

An Overview of the *Creating Level Pull* Workbook

By [Art Smalley](#)

The basic methods for implementing level pull production are well understood, having been developed by Toyota and its affiliated companies over many years. There also is now a considerable base of experience in introducing these methods in firms outside of Toyota. The challenge therefore is to provide a simple recipe for introducing these concepts in your facilities.

Based on my experience in converting facilities from push to level pull-based production, I've developed 12 questions and a step-by-step example for the [Creating Level Pull workbook](#) to guide you through the implementation process.

What follows is a summary of the questions. It's not intended to be a complete guide to the workbook or the [associated training workshop](#). Instead this document is meant to be a high-level overview, emphasizing many shop-floor basics surrounding scheduling execution that are often broken or under-analyzed for improvement potential. The outline for this material follows the same 12 basic questions in the workbook. In addition, it emphasizes what I believe are some key points in implementation.

1. What components to make to stock and what to make to order?

The history of pull production is basically a replenishment model with inventory being held in a market location. Depletion of that inventory is used to signal replenishment. If this situation is true for your business or value stream then an important issue is to decide what items to hold in inventory and how much of each. In particular for finished goods (FG) in the value stream (and at the raw material and component level) the number of items in inventory needs to be carefully analyzed and established. Essentially you need to determine what items should be made to stock and replenished and what items should be made to order in the value stream. If the situation is fairly typical then about 20% of the part numbers in a value stream will probably account for 80% of the volume and the remaining 80% of the part numbers will account for 20% of the volume. *Note: In replenishment systems if this 80/20 rule is not true then the real issue here may actually be failure to standardize designs in engineering which is manifesting itself as a scheduling problem in manufacturing.*

In replenishment cases as a starting point, you should put the high runner end items into a market in order to provide good levels of customer service and create better schedule stability inside the four walls of manufacturing. When correctly utilized, inventory can help provide quick and reliable delivery to the customer. An equally important but often overlooked point is that inventory can also help buffer the plant against large swings in customer demand. If you don't buffer, the swings will pass directly through the value stream and potentially create large amounts of disruption and inefficiency. Toyota and other lean companies strive to create a more level daily schedule precisely in order to avoid this type of daily disruption.

An important end goal of lean is to reduce the timeline from the time an order is received until it is delivered through the elimination of waste. Of course this means improving flow and reducing inventory as much as possible. However, instead of thinking of inventory as a pure form of waste, I encourage you to remember that the actual quote from Taiichi Ohno, credited as the chief developer of the Toyota Production System, was that “inventory beyond what you need to smoothly run the process is waste.” Not having inventory in the right place at the right time is often a bigger waste than having too much of the inventory itself – both aspects must be improved.

Of course, in other nonreplenishment cases, production items are either made (i.e. engineered) to order, are too numerous, or are too expensive to hold in inventory. In these cases although everything may truly be unique at the final assembly stage often many things are more common at the sub-assembly, component or raw material stage. As a first step look for where it is possible to hold inventory in the value stream before the finished goods stage. Can you consolidate standard types at a mid-point somewhere? Or can you at least hold the raw material and purchased parts you need in inventory in order to simplify production? If you can't do any of this you are probably looking at what is known as a sequential pull system that is unfortunately much harder to operate (more details on this later).

Key Points: Look for the “80/20 rule” on finished goods, sub-assembly components, and raw materials in a value stream. If the rule applies then it is easier to set up simple replenishment pull systems. If not a more complex sequential pull system may be required. It often helps to think in terms of lean “A,B,C logic” for components as well. A items are high runner items that should be made daily, B items weekly, and C items monthly for example (in some environments it might instead be weekly, monthly, quarterly). You might hold your A or B items in inventory and make the infrequent C items to order as needed.

2. How much of each item to hold in inventory?

Companies often fail to establish the right level of inventory to hold in market locations. (This same logic basically applies to finished goods, component inventory and raw material.) Even if this amount was once calculated properly, it is often not revised over time and can lead to problems and shortages downstream. Careful attention at the part level is required to review current demand and any known forecast information to set inventory. (Unfortunately this is not an exact science since perfect information is not available and if you have a gazillion part numbers it is difficult to do this manually.) For the items put in stock, some statistical method of estimating what is known as cycle stock, buffer stock, and safety stock needs to be properly calculated and put in place. Cycle stock is based upon actual historical demand, buffer stock is an estimate put in place to cover variability of the customer fluctuation, and safety stock is the amount required to make up for any process unreliability that might additionally be required (downtime, scrap, rework, etc.). If these effects are not properly taken into account then the internal logistical system often will not work regardless of the effort applied.

Key Points: Analyze a few part numbers in inventory (FG, component WIP, raw material) and determine if they are indeed held in the right quantities. See how often these items are reviewed and updated as well and if they contribute to delays in delivering parts to downstream customers. You may have to invest significant time in “right-sizing” your inventory markets before proceeding. Make sure the lead-time assumptions for purchased parts are correct as well.

3. How is material stored and organized?

There is no magic dust or formula here. In general companies which both overproduce and allow random storage of inventory are prone to problems in merely locating the right part even if it exists in raw material, component work-in-process (WIP), or finished goods. The rules of 5S and visual control were developed in Toyota to highlight and identify problems such as these. Proper workplace organization, parts storage, and visual control are critical enablers for logistical efficiency. Otherwise material handlers and shop floor coordinators spend endless amounts of time looking for parts rather than making and delivering them to the downstream customer.

Key Points: Apply good shop floor 5S principles and visual control techniques to inventory storage, tool storage, schedule information, and anything else critical. If there is an element critical to performance then devise a method so that abnormal conditions become visual and noticeable by the person responsible for the area.

4. Where to schedule the value stream and what type of pull system?

The answer to this question actually depends upon the nature of your production environment and what type of pull system best applies to your situation. High volume discrete parts typically operate under a replenishment pull system described earlier where finished goods inventory (or component level inventory) is held and replenished as the customer consumes it. In this case, the end of the line (typically some type of final assembly operation) is scheduled and everyone upstream reacts to this processes instruction.

In low-volume, high-mix shops or custom order environments the opposite is often true. The production instruction has to go upstream to an earlier point in the line and signal the need to make something. This part then flows downstream through succeeding processes and must arrive in assembly on time in order to satisfy the customer. Typically this latter type of pull system is called “sequential pull.” It is more difficult to operate because it approximates a make-to-order signal and there is no extra inventory to buffer the system; parts must be made right the first time and be delivered on time.

Furthermore, maintaining clear visual control over the first-in first-out (FIFO) sequence is extremely critical. The only buffer that exists in this type of pull system is the notion of a time buffer and releasing things slightly early (technically this is early production and a form of waste) in order to compensate for operational inefficiency, mistakes, or delays. Additionally, adjusting end customer delivery dates (rather than having a fixed delivery promise) is the only way to maintain a level load of work in the factory. Otherwise the factory production is only as level as the incoming orders.

My advice for sequential production scheduling shops is usually: to make crystal clear where parts are to flow through the shop and adhere to the principle of FIFO and small lot sizes. This of course means precise and standardized routing information must exist and spending effort on reducing changeover times. When possible organize equipment as well into a simpler flow pattern to minimize travel distance between stations.

In sequential pull you are typically “hostage” to supplied parts, both internal and external. Even more than in replenishment systems, the component lead-times must be understood and reliable. Otherwise promised deliver dates cannot be met. Typically in sequential pull there is no inventory to buffer against such problems. The only alternative is to buffer the system with extra “time” in the equation. However, this is a form of waste and the root cause still must be eliminated to shorten lead-times. If this was not complex enough, sequential pull systems require very high levels of first time quality and equipment uptime. Often these two areas must be significantly upgraded in order for sequential pull to work well. Given all this difficulty, most shops practicing pull production start with replenishment pull and then migrate to sequential pull over time as they become more proficient at TPS.

Many shops have to operate under both types of pull systems. In other words, there is a combination of replenishment pull (make-to-stock) and sequential pull (make-to-order). This is called a mixed pull system and unfortunately it means that more than one signaling mechanism needs to exist in a production line. For example, a replenishment trigger needs to exist for make-to-stock items and a special signal indicating make-to-order items needs to exist as well. Often these signals have to go to more than one location and unless they are thought through carefully then can “collide” at some mid-point in the value stream and force the process to decide what to run on its own. For simplicity it is nice to separate the make-to-order and make-to-stock items into separate production lines where possible. However this is not feasible due to the cost of equipment in many cases and production assets must be shared or put in a simple value stream.

Key Points: At the value-stream level the exact type of pull system and where the line is scheduled needs to be made crystal clear. For replenishment pull, it is often the end of the line. For sequential pull, it might be the first operation or somewhere in the middle. If a mixed system exists then how are the inevitable problems of mixed signals handled in the middle of the line? Create clear rules for which parts run first and how to run them must exist when there is a conflict.

5. How level can you produce at the pacemaker?

Production leveling consists of two dimensions – quantity and mix. Normally emphasis is only put on the former element as companies attempt to smooth the workload in the factory based upon some notion of capacity or revenue. In order for production systems to work most efficiently at the value stream level, however, both dimensions of quantity and mix need to be analyzed and considered carefully. For example, companies often think that parts are late from an upstream components shop due to scheduling delays

when in reality the upstream process is overburdened by the actual demand quantity placed upon the process. (Older traditional MRP systems assume infinite capacity). This situation can also result if leveling is only measured in terms of dollar amounts or quantity of units to produce in final assembly shops. The reality is that not all dollar amounts or work units are equal – especially in upstream operations. Some components have more work content associated with them than others. Fifty parts today might actually be the same as one hundred parts tomorrow if only half the work content is involved the second day. Revenue can be equally misleading at the component level. For this reason leveling should consider the amount of work content involved and how the pacemaker process (and of course down stream) is affected.

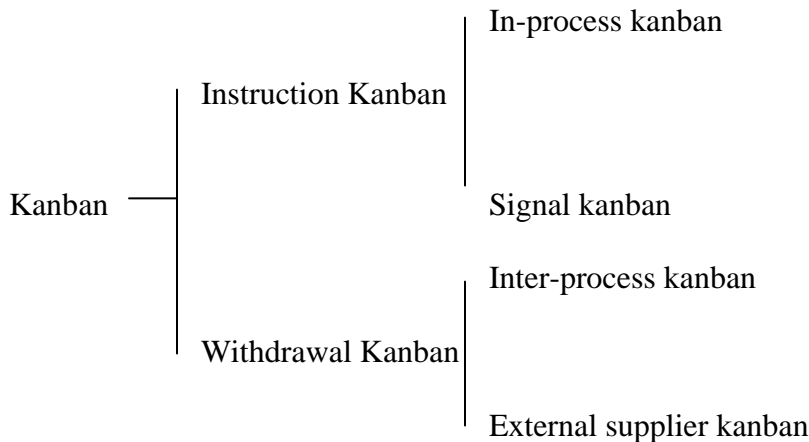
Beyond the issue of leveling *quantity* companies also often fail to figure out the right *mix* to produce at a pacemaker process. In other words, for the components made at the pacemaker process the right Every Part Every (EPE) Interval needs to be calculated. (See the Success Story: [Backbone of Lean in the Back Shops](#)). In general the more frequent the EPE interval the more likely you are to make the right mix of components and deliver the right parts on time downstream. In other words, increasing batch sizes conversely reduces the probability of having the right part at the right time. Combining the right EPE interval with better understanding of capacity at supplying processes will improve delivery performance. Unfortunately traditional EOQ logic drives an EPE interval that is optimal for some false notion of “product cost” and ignores the need and inherent changing requests of customers downstream. Toyota Motor Corporation figured this problem out in the 1950’s and since that time has worked for several decades on shortening setup and changeover times as well and designing machines differently in all of their manufacturing shops.

Key Points: Regardless of what type of operation or pull system you run, the EPE interval for the pacemaker process needs to be calculated and established. Some parts might be run on a daily basis, some on a weekly basis, others on a monthly basis and others as needed. However the more flexible the pacemaker processes (as well as any other batch machines in the value stream) the easier it is to deliver the right mix of parts on time. The price you have to pay to earn this benefit however is the challenge of quick changeover.

6. How to convey demand and what type of kanban to use?

It is important at all operations to be very clear about how demand information is conveyed and specifically what type of information signal is used. The word kanban is nothing more than the Japanese work for “sign board” or “signal”. Restaurants in Japan put out their “kanban” every day to indicate when they are open for business. Production processes often suffer from the problem of too many schedules floating around and thus no one puts much faith in one being accurate once the ink has dried. For reasons discussed earlier, it helps at the value-stream level if a single pacemaker exists for the area. Strive to avoid having multiple dispatch lists from a central location being issued and updated throughout the day. The unintended consequence of this system is often disharmony and less coordination than intended by well meaning parties.

Pull systems address this problem by designating a pacemaker process for the value stream. All other processes subjugate themselves to this process. While local efficiency is minimized total system efficiency is optimized instead. In practice there are different types of kanban that can be used for scheduling purposes. The major categories are as follows:



Instruction kanban are the category of kanban that tell a process “what to make.” There are in-process kanban that are geared towards making small lot or one-piece flow situations and signal kanban which are for inherently large lot or batch processes.

Alternatively, there are also withdrawal kanban that instruct “what to take” or in other words what to convey between a process and a market or a market and a supplier for example.

A kanban signal can be as simple as an empty space on the floor, a piece of paper, or a more complex electronic signal. It is merely a method to make a clear unambiguous signal to make or bring something to the next process. Depending upon the circumstances the right type of kanban needs to be designed and used accordingly. Using the wrong type of kanban or implementing it in the wrong way will not deliver results.

Key Points: It is not the intent of this document to describe the types and uses of kanban in detail. The real issue is how are you scheduling the parts of your entire value stream and how is material handling signaled to move items in a timely manner. If this flow of information does not occur in a stable and consistent manner then part shortages and delays will inevitably occur. This aspect of scheduling requires significant time and expertise to design a system that will reliably work and be trusted by all parties involved.

7. How will you control production between processes or departments?

In addition to the correct use of kanban, proper system design must also be placed upon the governing mechanisms between processes in order to avoid over production and

better synchronize production. In general there are two ways of governing this flow; adhering to a market mechanism or adhering to some form of FIFO.

In replenishment systems, or at least where there are many fairly standard items produced, a type of market mechanism can be used control flow between two points in a value stream. In this type of system, instructions to make and importantly instructions to stop making are governed by a supermarket of inventory that regulates flow. When some inventory is decreased and coupled with a kanban it acts as a signal to make some more of that item. Conversely, when the market area is full of a certain type it means to stop making that item. Although highly simplistic in concept a well designed market in conjunction with kanban is important to establish and can deliver significant results.

Sometimes between processes a market mechanism and kanban may not be the right answer. After all, markets hold inventory and it is desirable to produce with less inventory instead of more. Where possible a simple system of FIFO lanes can often be established as a way of regulating production between two points. This ensures that product is processed in the order it is received and instructs the following process what to make in a clear unambiguous fashion. In job shop environments where no supermarkets are feasible, FIFO lines are often the only way to clearly organize material flow between processes. In these environments FIFO lanes with designated spots of the floor at each process can be an effective technique for organizing what to make.

Key Points: FIFO lanes and market mechanisms often have to both be used in a value stream in order for it to flow as efficiently as possible. In order to make the lanes and markets work properly they have to be combined with the right amount of inventory (in the case of a market) and the right type of kanban in order to work correctly. 5S and visual control also play a role in making these simple items work since they do not work in isolation by themselves to govern production. Instead the FIFO lanes and markets are merely elements of a larger system that must be established (see next section on material handling as well).

8. How will you convey material between processes?

There are two simple ways to think about material handling. The normal case is a fixed time and unfixed quantity route. An alternative is the opposing fixed quantity unfixed time route. The easiest way to consider the differences between the two is by an analogy. A fixed time, unfixed quantity route is like a city bus driver on a standard route. The routes and pick up times for the route are known but exactly how many people will get on the bus is not known. Frequently this type of system works well in manufacturing for material handling especially in the case of purchased components being delivered throughout the shift to an assembly area. The material handler for a designated area can be assigned a specific route that forms a loop. The loop can be designed so that the material handler repeats it a fixed number of times per day (for example a one-hour loop would usually entail eight standard trips during the shift).

In this type of case establishing the standard route with a parts withdrawal kanban and the regular delivery times leads to defining the amount of material required at the point of

use location. Since any withdrawal kanban triggered on the last route should be delivered on the next route there is little reason to hold more inventory line side at the point of use than the hourly equivalent of the material handling cycle. If the material handling cycle is a fixed time route of one hour then one to two hours is all that needs to be held line side for safety. The rest should be stored in a market location and subjected to rules regarding withdrawal and signals to either suppliers or internal processes as required. Frequently this type of fixed time, unfixed quantity system is used with a dedicated material handling person utilizing tugger-style vehicles or other mechanized devices.

However, this application does not work perfectly in every case. Frequently parts are too large to be easily conveyed and lifted by a person. Some machines do not run every day due to low demand and are not staffed daily. Also regular and predictable frequency of need is not always well known in advance. In this instance a fixed quantity but unfixed time based system works better. Once again an analogy is useful to understand the difference. In this case, a city bus driver is not the right method for delivery. Instead a taxi driver “on call” is the right solution. The taxi driver waits until a specific call is made requiring his attention. In other words, a standard looping route is not necessary. The passenger calls and orders the pickup, telling the driver when to come and where to go. Until then, the driver did not have any advance knowledge of the situation. In some cases in manufacturing, this application works quite well. Rather than standardizing the route a defined number of material handlers (on fork lifts for example) are available “on call” for a signal to move material between locations instead.

Key points: Most companies pay too little attention to material handling when attempting pull or synchronized production. Material handling is like blood flowing through the human body paced by the heart. If the flow of blood in the body stops for any extended period of time the body will die. Manufacturing processes also starve in the same way for lack of material and poor pacing. For this reason material handling needs to be designated as either a fixed time, unfixed quantity or fixed quantity, unfixed time based system (each method has pros and cons). The key is to standardize the approach taken and ensure that a clear unambiguous signal to move the right part at the right time exists. Otherwise downstream processes will grind to a stop multiples times during the shift and destroy productivity with minor stoppages. (See the LEI workbook *Making Materials Flow* for details about implementing a material handling system.)

9. How to schedule batch processes?

Inability to properly schedule and get material through a batch processes is often a primary cause for part shortages. This is particularly true when there are certain constraint processes in a value stream that require production on some type of lot size for reasons of either quality or efficiency. The goal in these instances is to figure out the right lot size (ideally fairly small) to run, continually reduce the changeover time at the process, and enact the right type of scheduling method with the machines.

In general, there are three different ways to conceptually deal with a batch style production process. The machine may operate on a fixed sequence basis (but unfixed quantity of work), or it may operate on an unfixed sequence with an established lot size.

Regarding the latter case, there is the special case of signal or triangle kanban that can be established if applicable.

In order to ensure on-time delivery downstream from a batch process, you must understand the lead-time the process requires to deliver an item once it receives that signal. There are then specific logical assumptions that can be made to release production in time to arrive at the downstream process. Alternatively in replenishment environments, specific re-order point and re-order quantities can be established as well. (A specific example is provided in *Creating Level Pull* and a calculation spreadsheet is provided on www.artoflean.com in the documents section.)

Key Points: It is critical to first designate the type of scheduling method that will be used at a batch machine. Then accurate measurements and assumptions must be established for the lead-time to respond for each process in the value stream. Running batch machines on EOQ logic and not having balanced this with the right EPE interval for the value stream can lead to massive delays and forced schedule disruptions due to part shortages. It takes some expertise to schedule these machines and synchronize production accordingly – merely using rules of thumb will typically either result in delays or excess inventory.

10. How to implement and expand pull systems?

The typical causes for problems and suggested remedies in internal plant scheduling and logistics are highlighted in the previous nine questions. Addressing them one by one can often lead to improved internal on-time delivery performance, reductions in inventory, reductions in lead-time, and improvements in both direct and indirect worker productivity. Most people should start off with some sort of pilot line and a detailed value-stream map combined with further specific demand and inventory analysis needs to be executed. Each of the previous nine questions needs to be considered in light of the existing production environment and a proper future state scheduling system should be designed in concept.

There is only so much “book knowledge” one can accumulate on this topic and at some point “learning by doing” is the only way to master the topic and obtain results. The best way to learn is by implementation in a controlled pilot area that is representative enough in scale to encompass all the elements discussed in the previous questions. Inventory levels need to be defined and located, the right type of pull system needs to be designed, the mix and quantity produced at the pacemaker needs to be established, the right type of kanban must be implemented, and the right flow methods and material handling need to be created.

Once the above items are implemented in a pilot area there is still the need to roll out the system more broadly across the entire plant. Here again two different methods can be employed. One can either follow a value stream by value stream approach (customer specific approach) or roll things out department by department (process approach). There are valid cases for both arguments. The value-stream approach works well when there are relatively few value streams in question and processes can be easily designated to a

value stream (i.e. there are not many shared processes). Conversely when there are many value streams or many shared processes exist it is often easier to fix problematic areas department by department.

Key Points: First, make sure that you actually have managed to improve production in the pilot area. Don't rush forward and just try to implement tools as fast as possible. The goal is to achieve results that can be measured in terms of improved on-time delivery, inventory, and lead-time. If these show no signs of improvement, take time out and problem solve why. If improvement has been demonstrated, then make a plan on how to move forward. Either the vertical value-stream approach or the more horizontal approach is possible. I suggest that you begin where the greatest need to improve actually exists.

11. How to sustain level pull systems?

Level pull systems are fragile devices – by design they surface problems for organizations to observe and repair. If the organization does not react to the problems being surfaced, the system will not function in an optimal manner. There are several stereotypical reasons why level pull systems tend to break down.

First, there is often insufficient process stability to support the system. If there is massive downtime or quality problems in an operation, a pull system will not work well. It is worth time stabilizing these problems in order to achieve more consistent and predictable levels before attempting pull production. In other words stability is required before agility.

Secondly, there is a failure to establish and monitor critical process metrics after the system is established. If the system was designed with certain lead-time, quality, or downtime assumptions, then these items need to be monitored and tracked over time. If process stability deteriorates then the logistical system built around it will break down as well. Metrics for quality, downtime, changeover time, delivery, productivity, and other critical parameters need to be created and tracked over time for improvement.

Third, customer demand changes! Once a scheduling system based upon level pull production is set up it needs to be monitored and reviewed over time. Customer demand quantity and mix change over time based on the business environment. When a change in demand occurs, a corresponding change must occur in staffing levels, equipment capacity needs, inventory levels, and delivery from suppliers. A proper review interval and the key items to check needs to be established and ownership clarified. Otherwise the system will stop working and someone will inevitably say, "Well we tried pull production here once and it stopped working after a while."

Fourth, daily supervision is required to make the system run on a consistent basis. The internal workings of a level pull system are only as strong as the daily supervision governing the environment. On a daily basis problems will be surfaced that will require action. The ability in the supervisor ranks to spot problems and take action is critical to ensure success.

12. How to improve the level pull system?

How to improve the level pull system is difficult to address without specific data and situational information. The actual levers to improve differ case by case. In general, begin with the following check points to identify improvement opportunities.

- What is on time delivery to internal and external customers?
- What stops us from being at a higher on-time level than today? (Get facts not opinions.)
- What are our levels of raw, WIP, and finished goods inventory?
- How can we reduce the inventory level and maintain on-time delivery?
- What is the lead-time through from order to delivery?
- Where is the majority of the product lead-time spent and why? (Get facts not opinions.)
- How much time is lost looking for parts and why?
- How much downtime in final assembly is due to upstream part shortages?
- What is supplier on time delivery performance?
- What aspects of process stability (scrap, rework, downtime, etc.) can be improved?
- What is direct and indirect labor productivity? How can better delivery of materials impact this metric?

SUMMARY

The points contained in this document are intended as an aid for individuals attempting to improve internal aspects of manufacturing logistics specifically related to shop-floor scheduling, basic inventory control, and material handling in a value stream as starting points. This is not an exhaustive guide for all aspects of scheduling or logistics by any means. It is intended as a practical starting point for questioning some basic shop-floor practices and for identifying areas that typically cause problems in both low-volume, high-mix and high-volume, low -mix production settings.

For more information, refer to the [Creating Level Pull](#) workbook, the [Lean Enterprise Institute Library](#), and [Artoflean.com](#) .