It is a misconception, perhaps stemming from the pull system idea, that Toyota assembles vehicles in the same order in which customers buy them. Someday Toyota would like to have achieved that kind of 1x1 flexibility in its production operations (and also to have smoothened customer demand in the market). Today, however, Toyota
is at the point that it strives to run an intentionally levelized schedule in many of its assembly processes. The diagram in Figure 5-10 illustrates, in a simplified fashion, the basic mechanics of *heijunka*, or leveling an assembly process.

The schedule for the assembly process—in this case kanban cards that circulate when customers pull corresponding items from the finished goods inventory—is not sent directly to the assembly process. Instead, the cards are routed through a kind of sorter, depicted in the diagram as a box.

This sorter typically levels two things, the mix and the quantity:

1. **Leveling the mix.** The sorter rearranges customer orders (the kanban cards) into a predefined sequence by item type. The sequence could be selected, for example, to minimize total changeover time or to break up large batches of demand and spread them across the day. In the example, the predefined sequence is A→B→Z→E→D→F→G→H. Two additional slots in the example sequence box are left open for various low-volume items, which are ordered only occasionally.3

   The assembly process will try to produce items in this predefined sequence. The intended time to get through the entire sequence depends on the lot size. If, for example, the process can change over often enough in one day to make every type, called “every part every day,” then the lot size for any item is one day’s worth. In this case the process would try to get through the entire sequence each day, and begin again at the front of the sequence the next day.

2. **Leveling the quantity.** The sorter also defines for each item the maximum quantity of that item that should be produced on one pass through the sequence. This is based on the production lot size and the current customer demand rate for each item. If the lot size is one day, then on any one pass through the sequence the maximum quantity that the process should produce of any item is the average one day demand for that item. In the example, the average one-day quantity of item A = eight boxes, item B = seven boxes, item Z = nine boxes, and so on.
In the diagram, you can see that a customer has purchased eight boxes of Item B, pulling them out of inventory, which puts eight kanban cards for B in circulation. However, according to the leveling pattern (the sorter), the current average demand for item B is seven boxes and the assembly process should only produce a maximum of seven boxes of B before changing to item C. The eighth kanban card for item B should be filled on the next pass through the sequence.

This is where Toyota’s leveling efforts get counterintuitive from the process perspective. Imagine the assembly supervisor having eight kanban cards for item B in hand, the assembly process is currently making item B, all is running well, and now we are telling the supervisor that he should only produce seven boxes of item B and go on to item C.

Why do this?

Two already well-known reasons for leveling production in an assembly process are to be able to serve a variety of customers in a short lead time, and to limit the bullwhip effect, aka the “Forrester effect.” The latter states that any unevenness in assembly is increasingly amplified as the demand is transmitted to upstream processes. Since upstream processes must hold enough inventory to meet demand spikes, the amount of inventory—that is, lead time—in a value stream will be lower when the downstream assembly process operates in a level

![Figure 5-10. Example of a leveling scheme](image-url)
fashion. For this reason, leveling in assembly is often a prerequisite for introducing kanban to upstream processes, since without it the upstream kanban-system stores ("supermarkets") may have to hold an unacceptably large amount of inventory.

In the example, a customer has ordered eight boxes of item B, which is one box more than his current average daily demand rate. The assumptions are that over a period of time the customer will be buying the average quantity, and if the customer buys one extra box today, he will buy one less box in the near future. If the assembly process were to produce that extra box right away, it would send a demand spike upstream, which will be amplified and generate waste and extra cost the farther upstream it goes.

Smoothing production activities is the prevailing rationale behind leveling, or heijunka, but despite many attempts, I was never able to make heijunka work for very long. And neither can a lot of factories that I visit. I understood how to lay out the leveling sequence and lot sizes just like Toyota, but within a very short time we would have to deviate from the intended sequence because of problems, and we would quickly be back to frequent schedule changes, expediting, and firefighting. It seemed that Toyota must be experiencing fewer problems and practicing more stick-to-the-leveled-schedule discipline. But how?

The answer came to me because of two events.

When I paid my second visit to a factory in southern Germany, the production control manager met me with an angry face and the exclamation, “Please go away with that leveling concept!” On my first visit we had established a leveling sequence for one assembly process, but it did not last long, as usual. “We constantly experience part shortages,” the manager told me, “so if we try to stick to a predefined assembly sequence, we would lose valuable production capacity.” I had to agree.

The manager went on to show me the scheduling software program they had developed and were using instead. Every day, customer orders, inventory quantities, and parts availability are entered into the program, and from that information the next day’s assembly schedule is generated. “See,” the manager explained, “this is an assembly schedule
that we know we can run.” Of course, the assembly schedule was different every day.

The second event came later that same week. I was having dinner with a Toyota person, and at one point in our conversation he said, “Well, so many of the things we do at Toyota we do so there is a pattern.” The penny finally dropped for me at that moment.

**The Intention Behind Heijunka**

What the heijunka leveling sequence provides is a pattern, or in other words, a target condition. It is something to strive for; something that helps us see what we need to work on, and to focus our improvement efforts where they are needed. Here is how it works:

1. Load the leveling device, the sorter, with kanban according to the intended sequence and maximum lot size specification.
2. Ask, “Can we run this way today?”
3. If yes, do so. If no, ask, “What is preventing us?” Pursue one problem, and meanwhile temporarily go off the intended sequence. Strive to get back on the intended sequence as quickly as possible.

At the beginning, the answer to the question, “Can we run this way?” will be no more than yes. But if you do this over and over and tackle the obstacles one by one, the yeses will increase. What you are doing is improving the associated processes step by step in a systematic way—by leading people in a direction.

Now we can see that while the scheduling software at the German factory does ensure a feasible schedule every day, in doing so it works around problems and leaves the factory standing still rather than improving its processes.

“Ah, I see,” the German production control manager said when I explained what I had learned. “Toyota is trying to get to the point where the answer to the question, ‘Can we run this way?’ is always yes.”

In fact, that is not the case. At that point we will not yet have reached the ideal state, and there must still be waste in the system. If the answer to the question, “Can we run this leveling pattern?” is
almost always yes, then Toyota might, for example, reduce the lot size further in order to get back to a situation where the answer is no occasionally. Otherwise we are likely to stop improving.

One way to assess your efforts in leveling an assembly process is to measure sequence attainment per day. At one factory in the United States we set up a leveling sequence, and after one month the team proudly reported a sequence attainment of 73 percent. But I have never seen sequence attainment get so high so fast. We took a closer look and realized that the team was not measuring sequence attainment, but rather the old outcome metric—schedule attainment. That is, no matter how you did it, if today’s shipments went out on time, you have schedule attainment for that day. Sequence attainment is a tighter process metric, which means if the assembly process has to deviate from the intended leveling sequence, then even if shipments are still made on time, you do not have sequence attainment for that day.

The team recalculated and was crestfallen to find that their sequence attainment after one month was actually only 13 percent. But this is not a reason to be sad. It is simply the current situation, nothing more and nothing less. The only thing to think about is, “Okay, what is the first obstacle that we need to tackle?” By thinking this way, the team began to go after the obstacles to achieving a level pattern one at a time, with considerable enthusiasm and on their own initiative. It became a challenge. After one year the sequence attainment for this assembly process was in the 60 percent range and the team continued to work. Each step forward represented a true improvement for the factory, and people were adopting a new way of thinking. Not bad.

After almost 20 years of benchmarking Toyota, we have set up a lot of leveling boxes and schemes at various production processes in factories all over the world. It is revealing, however, to observe what happens when a senior manager visits the factory. The leveling boxes are cleaned up beforehand and put in perfect order. All the right kanban cards are in the proper slots, and someone explains to the guest how the leveling system works. The visitor asks a few
probing questions about the mechanics of the leveling scheme, eventually nods in approval, and then everyone moves on to the next stop on the facility tour.

What is happening here is a bit of a charade. Many leveling schemes are not working (which is, in fact, what one would expect, at least at the start) and are not actually in use, because we have misunderstood what they are for. The reality in many cases is that the assembly process still decides what to produce based on a schedule prepared anew every day. What the senior manager or executive who is being shown a leveling scheme should be asking is, “Okay, and what is currently your biggest obstacle to being able to run in this fashion?” There are a lot of heijunka boxes in our factories, but for the most part we are not yet using them as target conditions.

Contrary to what we might have thought, the heijunka pattern itself is not why production processes at Toyota factories run more level and more on time than in our factories. Establishing the heijunka pattern changes little in most cases. The point is how Toyota utilizes the heijunka pattern as a target condition to drive process improvement (Figure 5-11). It is the process improvement—the striving toward the target condition—that makes the difference.

Heijunka is one of the most far-reaching techniques in the Toyota toolkit, and a particularly useful target condition because pursuing it sheds light on so many elements of an assembly process and its associated value stream. Once we understand that heijunka, or leveling, is not a straitjacket, but a target condition, we can better reap the benefits of pursuing it.

**Figure 5-11.** The heijunka pattern as a target condition