footsteps of Taiichi Ohno? Shouldn't we go back to the supply chain concepts and derive an effective application that is suitable for the relatively unstable environments?

### *A time-based application of the supply chain concepts*

The most intuitive base for the mechanism to restrict over-production is not space or inventory but time – if one wants to prevent production ahead of time one should not release the material ahead of time. Using time as the base is not only more intuitive and, therefore, more easily accepted by the shop floor, it has an advantage that makes it suitable for unstable environments – it is much less sensitive to disruptions in flow.

The robustness of the time-based mechanism stems from the fact that it directly restricts the overall amount of work in the system rather than doing it through restricting the amount of work between each two work centers. In flow lines or Kanban-based systems the allotted inventories between work centers is restricted to the bare minimum (usually corresponding to much less than one hour of work). Therefore, when a work center is down for more than a short while the succeeding work centers are almost immediately starved for work and the preceding work centers are “blocked” from working. When, for any of the work centers, the accumulated time consumed by starvation and blockage is more than the excess capacity of that work center, the throughput of the company is reduced. The sensitivity of flow lines and Kanban-based systems stems from the fact that a disruption that occurs in one work center consumes capacity also from the upstream and downstream work centers – a phenomenon that (almost) doesn't exist for the time-based systems since the work, once released to the floor, is not artificially restrained.

The difficulty in using a time-based system is that, for each order, we should restrict the release of the corresponding material to be an appropriate time before the due date of the order. But, how does one go about computing the appropriate time? When computers appeared on the industrial scene (in the early sixties) it looked like we, at last, had the proper tool to handle the immense amount of details and calculations needed to compute the appropriate times for each material and order. Within ten years many computer programs, to do just that, were developed in numerous companies around the world. Unfortunately, the expected results of better flow and less work-in-process did not materialize.

The problem is that the time it takes material to be converted to a finished product, ready for delivery to the client, depends more on the time it has to wait in queues (waiting for a resource that is busy processing another order or waiting in front of assembly for another part to arrive) and not so much on the touch time to process the order. It is commonly known that in almost any industrial operation (except for process lines and companies that use the Kanban system) the time that a batch of parts spends being processed is only about 10% of the lead time. As a result, the decision of when to release the material determines where and how big the queues will be, which in turn determines how much time it will take to complete the order, which determines when to release the material. We were facing a chicken and egg problem. In the seventies it was suggested to handle that problem by reiterating the procedure (closed loop MRP) – to run the computer system, to check the resulting planned overloads on the various resources (the size of the queues), to adjust the due dates to eliminate the overloads, and to repeat this process until all meaningful overloads were eliminated. This suggestion did not last long since experience showed that the process doesn't converge; that no matter how many iterations are done the overloads just move from one resource type to another.

As a result, already in the seventies, the usage of these computer systems was not to guide the precise timing of release of material to the shop floor but rather it was confined to giving better information on the quantities (and timing) to order material from the suppliers. The official name of these systems was coined to reflect their major usage – Material Requirements Planning (MRP).<sup>13</sup>

---

<sup>13</sup> Orlicky, Joseph, *Material Requirements Planning*, McGraw-Hill Book Company, 1975.

The fact that such a mammoth effort did not yield a practical time-based mechanism to guide operations when not to produce, should not be taken as a proof that such a mechanism cannot be developed for the less stable environments – environments that must meet the due-dates of an uneven flow of clients' orders. It should not even discourage us from attempting to use time as the base for a practical mechanism. But, it should be a warning against an approach that tries to develop such a mechanism through handling the immense amount of details and calculations. What is needed is more of a bird's eye view approach.

Going back to basics, following the concepts of supply chain, the objective is to improve flow – to reduce the lead time. Taking time (rather than space or inventory) as the base for the mechanism to guide the operation when not to produce mandates that we should strive to release the corresponding material an appropriate short time, just-in-time, before the due date of the order. But, what do we mean by 'just-in-time'? Even though the term 'just-in-time' is a key concept in Lean its use is figurative and not quantitative. In Lean, by production just-in-time we certainly don't mean that the part that was worked on just now is needed to be at the loading dock ready for shipment in the next second... or minute... or hour. Actually, it is likely, that even under the best Kanban systems, this part will not be worked on right away by the succeeding work center (as can be deduced from the fact that full containers are routinely waiting between work centers). So, what time interval will we consider to be 'just-in-time'? More explicitly: if we want to restrict overproduction by restricting the release of the material, how much time before the due date of an order should we release the material for that order?

One way to reach a reasonable answer is through examining the impact the choice of that time interval has on the magnitude of the management attention required to meet all due-dates. Suppose that we release material before the due date by just the time it actually takes to process the order. Such a choice will necessitate a lot of management attention to closely monitor operations, since any delay in any operation or even a delay in moving the parts between operations will result in missing the due date. Moreover, precise scheduling will be needed to ensure that no queues will occur since any queue causes a delay for the parts waiting in the queue. This is certainly not a practical choice, even infinite management attention will not be sufficient to meet all due-dates. We must choose a longer interval of time; an interval that contains safety to accommodate delays. The need to include safety is the reason for

referring to the time interval of release of material before the due-date as the 'time buffer'.

Choosing longer time buffers elongates the lead time and increases work in process, but since longer time buffers means more safety time, expectations are that, with much less management attention, a higher percentage of orders will be completed on or before their respective due-dates. This is correct for relatively short time buffers, but when the time buffers are considerable, another phenomenon starts to raise its ugly head. What we have to bear in mind is that the longer the chosen time buffer, the earlier material is released which means that more orders are simultaneously present on the shop floor. When there are too many orders on the floor, traffic jams start to occur. The more traffic jams, the more management attention is needed to sort out the priorities. The magnitude of the required management attention as a function of the length of the chosen time buffer is shown schematically in figure 1.

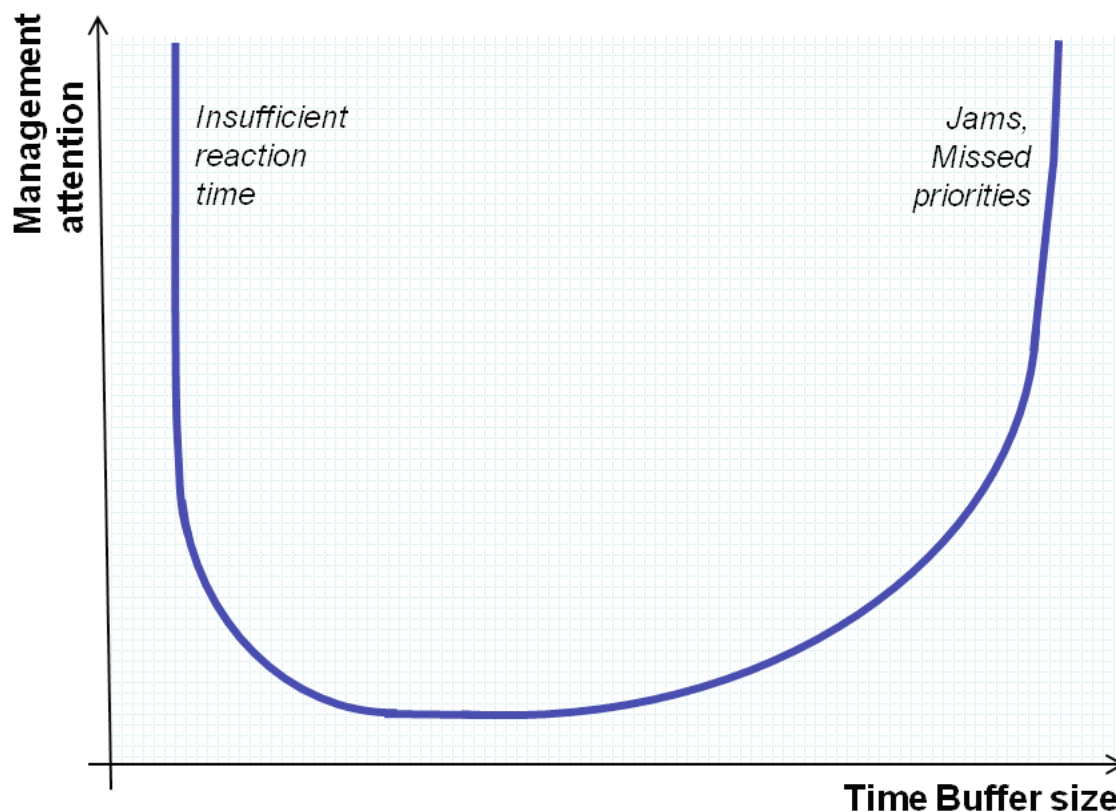


Figure 1

Operations that implemented Ford's or Ohno's systems enjoy an average lead time which is only a few times longer than the actual touch time and management do not have to invest almost any attention to guide the shop floor

personnel on what to work on now. They definitely reside at the left hand side of the low plateau of that graph.

But, where on the graph are the vast majority of operations, operations that are using the more conventional practice?

As we said, in conventional plants batches of parts spend only about 10% of the time being processed. About 90% of the time the batches are either waiting in a queue for a resource or waiting for another type of part to be assembled together. What we learned from Ford, and more so from Ohno, is that we shouldn't accept the size of batches as given; that economical batch quantities are not economical and instead we should and can strive to reach a one-piece flow. Armed with that conviction it is easy to realize that when a batch of parts is being processed (except in processes like mixing or curing) only one item is actually worked on while the other items in the batch are waiting. That means that in conventional companies that use batch sizes of more than ten units in a batch (which is the case in the majority of production environments) the touch time is actually less than 1% of the lead time. There is another phenomenon that typifies these companies; whatever the formal priority system is, if a formal priority system exists at all, the actual priority system is: "hot", "red hot" and "drop everything - do it now". These companies are apparently high on the right hand side slope of the management attention versus time buffer graph (figure 1).

Being on the right hand side slope means being in a lose-lose situation; lead times are very long (relative to the touch time), inventories are high and in many cases the company suffers from poor due-date performance (<90%) in spite of high management attention. Bearing in mind that if management would have chosen a shorter time buffer (moving into the wide plateau region of the graph) the situation would be remarkably better, how can it be that the vast majority of conventionally run companies are in that lose-lose situation?

The answer was given by Ford and Ohno. Through their work they, decisively, proved that contrary to the common belief, striving to constantly activate all resources all the time is not a recipe for effective operations. On the contrary, the exact opposite is true; to reach effective operations, local efficiencies must be abolished. But conventional companies do try to reach full activation of resources. Whenever the upstream resources are not bottlenecks (and that is the case in the vast majority of environments) they will, from time to time, run out of work. To prevent it, material is released; material that is needed for more remote orders (or even for forecasted orders).

The unavoidable consequence is longer queues. Longer queues cause some orders not to be fulfilled on time which in turn is interpreted as: we should release the material earlier. And is also interpreted as: we don't have enough capacity. It is not difficult to envision how such forces push companies up the slope.

A good starting point for improving flow will be to choose the time buffer to be equal to half the current lead time; such a choice will ensure that the company will find itself somewhere on the plateau of the graph. There is no point wasting time by trying to find or calculate the optimum point, the immediate benefits are too significant to postpone and the next efforts to balance the flow will modify the graph itself.

Restricting the release of material to be just the time buffer (half the current lead time) before the corresponding due-date of the orders will considerably improve the due date performance, will reduce the lead time to half of what it is now, and therefore as the excess inventories are flushed out, will shrink the work-in-process inventory to less than half of its current level.

But one cannot expect that this change alone will bring the due-date performance to the high nineties. Simply there are still many orders on the shop floor, there are queues in front of resources and leaving to chance the sequence in which the work is processed will cause many orders to finish late. A priority system is needed. The need for a priority system should not open the gates for sophisticated algorithms to set the priorities. The number of orders coming in is constantly changing, the content of work differs from one order to another, the length of the queue is constantly changing and let's not forget that disruptions still occur; in short, this is an environment with high variability. The lesson that Shewhart brought to manufacturing from Physics, and Deming made known worldwide, is that trying to be more accurate than the noise (in our case, trying to use sophisticated algorithms that consider every possible parameter in an environment of high variability) does not improve things but makes them worse – the results will most certainly not be an improvement but a deterioration in due-date performance.

A straightforward priority system emerges when we recognize that the time buffer, being half of the current lead time, is still much longer than the touch time and since it dramatically reduces the traffic jams, without any interference, many orders will be finished within just one-third of the time buffer and the majority will be finished within the first two-thirds of the time

buffer. Based on that realization, priorities are assigned by ‘buffer-management’. Per batch, the time that has passed since its release is tracked. If less than one third of the time buffer has passed the priority color is green, if more than one-third but less than two-thirds the priority color is yellow, if more than two thirds the color is red, if the due date has passed the color is black. Blacks have higher priority than reds, etc. If two batches have the same color, to try and decide which one should be worked on first is an excellent example of trying to be more accurate than the noise.

Putting such a system on the shop floor is relatively easy. In the first step, there is no need to do any physical changes, just to choke the release of material to be half the historical lead time before the corresponding due-date and to guide the shop floor to follow the color code priority system. The impact is impressive, especially compared to the efforts. To get a first hand impression of the impact (and the speed) from just the first step, Figure 2 gives the actual percentage of late orders of a 2,000 worker plant that produces thousands of different types of metal kitchenware.

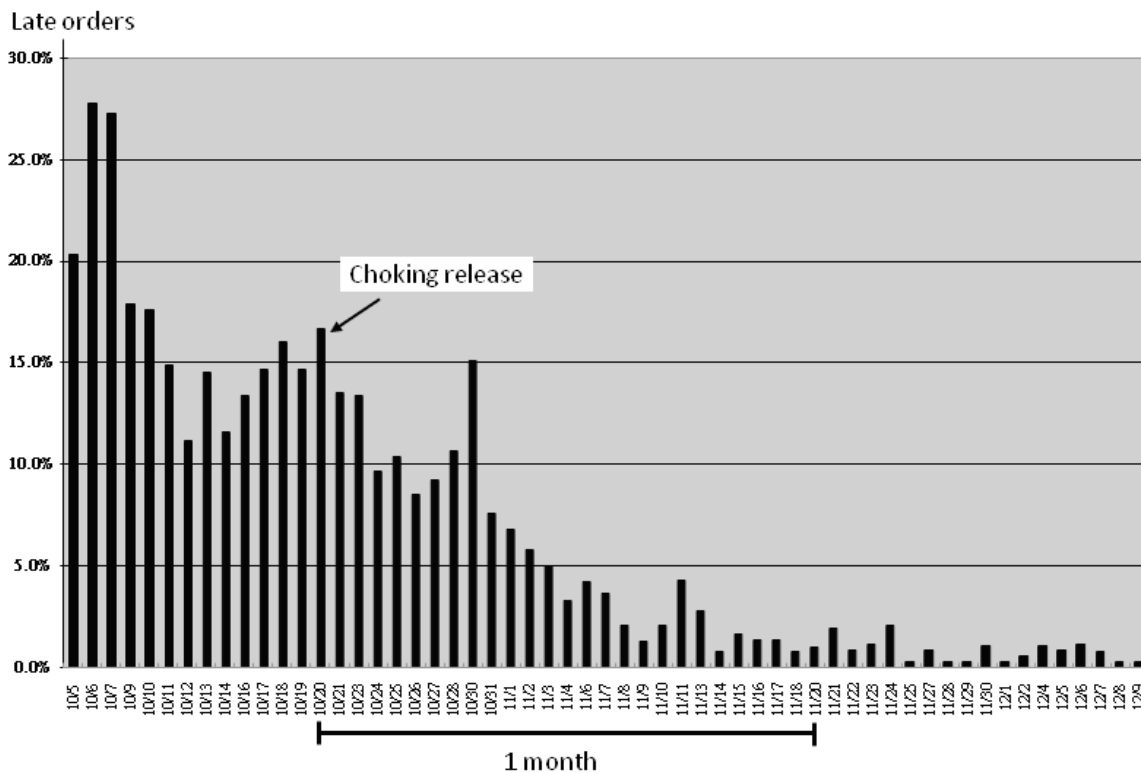


Figure 2

Of course local efficiencies must be abolished, otherwise the pressure to release material too early will continue. Experience shows that the speed at which everybody on the shop floor realizes the positive impact, makes that change almost resistance free.

But, in most environments there are still orders that miss their due dates and there is still enormous potential for improvement to capitalize on. The fourth concept must also be translated into practice – a focusing process to balance flow must be in place.

The first step in balancing the flow is relatively easy. Choking the release of material exposes the abundant excess capacity that was masked before. But it is likely that some work centers have less excess capacity than the rest. These work centers are flagged since they do have a queue of inventory in front of them. The fact that local efficiencies are abolished helps to identify the simple actions required to increase their capacity – simple actions such as ensuring that a capacity-constrained work center will not stay idle during lunch break or shift changes, offloading work to less efficient work centers that have ample excess capacity, etc.<sup>14</sup>

Since the above actions add effective capacity to the work centers that cause queues, the queues become shorter and fewer orders reach the red status. This means that the time buffer becomes unnecessarily long. An effective rule to adjust the time buffer, without taking a risk of deteriorating the high due date performance, is to decrease the time buffer when the number of red orders is smaller than 5% of the number of total released orders and to increase it when the proportion of red orders is more than 10%.

A company that follows the above will find itself, within a few months, with very high due date performance, considerably shorter lead times and ample excess capacity. This is when the real challenge begins. In the past, sometimes (too many times) the reaction of top managers to the fully exposed excess capacity was to ‘right size’ the capacity and gain cost savings. This is a grave mistake. The ‘excess capacity’ is employees – employees who just helped the company to improve and as a direct consequence are ‘rewarded’ by losing their or their friends’ jobs. In all the cases in which such an action was taken, the unavoidable backlash quickly deteriorated the plant performance to worse than the starting point. Hopefully such top management behavior is behind us.

---

<sup>14</sup> Goldratt, Eliyahu M. and Cox, Jeff, *The Goal: A Process of Ongoing Improvement*, North River Press, 1984.

The more sensible way to deal with the exposed excess capacity is to capitalize on it; to encourage the sales force to take advantage of the improved performance to gain more sales. The increased sales can easily cause the emergence of a real bottleneck. Ignoring the bottleneck's limited capacity when giving due-date commitments to new orders will deteriorate the due-date performance and sales from disappointed clients will plummet. It is essential to strengthen the tie between sales and operations – that is the real challenge. A system must be put in place to ensure that every due-date commitment is given only according to the yet unallocated capacity of the bottleneck.

The bottleneck becomes the 'drum beat' for the orders, the 'time buffer' translates due-dates into release dates and the action of choking the release becomes the 'rope' that ties the order to the release of work. That is the reason this time-based application of the Theory of Constraints became known as the Drum-Buffer-Rope system or in short DBR.

Currently there is widespread experimentation to polish a process to further improve operations based on recording and analyzing the reasons for the red orders.

### ***Example of Hitachi***

Hitachi Tool Engineering Ltd., a 24 billion yen company, designs and manufactures over 20,000 different cutting tools. The demand for most products is sporadic, and the customs of their industry force them to launch new product families of tools every six months. When new product families are launched, the older families become obsolete. No wonder their efforts to implement Lean were unsuccessful.<sup>15</sup>

Hitachi Tool Engineering Ltd. started implementing DBR in one of their four plants in Japan in 2000. The jump in due-date performance (from 40% to 85%) associated with cutting WIP and lead times in half along with the ability to ship 20% more products with the same labor force encouraged them to

---

<sup>15</sup>Umble, M., Umble E., and Murakami, S., "Implementing theory of constraints in a traditional Japanese manufacturing environment: the case of Hitachi Tool Engineering," *International Journal of Production Research*, Vol. 44, No. 10, 15, May 2006, pp. 1863 – 1880.

expand the implementation. By 2003, they had implemented DBR in all four plants.<sup>16</sup>

The drastic reduction in lead time and the much better responsiveness enabled a reduction of inventory in the supply chain – the distributors – from 8 to 2.4 months' worth. The reduction of inventory improved dramatically the distributors' return on investment, freed up their cash and strengthened their relationships with Hitachi. No wonder the distributors expanded the range of Hitachi tools that they were offering, leading to an increase of 20% in sales (in a stable market).

The true impact is revealed when we evaluate this company's bottom line performance in light of the fact that during the period of 2002 to 2007 the price of raw materials (metals) increased much more than the increase in the selling price of cutting tools. Under such conditions the profits of the company should have vanished. Instead, the annual net profit before taxes of Hitachi Tool Engineering Ltd increased from 1.1 billion yen in the fiscal year ending March 2002 to 5.3 billion yen in the year ending March 2007 – a five-fold increase in net profit in five years. The profit ratio of Hitachi Tool Engineering Ltd increased from 7.2% in 2002 to 21.9% in 2007, the highest ratio ever reported in this type of industry.<sup>17</sup>

### *The boundaries of DBR*

As was highlighted before, an application makes assumptions (sometimes hidden assumptions) about the environment and we should not expect the application to work in environments for which its assumptions are not valid. The assumption that DBR makes is apparent, it assumes that the touch time is very small (<10%) compared to the current lead time. This assumption is valid for many, if not most, typical production environments. But, definitely it is not valid for a very broad range of environments that are traditionally called project environments.

In project environments the touch time is relatively long and the eagerness of the clients to get the project completed forces operations to promise lead

---

<sup>16</sup> Ibid.

<sup>17</sup> A GUIDE TO MAKING EVER FLOURISHING COMPANY - PRODUCTION, DISTRIBUTION, MARKETING AND SALES. Chukei Publishing, 2008. Satoru Murakami, Jun Takahashi, Shotarou Kobayashi p196~p207

times which are only twice (or, rarely, three times) longer than the touch time. It is no wonder that the performance is bad to the extent that no one expects to get the project completed on time, in budget and with the full content; something is expected to give. But, that fact shouldn't distract us from the conclusion that since DBR's assumption is not valid, DBR is inappropriate for project environments. A different application, an application that directly addresses the relatively long touch time, is needed.<sup>18</sup>

---

<sup>18</sup> Goldratt, Eliyahu M., *Critical Chain*, North River Press, 1996.