CHAPTER 13

LEAN PRODUCTION AND QUEUES MANAGEMENT

Annotation

Queuing Theory (also called Theory of Queues) was founded in 1908 by Agner Erlang, a Danish engineer and statistician, to resolve mathematical problems for managing traffic calls to and from a rural automated telephone exchange.

Queuing Theory generated worldwide interest and gave strong impetus to the development of a number of important applied aspects of Mathematical Statistics. Queuing Theory was completed in general terms in the late 1930s.

This theory is a mathematical theory, but we are not interested in its mathematics. We are interested in its philosophy regarding the queues, waiting, and downtimes that are formed in every real production process and unbalance and slow it down.

Queuing Theory helps Industrial Logistics and Industrial Engineering. It also has an important place in Lean Production. The task of the Queuing Theory is to reduce queuing. A big Muda in any production are the queues, downtimes, and waiting.

The three main ways of shortening queues are considered:

1) Regulation of production capacities.
2) Disciplining queues of customer and production orders.
3) Shortening of times of operations.

Theory of queues defines various ways of disciplining the queues of customer and production orders, aimed to increase the efficient use of production resources.

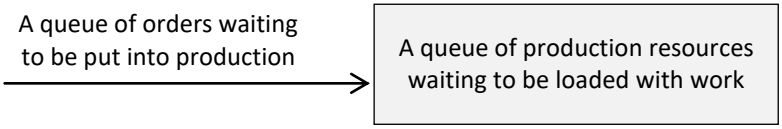
Useful guidance is also given on what to do to drastically shorten the various types of auxiliary times that always delay the work operations and processes.

Introductory Words

We will not need to define the word "queue". But we don't mean an animal tail…

Each production is characterised by two types of connected queues. On one hand, there is the queue of orders waiting to go into production. On the other hand, there is the queue of production resources, which have not yet been loaded and are waiting to be loaded. In both types of queues, there are losses from waiting.

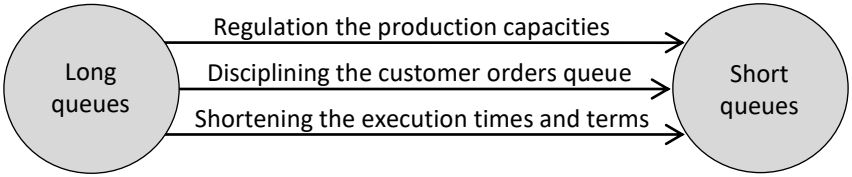
Two Types of Queues



13.01. The Main Task of Queuing Theory

The main task of Queuing Theory is to minimise the total losses of waiting, but not only in one or only in the other queue but in the two queues, taken together – as in the queue of orders, waiting to be put for execution in the production, also in the queue of production resources, waiting to be loaded with work on orders execution.

Which Way Should We Go...



The main task of Queuing Theory can be seen and solved in three different ways. These three ways are, on the one hand, independent, while on the other, they can perfectly complement each other. The first possible way is to regulate the production capacities to balance them against the intensity of the orders queue. The second way is to introduce "discipline" in the queue of customer orders. These are the rules of transformation for the sequence of customer orders in sequence of production orders. The third way is to shorten the execution times of operations and processes. In this way, we will shorten the orders fulfilment terms.

By adhering to these three ways, individually or jointly, we will reduce both the order queue and the queue of unladen production capacities. Explanations and details of the three different and independent ways are given in this chapter.

13.02. Production Capacities Regulation

First, we will talk about capacities regulation. There are a number of possibilities.

We can increase or decrease capacity or work with variable capacity, or secure some reserve capacities for our basic capacity, or attract additional capacities.

We can increase or decrease capacity. Greater capacity can be achieved by a higher number of workstations and/or by installing additional equipment, and/or a higher operating speed. Lower capacity can be achieved by fewer workstations and/or lower operating speed.

We can reduce the capacity. We bought a new machine, but it is of excessively high performance, regarding our needs. It is better to sell it and return the old machine, because we do not need a machine with such high productivity.

We can use variable capacity. I give an example with the number of working cash registers in the neighbourhood supermarket according to the number of buyers. By 4.30 p.m., only one cash desk is open. At 4.30 p.m., a second desk is included, at 7.00 p.m., all desks are open. After 8.30 p.m., only one cash desk remains open.

We can use reserve capacity. We receive a large order, and we assign part of this order to other companies – our partners or even competitors. In the mechanical and electro technical industries, this idea is only now becoming more accepted since we are still playing the game of being competitors. However, there are other industries that have learned the lesson the hard way, and that painful lesson has helped them wise up sooner. For example, in the clothing industry, when a large order comes, they do not hesitate to outsource a significant part of it to competitors, even though at other times they aim at them below the belt. But why should they miss out on a good order? Competitors can have a common interest.

Additional Capacity. In a five-star hotel in the town of Pravets in 2016, an international seminar on the Theory of Constraints was held with the participation of assistants of Dr Eliyahu Goldratt. The seminar was held in a hall on the second floor of the hotel. There were one hundred and twenty participants. They all went for a break at the same time. There were huge queues in front of the toilets. Surprisingly, there were very few who realised that the hotel, besides the second floor, has a first floor and a third, and a fourth floor, and that there are also toilets.

We often do not see these useful additional capacities wandering all around us. Unable to look beyond our own four walls, we are blind to another world around us.

13.02.01. Why Don't We Regulate the Demand...

Apart regulation of production capacities according to the intensity of demand, it's not a bad idea to try to smooth out uneven fluctuations in market demand. This, is how we try to influence demand to load our capacity more evenly. We have already seen it with the Heijunka dialogues (see Chapter 09, pages 266 and 267). Active communication and interaction between customers, traders, planners, and production executives are needed. Otherwise, there will be no Heijunka dialogue. Unfortunately, there are not a few companies that are forced to put up with being voiceless production appendages to an unscrupulous commerce bloodsucker.

Here's a question. Do we produce to sell, or do we sell to produce? In other words, should traders align their actions with the interests of production, or should production align itself with what contributes to the interests of traders? These are things worth thinking about and should be thinking about them more deeply.

What I'm about to say now runs the risk of some orthodox marketing professor giving me an E. The lion-company does not follow the market but adjusts the market for itself. The question is whether and to what extent the big company can influence the intensity and cycles of market demand. The answer is that it not only can but also does so. We see it everywhere, every day. Big companies do just that.

Given mentioned above another existential question arises. What can the smaller companies that are forced to work under the dictates of the larger companies do? The answer is easy to say, but it is not easy to do. A small company can become a big company someday only if it starts working like a real big company from today.

This applies not only to companies but also to real social life. If a small person does not behave like a big one, he will not grow up and will remain forever small.

13.03.01. Priorities of Customer and Production Orders

Let's recall that the main task of Queuing Theory was to minimise total waiting losses in the two queues – the queue of orders waiting to be put into production and the queue of unladen production capacities in anticipation of taking orders.

We said that there are three different ways to solve the main task of Queuing Theory. We already know the first of them – it consists of capacities regulation.

The second way we can solve the main task of Queuing Theory is to introduce discipline in the sequence of placing orders in production. This discipline is a kind of mechanism to transform company priorities into a sequence for placing the customer orders in production that fully and evenly loads production capacities.

There are many types of priorities. The big question is which priorities to choose. I will list some types of priorities and explain them further.

Priority may be based on the moment of acceptance. It is to convert customer orders into production orders in the sequence of their reception in the time.

The priority may take into account the time required for implementation. First, we start work on these orders which we will be able to fulfill in a shorter time.

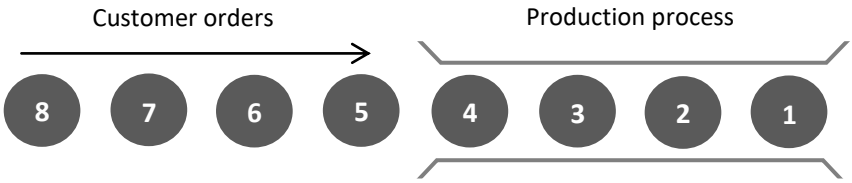
The priority can take into account the "price" of orders execution. We first place the orders that we will be able to fill with lower direct manufacturing costs.

The priority may also take into account the losses from delayed execution. We place these orders first, the delay of which would cause the greatest damages.

The priority may be related to our attitude towards the customer – with some customers we have a long-term contracts in which their orders are prioritised.

The priority may be influenced by the customer's persistent insistence that their order be executed as soon as possible. In cases where a customer cries and screams loudly, his order leaps over other orders and jumps forward in the queue.

13.03.02. Priority According to Acceptance Moments of Customer Orders



Two words about the priority, which we will call here First Come – First Served. In this case, the customer order first accepted is placed first as a production order.

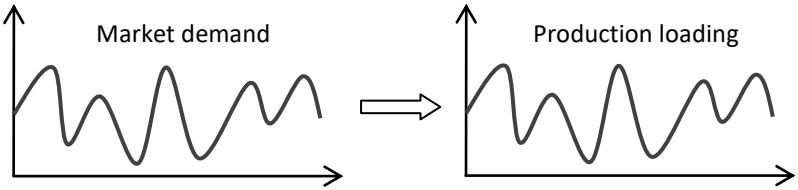
This is a special case, a special variety of the priority First In – First Out (FIFO).

FIFO is a commonly recommended order for move and processing of materials and products. There are even a number of stricter international and sector standards for quality management systems that impose the FIFO order as a mandatory order.

These are GMP (pharmaceuticals), ISO 13485 (medical devices), ISO 9100 (astronautics and aeronautics), IATF 16949 (automotive suppliers), and others.

Let's be careful! The problem is insidious...

In the case of uneven demand, the application of FIFO priority when transforming customer orders into production orders brutally shakes the entire production flow.



If we apply FIFO priority in the transformation of customer orders into production orders, the result is that as customer demand fluctuates (as input of production), so will production fluctuates. The best discipline for transforming customer orders into production orders is unlikely to be FIFO. This is especially true if we have not learned to execute the production orders in a significantly shorter time (much shorter than the deadlines for executing customer orders, which are established and acceptable to the customers in our business sector).

13.03.03. Priority by Order Execution Times

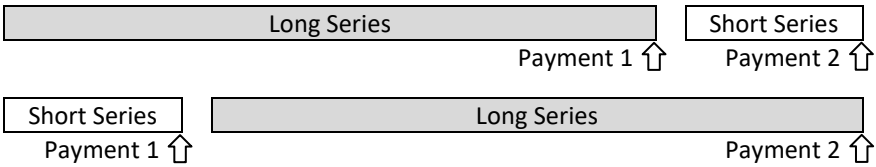
We first execute these orders that we will be able to complete in a shorter time. The orders that take more time are left to wait. This is almost the perfect priority from at least two points of view. On one hand, the total waiting time in the production order queue will be reduced. On the other hand, this priority will allow us to load our production capacities relatively evenly.

A dentist's office. There's a queue in front of the cabinet of patients who didn't book an appointment in advance, and there's no patient with an emergency. What's going to happen? The dentist comes out and starts asking, "What are you here for, what are you here for?". And no one patient is his friend. They're all equal. Who will the dentist invite to come in? The one he can serve most quickly.

Here we see an experienced production boss. He has a passion to aim and select slower for execution or non-urgent orders, and longer series. Such orders and series clog up and block the factory capacity, while at the same time, short series and urgent orders angrily knock on the door of traders and senior management.

The experienced production boss runs to the high-rank bosses and insistently cries for more people, more machines, more materials, and more areas to put out fires.

Priority of Short Series with Fast Payments

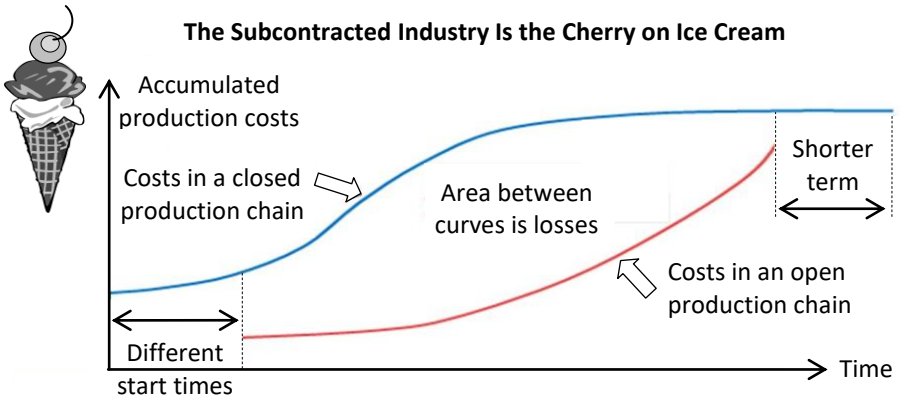


Prioritisation of customer and production orders according to the length of the series, respectively, according to the completion time there is another seductive advantage. First, we place into production the shortest series. And then we place the long series last. This allows us to receive a large part of the payments earlier.

13.03.04. Priority According to the Costs of Fulfilling the Orders

First, we place orders that we can fulfil at a lower cost. I remember a company, our client, received a request for a huge customer order. They prepare good offer, the offer was accepted but the company did not have enough money to cover the cost of executing the order. Fortunately, the customer was in no hurry for the order. This allowed the company to put into production other cheaper for execution orders, they took money from them and used this money to procure materials for the large order and fulfilled it. Otherwise, if they had started with the larger and more expensive order, they would not have had the resources to fulfil it. We are talking about the big factory Intelligent Security Systems in the town of Vratsa. Greedily gobbling up the biggest order first is an own-goal.

The Subcontracted Industry Is the Cherry on Ice Cream



The idea of prioritising the cheaper for execution orders also raises the question of whether large costs should be at the beginning or at the end of the production cycle. The two curves show two different ways of accumulating costs in two large lathe machine factories, ZMM Mashstroy in the town of Troyan and ZMM Sliven in the town of Sliven. ZMM Mashstroy has a completely closed production cycle. It also has a foundry and blacksmith shops. More than 80% of the production costs of the lathe are formed inside the factory. The top curve shows that the large costs are consumed at the beginning of the production cycle. ZMM Sliven is a factory of the same scale and with almost the same products, but its production logic is inversed. More than 80% of the direct production costs are generated by cooperating subcontractors. The lower curve shows that only towards the end of the production cycle the higher costs are consumed. I have to make a clarification that this was in the past years – at the beginning of this century. We see how the production costs are accumulated in the two factories. The area between the two curves is no costs but is losses. In previous years (ZMM Troyan factory no longer exists) production costs at ZMM Sliven were significantly lower, and the term to produce one lathe was significantly shorter. This is one of the main advantages of the open chain of production. Well yes! As long as there are stable partners...

13.03.05. Priority According to the Risks of Late Production

This priority is often applied in cases where late or delayed execution could result in unacceptable losses. For example, an unprofitable contract envisages severe fines for both non-execution and delay. Another example – the non-fulfilment or delay will be followed by long and difficult disputes and/or irreversible damages to the company's business image.

Similarly, is the priority logic for operations and orders which pose a risk of damage to materials if they were to remain in a state of waiting for a long time. The priority of orders which require personnel, workplaces, and equipment and where downtime is costly also has a similar logic. I don't know how managerial and technical personnel in mechanical engineering and similar industries would react if I said that the organisation of production should resemble the organisation for the production of perishable foods. Imagine a factory for perishable foods where there can be queues of unfinished products. Yes... they may have refrigerators, and the semi-finished products are stored in them waiting for their turn. Refrigerators are expensive, take up a lot of room, waste energy, you have to defrost them, etc. Now let's imagine the same factory, but without

the refrigerators in it – from raw material to final product, every operation is synchronised. There's an even rhythm; there are no intermediate stocks or queues in the way of the product, because if there are stocks and queues, the product will spoil. There is much to learn from the technological solutions in the food industry. People there know how to synchronise operations and how to shorten the cycle.

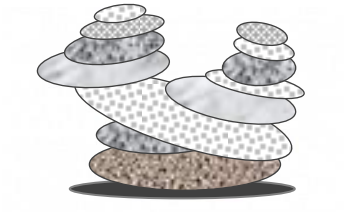
13.03.06. Prioritise Orders that Don't Shake Up the Flow

When we have a mixed range of old and new products or we produce in mixed series sizes, it sometimes (not always!) makes sense to prioritise the orders which do not shake up the flow. First, we place orders, in which we do not need to learn unknown processes, new materials, or other new production resources, and we do not need to change the logic of the technological sequence. Such orders do not require y1ou to make changes and readjustments of the equipment, they don't require tests and experiments or to validate new mastered materials and technologies. In other words, first we put into production those orders, where there is clarity in terms of the deadlines and the cost of resources needed, and where there are no technical and technological surprises lurking in wait for us. This orders priority will give us a sense of production comfort but this it can only be a short-term policy. Looking at the distant horizon, there are the risks of losing sight of commercial reactivity and developmental initiatives.

Here is an instructive example with Optix, a big optical factory in the town of Panagyurishte. One of the owners, who was also chief technologist, introduced the principle that up to 10% of orders with technological uncertainties should be scheduled for production during the intervals between the other 90% orders, for which the technology had been mastered (this would enable them to be produced cheaply, at a high quality, and on time). Ninety percent of orders are such that they know how to execute them. In parallel, they work a more focused manner on orders for which they miss technological clarity. The 10% of technologically unclear orders required a certain amount of risk taking. Sometimes some of these orders are almost losers, and there are even problems with its customers. It takes soft human skills to solve such always delicate.

As a result of nearly 25 years of consistently pursuing such a policy, there are no longer any technological secret for Optix, and they can produce everything. Even if they have unmastered processes, in the following years they will master them too.

Grouping of Orders

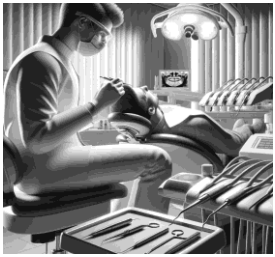


There are also some interesting priorities which depend on the possibility of grouping (by some criterion) a part of or all orders. This helps to load the production capacities and, at the same time, to smooth the flows. I will give examples of such type priorities.

13.03.07. Priority of the Technologically Validated Orders

First, we execute the orders with a known deadline. The orders with uncertain deadlines will wait. Once again, let me reiterate that a priority with such logic can be allowed only as an episodic policy – for example, in times of high market demand. But if we constantly rely on the policy to accept only orders that we know how to seamlessly execute cheaply, qualitatively, and on time, on the long horizon, this policy will come back to bite us. It will dull the agility and inventiveness of designers and will hinder the flexibility and reactivity of traders.

13.03.08. Grouping of Orders by Technological Similarities

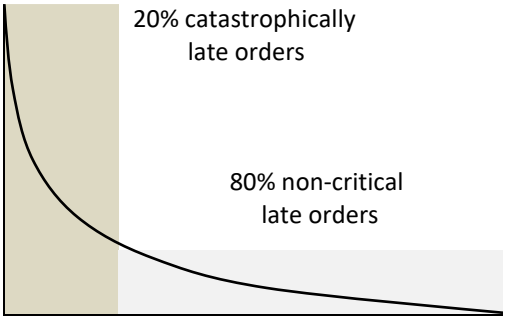


This priority is often applied when it is possible to group some of the orders by a similar mix of materials used and/or by similar equipment modes or settings. This priority aims to minimise the total readjustment time in a group of similar orders or to minimise readjustment time between sequences of different orders.

You call your favourite dentist and whine about a sudden, urgent problem, "My dear doctor, I've lost a filling!". The dentist grunts, "Let me see how I can accommodate you, I don't have any free appointments". He is silent for a while, he thinks, he pretend that turns over his notebook, keeps you on your toes, and says, "Come on Thursday". Not because there's free space in his schedule then, but because he only does fillings on Thursdays. The dentist's schedule looks like this – Monday takes measurements, Tuesday pulls teeth, Wednesday fixes dentures, Thursday does fillings, etc. For each day of the week, he groups related

operations – Monday, Tuesday, Wednesday... and puts you in his patient schedule where he is comfortable. That's one good way to do things. The dentist's schedule is convenient for him, not for you. We all know what the work table next to the dentist's chair looks like – it is narrow – 70 cm by 40 cm or something like that. This work zone is such a perfect 5S System that you will rarely see it anywhere else. If he has patients with different cases and different types of treatment, the dentist has to reorganise his workplace ten times a day. But no! He groups the "orders" according to similar materials and similar equipment adjustments.

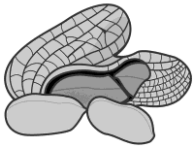
13.03.09. Anti-Pareto Priority



Vilfredo Pareto is an Italian sociologist, engineer, and economist known for his analysis that 80% of wealth is in the hands of 20% of the people. That was at the end of the 19th century. Now the distribution of wealth is even more deplorable. The Pareto Principle says that we need to separate the few important things from the many unimportant things, to take the bull by the horns, to attack the few important things, and to leave the large number of unimportant things for later.

Anti-Pareto principle states that if there are a large number of slightly late orders and a small number of critically stuck orders, we first deal with the large number of slightly late orders, with all the risk of full clogging up the critically late orders. As so, we might enrage some customers, even some customers will leave us, but there will be smaller total number dissatisfied customers. Pareto principle helps us to separate the small number of big problems from the large number of unimportant problems. But this does not always mean that we should first focus on solving the important problems. The Pareto Principle must not be perceived and applied mechanically. The problems are more or less interdependent. The small, unresolved problems sometimes get in the way of solving the big problems.

13.03.09.01. Priority with Elimination of the Critical Delays



Priority with eliminating the critical delays somewhat resembles Anti-Pareto priority. However, it also includes the idea of rapidly collecting ours money from urgent orders at the expense of a permissible delay in the deadlines for other less urgent orders.

Varna is the city of the densely clustered roasted peanut sellers. I love peanuts… There are a lot of small stalls for roasted peanut in the huge central city square.

I noticed crowds of people in front of some stalls, while other stalls were deserted.

I'm always in a hurry and go where there's no queue. It turns out I bought mouldy peanuts. The next day, in another stall without a queue – again mouldy peanuts…

I decided to find out why and in what places people tend to queue. I exchanged a professional conversation with a peanut seller who had a queue at his stall. He said he sells the fresh peanuts first. He always sells the most recent deliveries first.

He said there was a risk of some peanuts moulding and having to be thrown away, but he always sells fresh peanuts, and that's why the queue is in front of him.

At that time, we were working with a bakery that had company stores all over Sofia. We offered them an experiment – not to sell the bread in the order in which it was produced but to have the retailers follow the principle "The freshest bread should be sold first". They increased their sales by over 10-11%.

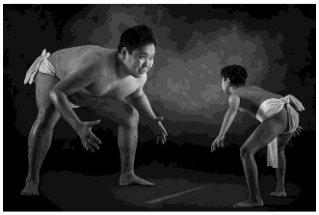
Sometimes priorities of customer orders don't seem as logical as we would expect.

Such a priority which leads to elimination of the critical delays is a delicate matter and can be disadvantageous, especially if the losses from spoilage of the product can be significant. Customers do not lose anything as a result of this priority.

The question that the frugal trader asks himself is whether losses from spoilage of the product are covered by the additional profit from increased sales volumes.

By implication, such a priority is mainly applicable to these products that are at risk of losing their properties if the time between production and sales is longer.

13.03.10. Priority of Small Orders



The prioritisation to some smaller orders (if they do not interfere with the large orders) is a policy that requires the ability to execute a large and appetising order in parts or with interruptions, thus releasing up free capacity for other orders, which may be smaller but also quite interesting.

On the other hand, there is a risk that large customers will drive away smaller customers. But some of small customers in the future they may grow up and become important. In the case of increased demand, we should protect ourselves from the appetite for large orders and thus attract a greater number of customers. With increased market demand and decreased supply, we observe two radically different business manners. Most companies are in a rush to profit from large orders and actually achieve high temporary profits, but by doing so, they drive away some of the smaller customers. Other companies are trying to protect and even expand their market foothold and they do this in two ways. The first way is to play with the prices and deadlines so as to get as many orders as possible. The second way is to announce that they supposedly have reserve capacities, thus attracting customers who have been possibly rejected by other companies.

In a small neighbourhood store, the kind saleswoman politely invites customers with small purchases to skip the queue and be the first to pay for their purchases. Why? Because the store is small and has no space for more customers, it will sell less goods. If there is a big queue at the cash register, there is no room for more customers to enter the store. In some hypermarkets, parallel to the main cash registers, there are also cash desks for single or small purchases. They take the pressure to the main cash registers, so the flow of customers flows quickly.

Let's continue by talking about the interesting priorities of fitting in small orders.

We're taking our car for a long major overhaul. The mechanic says it's going to take a long time. Two months the car is with him. The time needed for the real repair can't be more than ten hours. It's not just our car he's working on. He'll be working on at least fifteen or twenty other cars at the same time as our car.

Before I met Dr David Mossop, whom I thank for translating this book, I have worked with other translators. They were professionals, but was taking a month to translate a few pages. Like our car mechanic, they was taking on many clients at once so not to miss out any opportunity. That's how they lose clients. I hope one of those car mechanics reads these lines. The Bulgarian builders are the worst example of how work for three days can be dragged out for three months.

People pour mineral water from a spring into canisters and large bottles. If a child holding a small bottle begs to be allowed to fill it, the ones with the larger vessels will invite him either kindly or with a grimace of annoyance, some benevolently, some not so.

Insertion of small orders into a large order is a variety of Short Lead Time priority.

Insertion of small orders implies flexibility of resources. Example – fish restaurant only takes advance reservations. Before 9 p.m., many tables, signed "Reserved", are empty. This is a nearly 20% unoccupied resource. There are two solutions. One solution is a condition for an exact arrival time – if you're late, you lose the reservation. The other solution is to work with small tables, allowing regrouping and maintain up to 20% free capacity. In this way, the restaurant will fully utilise its capacity. I pointed this out to the head waiter, but he didn't understand me.

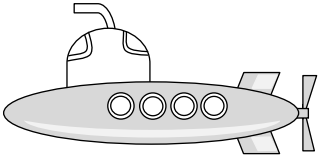
There is a small wine cellar in Brestnik, a village near Plovdiv City, where I buy wine. It's a magically dense and strong wine called Stanimashka Malaga. This is a small winery. It produces 15-20 kilolitres of wine, and there are a hundred customers like me who buy 100-150-200 litres. One impertinent merchant comes to the manageress and offers to buy all the wine production for the next three years.

The landlady (she has only primary education!), with the shrewd mind of a tired country woman, is pragmatic and answers him like this, "My dear! You'll buy all my wine for three years in advance, but how I can sure, that when the third year comes, you will buy from me again? By then, I will have lost all my clientele".

There's a kiosk in the neighbourhood where I buy cigarettes. I pass by and ask, "Have you got any of my cigarettes?". Answer, "I have, but only four packets". I said, "Give me all four". The salesman smiles, "I won't give you four, I'll give you one. I have other clients…".

In the case of cigarettes and similar type of orders, the smaller inserted orders may also acquire higher priorities according to the regularity of the customer. If we learn to produce efficiently in a short series and to deliberately divide the large orders into parts, we will also fulfil the large orders, and along with them, we will fit a big number of smaller orders. So, all our customers – both small and large – will be satisfied. Production planning skills help us be good salesmen.

13.03.11. The "Customer's Submarine" Priority



There is such a tale: He got his foot through the door, and now he's sleeping in my bedroom... Informal relationships can and should help, not hinder, good business. So, let's be cautious, especially when dealing with larger customers.

13.03.12. Increasing Priority of the Long Pending Orders

Ski lift. The skiers are lining up in the queue. The lift operators have friends who brazenly jump the queue. The lift operators have lots of friends. If they put only their friends on the ski lift, the main queue will never move. But lift operators are intelligent and resourceful people. They introduce an order – one friend, one from the queue, one friend, one from the queue... So, the two queues are served, one quickly and the other slowly. People in the main queue wait patiently for their turn to board the ski lift. Everyone is happy.

Increasing the priority of long-pending orders is a double-edged sword. One edge of the blade allows us to insert some short series into the pauses on longer series.

So, the number of dissatisfied customers is reduced, and our market presence is expanded. But if one of the skiers is fed up with other people jump the queue, he'll look for another ski lift. If he sees people standing in a long queue there too, he might even give up skiing. So, the other side of the blade, which has to be very good sharpened, is not to let an order wait too long since we risk losing the corresponding customers. I've said it once, but I'll say it again. If we have learned to work effectively in a short series and if we are able to divide orders, we will also take large orders, and we will fit also the small orders into the production.

13.03.13. Random Selection Priority

The priority of Random Selection may have a place, but rather in services and in retail but not in serious self-respecting industries. In industry, this priority (by the way, it is a special case of orders with increasing priorities) would create much more problems than comfort because there are all the risks of upsetting the flow if an urgent order stops and disrupts orders already placed for production.

13.03.14. What Determines the Discipline of the Queue?

Criteria for queue disciplining (individually or together) can be determined depending on the nature of the demand – on its intensity and fluctuations, as well as on the structure of the orders – in terms of product gamma and series lengths.

The discipline in the production order queue can account for possible clustering of homogeneous orders – once, this shortens the total change and readjustment time, and secondly, which is by no means unimportant, it facilitates the documental dispatching and physical removal of materials from the warehouses.

Formalised criteria for order prioritisation may be introduced. Order queue discipline may help you achieve higher yields on the utilisation of production resources. Finally, but only as an exception, order queue discipline may be dictated by some special motives of senior company management.

Order queue discipline may change over time if new or special circumstances arise, but it should not be determined or changed for emotional or subjective reasons. First, we need to clarify who sets the priorities. Secondly, it's a good idea for them to be documented. It should be quite clear who and under what conditions, has the right to ignore such priorities. At best, no one should. If priorities become obsolete and seemingly unsuitable, we can change them. But as long as the set priorities are valid, we must not ignore them.

13.04.01. Shortening of Operation Times

The main task of Queuing Theory is to minimise the total losses in two related queues – production orders queue and production resources queue. We looked at two ways of solving the task – by regulating capacities and by disciplining orders. We will consider the third way – by shortening the times of all operations.

Here total time is 120 seconds, of which 15 seconds is the time for the main operation – the chip removal, which makes 12.5%. There is only one really effective main operation (his time can be shortened by the technologists without much benefit from it) and many auxiliary operations (their times depend on the constructors of technological devices and on industrial engineers, not less).

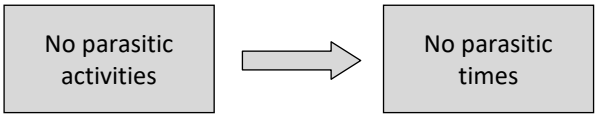
Operation Properties	Operation type	Time (sec)
Taking a blank	Auxiliary	5
Transferring to the lathe	Auxiliary	10
Loading the spindle	Auxiliary	15
Initial tightening	Auxiliary	5
Checking for coaxiality	Auxiliary	15
Tightening once more	Auxiliary	5
Bringing the knife	Auxiliary	5
Chip removal	Main	15
Dimensions check	Auxiliary	5
Loosening of the spindle	Auxiliary	5
Shooting the piece	Auxiliary	5
Transferring to the cart	Auxiliary	5
Stacking in a cart	Auxiliary	5
Data entry in the journal	Auxiliary	20

Now I will make an important distinction. The time of an operation is the sum of its effective time (time of execution of the operation itself) and its auxiliary times (preparatory times, times for results checks, times for adjustments, finishing times, times for readjustments, and any other times that do not add customer value). Let me say it again. When thinking about how to shorten the times of work operations, it is important to distinguish the main times from the auxiliary times. This is a lathe operator, next to him in the table are consecutive operations, and their times are rounded up to 5 seconds. Let's look at the numbers for the times. Notably, out of the total time of 120 seconds, only 15 seconds, or 12.5%, are effective time, i.e., the time for chip removal. I'll leave the question of how to shorten the chip removal time to the technologists. How will they shorten that time – with forced cutting modes, forced cooling, forced lubrication, and so on…

Whatever they do, the most they can do is reduce the time of the main operation from 15 seconds to 10 seconds. This is not a significant effect in the context of the 120 seconds total lead time to perform the entire operation from beginning to end.

The logic of what I'm saying is that the key reserves are not to be found in the effective time but in the auxiliary times. That must be the logic of thinking.

13.04.02. How to Get Rid of Parasitic Times?



Parasitic activities are those that are not at the core of production process. But we have to resort to them for the purpose of compensate for the consequences of some or other shortcomings of the technology and production organisation. Parasitic activities not only swallow up material resources but also time that can extend the total execution time of main operations and processes.

How to rid us parasitic times? There is nothing simpler! We perfect production technology and organisation. This eliminates parasitic activities, as a result of which there will be no parasitic times. Why would we perfect the production technology and organisation? Because they're imperfect, because when they were conceived, the main focus has been on costs and quality, but not on times, without thinking that the parasitic activities extend the local and total lead times.

13.04.03. Ways to Shorten the Operation Times

There are many ways of shortening operation times, it is enough just to list them. I have given more details about them in separate chapters of this book.

I list them. The personnel are competent, therefore skilled. Not that they work faster, but when they work they make fewer mistakes, because they understand what they are doing and why. Equipment is high-performant with high working speed. Materials are technologically tested for short processing times. Devices and tools are suitably adapted to the nature of the work. Production, control, and logistic technologies are verified and stabilised. Work conditions are ergonomic and comply with the requirements for the work environment. Operative management is simplified. Procedures, documents, and records are few. Last but not least, projects directly aimed at improvements and shortening times.

13.04.04. Squeezed Production Factors

All the materials and technological consumables have a high degree of readiness for use. The work tools and technological devices are designed with the idea of being installed and removed easily and quickly. When technological cost limits are determined, it is also taken into account the times of the inherent auxiliary and additional operations.

The organisation of the production process and the topology of equipment and workplaces bring workers physically and quickly closer to the objects of their work.

13.04.05. More Tools for Shortening Times

All the elements used in the work are arranged in FIFO order but formulated in words different from the original, and in particular First Needed – First Served or First Needed – At The Front. This model of arrangement shortens the times for searching, finding, taking, placing, and using these elements.

The workplaces are organised according to the first three S's of the 5S System. Sorting! (1S). There is no lack of anything necessary, but there is nothing superfluous for the actual work to be done. Set in order! (2S). Everything is arranged conveniently and ergonomically and has quick and safe access. Shine! (3S). The workplace and everything around it have been cleaned and shined. For workers, this has a stimulating effect and reduces inattention errors.

When changing and readjusting the equipment (machines, energy, tools, technological devices, etc.), all applicable SMED organisational ideas and technical solutions are used in order to reduce the times. Not only there, but also when uploading and downloading, when inputting and outputting materials, in the maintenance of equipment, and in many other important places and work activities.

The materials, products, tools, devices, documents, etc., needed for the work, are brought to the workplace on time and are removed as soon as the work with them is finished. The same is true for every production order and series. This is to ensure that there is everything you need in the workplace and nothing more.

A longer series is purposely divided into shorter series so as not to clutter the workplace with things that will not be used immediately. This saves work space. In addition, surpluses beget conditions for inattention errors and labour accidents.

When preparing for a given operation, checks are made for the availability and suitability of all the elements necessary for it. So, you won't have to waste time looking for something you need but it's not there, to correct something that's unsuitable for the work, or to look for a place to remove something we don't need at this moment in this workplace.

Before starting to work another order or batch we need to remove everything from the workplace that we don't need. This may be all kinds of materials and products, consumables, technological devices, work tools, packages, transport means, measuring and control means, technical work documentation, records, remnants from the previous batches, and anything else that would prevent us from working without wasting time.

13.04.06. The Important Role of Technological Devices

The technological equipment is designed for fast and error-proof working with it.

This is important assurance for shortening times, especially for manual operations.

In assembly industries and in manufacturing industries, this is very much true for the technological equipment for uploading, positioning, fixing, and downloading.

In the chemical, thermal, and varnish painting processes, there are technological tools for submission and withdrawal, removing and arranging group processing.

Although that are quite good technical solutions, there is still much to be desired.

Here an example from one ammunition factory. A plastic part shaped and sized like a melon. With a flat file, they remove the fallout from injection moulding.

This is done on a smooth desktop without using any means for positioning and fixing. Worker struggles to press the part against the desktop so it doesn't move, but nevertheless, it keeps turning by itself to the wrong for the next sawing position.

Solution is to have a worktop with indentation like the form of a melon where the part can be placed and where you can file it without needing the juggler talent.

Here one more example. A worker solders miniature electronic components to the upper surface of a control board. The circuit board has a flat bottom, and the soldering is carried out on a smooth horizontal top. While it is being soldered, the board moves here and there.

The worker holds the board in the left hand while, at the same time, keeping it away from the hot soldering iron. It would be a piece of cake to attach a sheet of felt to the counter to stop the board from moving.

In a number of factories, technologists are indebted to the workers for such technological devices which help ease labour and shorten the times of operations.

And if the role of the devices is more or less understood for the main operations, it is also very important to understand it for the auxiliary operations as well.

Technologists in quite a few industrial enterprises need serious additional training to rethink the powerful potential of technological equipment and to contribute not only to the proper execution of operations but also to shorten their times.

13.04.07. Shortening the Total Time by Process Decomposition

Wash and Dry

10 + 5 — Fifteen minutes by car, four cars per hour

10 → 5 — Ten minutes by car, six cars per hour

Combined operations and workplaces lead to local savings (which is not always true) but may also lead to a deterioration of the throughput of the whole process.

A Bulgarian man goes abroad to work, earns some money, returns to his hometown of Cabrovo and decides to invest his earnings into a profitable car

The owner is pragmatic and faces a dilemma. Should he buy a combined single-chamber installation that washes and dries? Or buy installation with two separate chambers – one washes, then the car passes to the other chamber for drying.

Combined washing-drying installation is cheaper, plus it occupies a smaller area and so more terrain will be freed up for more cars to wait calmly in a long queue.

In the other installation, where one chamber washes and the other one dries, we see what the picture tells us. In the combined chamber, washing takes ten minutes and drying takes five minutes. This is a total of fifteen minutes per car, that is, four cars per hour. If there are two separate chambers, we have ten minutes per car for the two operations, and that's six cars per hour.

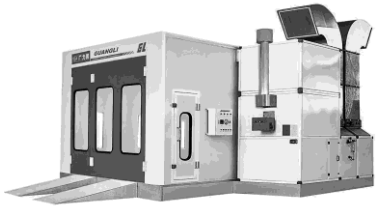
After a year, the business acquired a new dimension and developed to two washing chambers with a common dryer. The capacity has doubled – twelve cars per hour.

Bravo to our countryman! He has a flair for the non-trivial. What at first glance seemed like an attractive decision, then turned out to be an unfortunate decision.

Anyone can succumb to fraudulent delusions. Things are not always as they seem at first glance. See more on the topic on page 545 of Chapter 17 on Lean Thinking.

Painting and Kilning

Here's a similar case, like the one with the car wash. A factory for cooking stoves which was about to buy a high-productivity painting chamber, also had two possible solutions – one chamber where the paint is applied and baked, or two separate painting and baking chambers. The numbers indicate the solution.



10
+
20

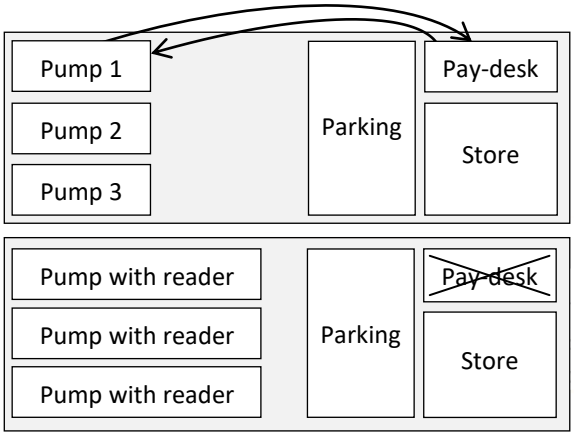
Thirty minutes per piece,
two pieces per hour

10 → 20
10 → 20

Ten minutes per piece,
six pieces per hour

Car washes and painting chambers have nothing in common. But the task is similar. Let's repeat that combined workplaces lead to local savings but can make the throughput of the entire process worse. A hint – the interesting experience is not always in our branch but in another branch that has nothing to do with ours.

Reducing the Total Time Spent at a Gasoline Station



Total time =
charging time
+ pump-cashier route
+ payment time
+ cashier-poump route

Total time =
charging time
+ payment time

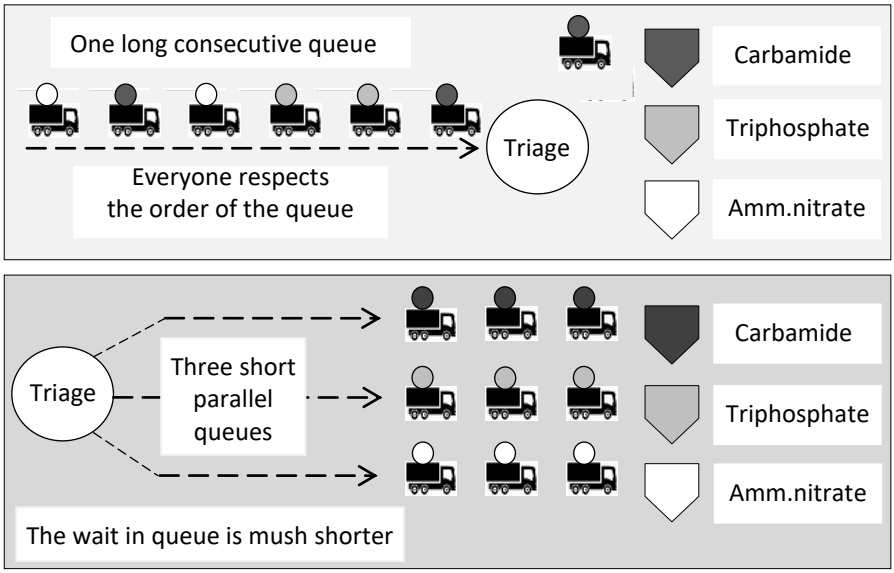
In today's Bulgarian petrol stations, the servicing process goes like this. You stop at the pump, you leave the car to be filled up, you go to the cashier, you pay, you come back, you pick up the car, you park, and you sit in the cafeteria to have tea.

There are petrol stations which are set up differently. We still don't see them in Bulgaria. The pumps are equipped with a card reader. You put the payment card in, fill the tank, remove the card, park, and sit down to drink tea. The process goes in one direction, there is no going back. You don't have to waste time running to go to the cashier and come back again. The overall service process is simple and fast. A larger number of cars can pass through a petrol station organised in this way.

13.05.01. Relieving the Queue of Orders

Another way to shorten the waiting time is to lighten the order queue itself.

Possible solution to lighten the queue of customer orders is to redirect some of the orders to other divisions of the company (if any), to entrust them to partners or subcontractors, or even to submit them to our loved competitors (why not!?).
A second solution is to temporarily stop or delay some orders if they are blocking a large production resource or are lacking in the necessary technological preparation. But these can only be orders that will not suffer from the delay, i.e., orders whose customers would be tolerant enough to accept a longer deadline.



Nitrogen fertilisers factory produces carbamide, triphosphate, and ammonium nitrate. There are three separate silos for the three types of fertilisers. The three silos have a common entrance. The fertilisers are loaded on lorries from the silos.

The lorries, regardless of which silo they are loading from, wait in order of arrival. In the picture, the circles above the lorries are white, grey, or black according to the type of fertiliser they will load – black for carbamide, grey for triphosphate, and white for ammonium nitrate. You can see a lorry waiting for its turn to load carbamide, but the corresponding silo is busy, and this lorry creates a long queue behind it. The silos for triphosphate and ammonium nitrate are free, but the lorries waiting for them are stuck in the queue and cannot be positioned under the silos.

The solution is to have a triage point 300 metres before the silos and also have separate entrances for the three silos. There will be three independent, parallel, and shorter queues, and as soon as one lorry is loaded, the next one comes.

13.05.02. Slowing Down the Flow of Orders

There are a few more non-trivial opportunities to artificially slow down the flow of customer orders... as long as we can afford to use these opportunities.

We break the flow of orders – for example, by entering fixed dates for accepting inquiries, returning offers, and accepting orders. Prices are in function of the delivery time – for example, by applying price discounts for orders with a longer deadline and by increasing prices for orders with a shorter deadline. We discount for a fixed delivery time – not as a fixed length of time but as fixed calendar dates or days in the week or month. For example – we ship only on Tuesdays or only in the first week of the month. We introduce of formalities for the acceptance of the order – for example, the requirement to coordinate the request and to confirm the offer. We delay communication with the customer. For example, we introduce the practice of interim clarifications during the fulfilment of the order, or we require the customer to participate in its final control and/or in its shipment.

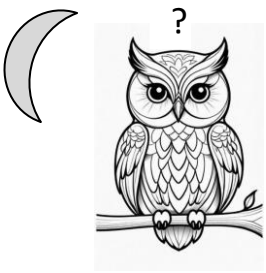
It's understood, these cheeky tricks can only be afforded by a powerful producer, before whom the queues are piling up. But some of the ideas can be borrowed.

Production Comfort and Market Adaptability

Catalogue-type production is "calm production", but it is economically justified only if pre-shipment stocks are moderately large. In the long run, it can dull market reactivity and development initiative. Compared to this, the Make-to-order type production is accompanied by difficulties in planning and operative management but keeps product development and technological innovation alive.

It also keeps the senses of the marketing and commercial functions and activities in heightened condition. Let it not be taken as a dogma but as if the proportion of 80/20 between Make-to-Catalog production and Make-to-Order production is a good pragmatic ratio between production comfort and market adaptability.

13.06. Who Is Waiting for Whom in Production?!



The main question is who is waiting for whom in the production process. There are two sides to question. One side of the question is whether the product is waiting for the work operation, or it is the work operation waiting for the product. The other side of the same question is whether the worker is waiting for the work tool, or it is the work tool waiting for the worker.

In the spectacular construction of reservoirs in the Far East in the 1950s, three poorly paid workers queued for an expensive pickaxe.

Nowadays, in a modern West European factory, three cheap machines not working while they were waiting for an expensive paid worker to come and work on them.

There are already such factories in Bulgarian industries. It all depends on what's more expensive – the cost of human manual labour or the cost of machine time.

The general answer is that cheaper queue will wait.

This is to reduce losses in the more expensive queue. It's all right to wait in such a queue, where relatively smaller losses are incurred. Where the losses are greater, no waiting should be allowed.

The hastily conclusion that the cheaper queue has to wait is not always true and correct from the Lean point of view. There are Lean tools aimed at reducing losses in product and order queues. From a Lean point of view, it is worth having unloaded resources if this helps to smoothen, compact, and speed up the flow.

Again, from this point of view, unladen production resources mean that there should be no waiting in the queue of customer orders. Isn't that where the money comes from? The main purpose of Lean ideas and tools is to constantly and always maximise the throughput of the entire production, seen in its fullest potential.

Conclusion to Chapter 13
Lean Production and Queues management

Remains is to draw a conclusion about the queues management. We see the Danube River in the spring. It flows quickly with a large and even flow. Like the Danube, which is full of water in the spring, this is how our production should flow.

That's attractive and seductive! And it is very likely that we will be tempted to try to over-pump our production in order to increase and accelerate the flow even more. Then there is a risk that the high flowing, dense, and uniform laminar current will become turbulent – with breaks, rapids, and eddies. I don't hink a chapter on queue management is the right place to say it, but on the one hand, you suffer losses from waiting in queues, while on the other, serving queues constantly is exhausting. In the sense of Lean's ideas, efficiency does not mean production resources that are loaded constantly and to their limit but rather a good product that does not wait in queues on its way but flows calmly and quickly in an evenly flowing production current. Over-pumping the production, even if we allow it sometimes, can only be a short-term policy with measured risks. It is unreasonable for this to become a regular practice – it leads to accelerated wear of hardware resources and, above all, to exhaustion and demotivation of people. The occasional and temporary high efficiency, which then fades, is exhausting. What we need is constant and steadily increasing efficiency.

Discussion questions, homework tasks, practical assignment, and exercises

Discussion questions

What are the ways to shorten the times of auxiliary operations?

Are there any queues in production that cause other queues?

Which production units have the longest queues?

What has been done so far to reduce queues?

Homework tasks

Together with colleagues, calculate total product waiting time
and compare it with the total net technological time.

Together with colleagues, assess which ways to shorten times
of operations are really applicable in the conditions of your company.

Identify together with colleagues which of these ways you will apply first.

Practical assignment

Together with colleagues, identify critically long and/or chronic queues
in one production site and identify measures to shorten or avoid them.

Create a plan for implementing these measures
within a short, medium, and long-term horizon.

Define measurable indicators to evaluate
the effectiveness of these measures.

Exercises

See two practical exercises on page 744 and the answers on pages 754 and 755