Lecture Notes on Operations Management

Christian Reinboth

23.09.2013
Wernigerode, Germany

These lecture notes were taken during the 2013 installment of the MOOC 'An Introduction to Operations Management' as taught by Prof. Dr. Christian Terwiesch at the Wharton Business School of the University of Pennsylvania via Coursera.org.

Contents

1 Week one

1.1 The four levels of performance

A look at various types of business operations - from a sandwich restaurant to a hospital - reveals, that the typical customer evaluates business processes based on **four basic levels of performance**.

(1) Cost

How efficiently can the business operation deliver goods and / or services?
(More efficient businesses can offer goods and / or services at lower prices)

(2) Variety

Can the business fulfil the specific wishes of their heterogenous customer base?
(Most customers do not care about variety itself, but want their wishes to be met)

(3) Quality

Performance quality = How good is the product and / or service offered?
Conformance quality = Is the product and / or service as good as advertised?

(4) Time / Responsiveness

How fast can the wishes of the customer be fulfilled?

There are **trade-offs** between the four dimensions of performance, e.g. the fastest service with the highest quality can usually not be offered at the lowest price in comparison to competitors. These trade-offs can be used by businesses to create unique business strategies (e.g. cost leadership) as well as to distinguish themselves from their competitors.

The tools of Operations Management can be used to help businesses to reach decisions about such trade-offs and to evaluate measures to overcome inefficiencies. One way to detect such inefficiencies is to compare your own business with those of your competitors - and especially those competitors, who are positioned on the **efficiency frontier** with no other company being pareto-dominant on a combination of two out of the four dimensions (e.g. no other company is cheaper and faster or faster and offering more variability).

1.2 Flow rate / troughput, inventory and flow time

The three most important performance measures of a business process are flow rate / throughput, inventory and flow time. In the following definitions, the term **'flow unit'** will be used a lot. A flow unit is the basic unit of analysis in any given scenario (customer, sandwich, phone call etc.).

Flow rate / throughput: The number of flow units (e.g. customers, money, produced goods/services) going through the business process per unit time, e.g served customers per hour or produced parts per minute. The flow rate usually is an average rate.

Flow time: The amount of time a flow unit spends in a business process from beginning to end, also known as the total processing time. If there is more than one path through the process, the flow time is equivalent to the length of the longest path.

Inventory: The number of flow units that are currently handled by a business process, e.g. the number of customers in a store, the number of enrolled students in an university etc. pp.

It should be kept in mind that the definition of inventory in Operations Management is different from the definition used in accounting. While the number of bottles on stock qualifies as inventory in both Operations Management and accounting, the number of patients waiting at a dentists office would not be seen as inventory in accounting - but is, in fact, inventory in Operations Management.

1.3 Capacity, bottleneck, process capacity, flow rate and utilization

In order to perform the following calculations, processing time has to be defined as the time that is spent on a certain task (e.g. one station in a sandwich restaurant). We will also need the previously introduced definitions of flow rate and flow time.

Capacity: The capacity can be calculated for every station in a business process. It is always m / processing time with m being the number of resources (e.g. workers) being devoted to the station. If, for example, one worker needs 40 seconds to put together a sandwich, the capacity of this station is 1/40 per second or 1,5 sandwiches per minute. If there are two workers on the same station, the capacity increases to 2/40 per second or 3 sandwiches per minute.

Bottleneck: The bottleneck is defined as the process step (station) in the flow diagram with the lowest capacity (the 'weakest link'). Although the bottleneck is often the process step with the longest processing time, it is important to always look at the capacities for making a judgement.

Process capacity: The process capacity is always equivalent to the capacity of the bottleneck. It is useful, to calculate a comprehensible number, such as customers per hour or parts per day (instead of a hard to comprehend number such as 1/40 customer per second or 1/345 part per second).

Flow rate: Even though the flow rate was previously defined, the definition needs to be augmented as the flow rate being the minimum of demand and process capacity. While the flow rate logically can never be higher than the capacity of the bottleneck, it can very well be lower, if the demand is insufficient.

Utilization: The utilization tells us, how well a resource is being used. It is calculated as flow rate divided by capacity (e.g. 1/40 / 1/25). The utilization always lies between 0% and 100%.

1.4 Labor content, cycle time and idle time

Cycle time, labor content and idle time are indicators for assessing the productivity of a process.

Cycle time: The cycle time is defined as the time between the output of two successive flow units (e.g. the time between two served customers or two treated patients). It is always equivalent to the time of the longest process step.

Total labor content: The total labor content is defined as the time sum of all process steps. If, for example, a process consists of two steps each claiming 20 seconds, the total labor content is 40 seconds.

Idle time: The idle time is defined as cycle time minus processing time. The idle time thus tells us for how long a resource (e.g. a worker) is not able to do anything, because he has to wait for another resource. If, for example, one worker in a sandwich restaurant prepares sandwiches while another operates the register, the second worker has to wait for a sandwich to be finished in order to collect on the customer. If the demand is maxed out, the idle time at the bottleneck is always 0.

Total idle time: The total idle time is the time sum of all idle time within a process.

1.5 Average labor utilization and cost of direct labor

Two important performance measures are the average labor utilization and the cost of direct labor.

Average labor utilization: The average labor utilization is defined as the total labor content divided by the sum of labor content and total idle time. If, for example, the total labor content is 30 minutes and the total idle time is 10 minutes, the average labor utilization is 30 / 40 = 0,75 = 75%. The average labor utilization thus tells us the overall performance or productivity of the process.

Cost of direct labor: The cost of direct labor is defined as the total wages per unit of time divided by the flow rate per unit of time. It tells us, how many Dollars (or Euros) are being spent in order to get one flow unit through a process (e.g. to treat one patient or to serve one customer).

Looking at the direct labor costs is very important, even though labor costs seem to make up - at least at first glance - only a tiny part of the overall costs. But since labor costs are hidden in all supplies and materials a company buys (that is, the labor costs of the suppliers), a company that might on the sheet be only paying for materials does in fact pay for externalized labor.

1.6 Little's law

Little's law was named after the American professor John Little (1950s). It defines the relationship between the inventory, the flow rate and the flow time, who have all been already defined previously.

inventory = number of flow units in the process
flow rate = rate at which flow units are being processed
flow time = time a single flow unit spends in the process

*Little's law: inventory (I) = flow Rate (R) * flow Time (T)*

Little's law is important, because it can help us calculate one of the three variables. Once two of the variables are known, the third one is set by the law. This also means that, form the standpoint of an executive, two variables can be picked by management while the third one then falls into place.

1.7 Inventory turns

One other important indicator for the evaluation of business organisations (not processes) is the inventory turnover rate (or number of inventory turns). The turnover rate answers the question of how much time a Dollar (or Euro) bill actually spends inside an organisation with the organisation being seen as a black box of sorts. A company has a competitive advantage if it can turn its inventory faster then the competitors can.

inventory turns = cost of goods sold (COGS) / average inventory

The inventory turnover rate is especially important because inventory creates costs. A business has to invest money in order to produce or buy an item and store it over a period of time, especially if the item might also loose value over time. The important indicator here are the inventory costs per unit, which are calculated as average inventory costs over inventory turns per unit time. If, for example, the annual inventory costs are 30% and there are 6 inventory turns per year, the per unit inventory costs are at 5%, meaning that for every unit sold the company has to calculate 5% inventory costs sort of as an internal tax rate. Inventory turns are therefore a good indicator on how productive an organisation uses its capital.

1.8 Why do companies hold inventory?

There are five major reasons for holding inventory:

(1) Pipeline inventory

A pipeline inventory is the minimum inventory an organisation needs in order to function. E.g. a producer of wine that needs to age for two years in order to be sold needs a minimum inventory of wine for two years in order to exist.

(2) Seasonal inventory

A seasonal inventory is helpful, if an organisation wants to produce at a constant (cost-efficient) capacity, yet the demand varies with the seasons. E.g. a toy company can cheaply produce at at steady pace and build up a seasonal inventory for higher sales during the Christmas holidays.

(3) Cycle inventory

A cycle inventory is helpful if keeping an inventory saves costs associated with buying supplies on time. A private household will, for example, keep a box of water bottles in the cellar for practical reasons instead of satisfying the demand for water at the store every time it comes up again.

(4) Safety inventory

A safety inventory is a buffer agains high external demand, e.g. a burger chain keeping an inventory of pre-made burgers so that customers can be satisfied immediately. Safety inventories are closely associated with the meme 'buffer or suffer', meaning that if a process is not able to buffer for variabilities (such as an unexpected external demand) it will loose on flow rate.

(5) Decoupling inventory

Whereas the safety inventory can be seen as the buffer against heightened external demand, the decoupling inventory can be seen as the buffer against heightened internal demand. Such an inventory decouples supply from demand and supports a higher (and steadier) flow rate.

1.9 Finding the bottleneck in processes with multiple flow units

Since many flow units (e.g. customers in a shop, patients in a hospital) do not at all need the very same treatment or service, business processes have to serve different needs. Processing times thus might be quite different for each flow unit as well, even the path of flow units through the process might differ. For such processes with so-called **multiple flow units**, we need new tools to figure out which process step can bee seen as the bottleneck.

Finding the bottleneck in a multiple flow unit scenario takes four easy steps:

Step 1: Determine the capacity of each resource in the process (m / activity time or units per hour)
Step 2: Calculate the service demand for each of these resources considering multiple flow units
Step 3: Add multiple unit demand for every resource to calculate total demand for the resource
Step 4: Divide the total demand by the capacity of the resource to calculate the **implied utilization**

The highest implied utilization indicates the bottleneck process step. To differentiate between the implied utilization and the 'regular' utilization, its best to keep this important difference in mind:

utilization = flow rate / capacity (always between 0% and 100%)
implied utilization = demand / capacity (can very well exceed 100%)

The same calculation can also be done by looking at the required work time:

Step 1: Determine the work time capacity of each resource (e.g. 2 workers x 60 min = 120 min)
Step 2: Calculate the demand for work time at each resource considering multiple flow units
Step 3: Add multiple unit demand for every resource to calculate total demand for the resource
Step 4: Divide the total workload by the available time to calculate the implied utilization

1.10 Finding the bottleneck in processes with attrition loss

A process with multiple steps might have an attrition loss (of flow units) on every step. Take, for example, a process in which 300 people apply for a job opening and go through a four-step process:

Step A: 300 people apply per mail
Step B: Out of those, 100 people are invited (100/300)
Step C: Out of those, 20 people can make an internship (20/100)
Step D: One of the people doing the internship is hired (1/20)

Unlike the previously analysed processes, not every flow unit makes it to the end of the process (in this example, actually just one flow unit comes through). In this case, it would be misleading to just look at the overall capacity of each resource involved to determine the bottleneck. Instead, the following three steps need to be taken:

Step 1: Determine the capacity of each resource in the process (m / activity time or units per hour)
Step 2: Calculate the service demand for each of these resources considering the drop-out of units
Step 3: Divide the total workload by the available time to calculate the implied utilization

The bottleneck is the resource with the highest implied utilization, which does - again - not need to be the resource with the lowest capacity.

2 Week two

2.1 What is productivity?

A very basic definition of **productivity** - the average measure for the efficiency of a production process - would just be the **ratio between process output units and process input units**. The labor productivity might, for example, be four output units (such as car or netbook parts) per labor hour.

If one does not focus on a specific form of process input units (such as labor hours), but instead takes all kinds of input units into account (such as materials, energy, labor etc.), we are talking about the so-called **multifactor productivity**. Since both process input units and process output units are often difficult to measure and compare, fiscal measurements such as revenue and costs can be used instead. Fiscal measurements are also needed for comparing different input units (work time, materials) in the already mentioned multifactor scenarios.

These considerations lead to some basic mathematical expressions:

productivity = output / input

labor productivity = output / labor time

transformation efficiency = output / capacity

multifactor productivity = output (in $) / (capital + labor + materials + services + energy) (in $)

The two main drivers that reduce productivity are **waste** and **inefficiencies**. Inefficiencies and waste can be seen as the distance between a company and the efficiency frontier.

2.2 The seven main sources of waste

During the first week, the concept of idle time was introduced. In general, idle time is the time during which a resource is not fully used because of low demand or bottleneck restraints, which, in turn, reduces productivity. Idle time is, however, not the only (and not even the biggest) source of waste. Taiichi Ohno - the father of the Toyota Production System - once stated that 'moving is not really working', which basically means that **a loss of productivity does not necessarily need to be connected to idle time**. While, for example, the re-configuration of a machine or the movement of materials is clearly 'working time', those activities are not directly adding value to the process.

The **seven main sources of waste** are:

(1) Overproduction: Producing sooner or in greater quantities than the customer actually demands. Overproduction leads to higher storage costs and material losses (e.g. because food products must be thrown away after a certain shelf time) and reduces the number of inventory turns. The solution to overproduction is to match supply and demand as closely as possible.

(2) Transportation: Unnecessary movement of people and / or materials between process steps. Transportation causes costs and production delays and thus often reduces the overall efficiency of a process (e.g. peeling crabs caught in the North Sea in Africa before shipping them back to Europe). The best way to reduce transportation waste is to optimize the process layout ('short ways').

(3) Rework: Correction processes within the main process. Rework is always the result of a failure to do something right in the first place. Reworking something creates double labor costs while not actually producing more, thus productivity is reduced. It also requires holding additional resources just for reworking, so that normal production processes are not disrupted by reworking processes. The solution lies in - very simply put - making it right the first time as wells as in analysing and subsequently eliminating sources of failure.

(4) Over-processing: Doing more than the customer requires, e.g. keeping a patient in the hospital for a longer period of time than medically absolutely necessary. Over-processing is a form of waste that might have positive reasons such as workers having higher standards than the customers or employees being overly proud of their work. Still, over-processing adds to waste and thus to the reduction of productivity. One possible solution to this problem lies in the definition of quality and process standards, to which all employees then have to adhere.

(5) Motion: Unnecessary movement of people or parts within a process (similar to no. 2 but on a smaller scale), e.g. an office worker having to switch rooms every couple of minutes because the files are stored in different cabinets. One solution to this problem lies in optimizing the layout of workspaces (ergonomic design) in order to minimize unnecessary motion.

(6) Inventory: Too many flow units in the system. Inventory has already been discussed at length during the first week. It is by far the biggest source of waste in most companies and costs a lot of money because inventory requires costly storage space and drives down inventory turns. Possible solutions include 'on time' production schemes and the reduction of so-called 'comfort stocks'.

(7) Waiting: Employees or resources just sitting around waiting for something to happen, e.g. for a meeting to start, a customer to call or a customer waiting to be served. Per definition, waiting is an unproductive use of time. Waiting waste includes idle time as well as long flow times.

While the first five sources of waste are resource-centred (What did the employee do? Did it add value to the process?), the last two sources of waste are really an expression of Little's law. Some scholars point out that waste of intellect (ignoring the ideas of employees for problem solving and process improvement because of hierarchical concerns) can very well be seen as an additional eighth source of waste. A production system that tries to reduce all types of waste at once is commonly called **lean production**.

2.3 Finance and productivity

For a for-profit organisation, the overall target is not and can not possibly be increasing productivity or optimising processes - it is making money. However, understanding the value-adding operations has to be a core goal for every business executive. This is especially the case when operations are constrained by the bottleneck (and not by low demand). In those cases, even minimal operational improvements at the bottleneck can have a very high impact on the revenue.

2.4 Overall equipment effectiveness framework

The **term overall equipment effectiveness framework** (OEEF / OEE framework) refers to a method for evaluating the productivity of machines and other resources. In a first step, all the time that is available on a certain resource is added up. During the following step, the waste times (idle time, time spent on failures, maintenance time etc.) are identified and also added. By subtracting the wasted time from the total time, one can then calculate the productive time of the resource. The **ratio between productive time and available time** is the overall equipment effectiveness of the resource. The OEE framework was first introduced by the consulting company McKinsey.

When calculating the OEE framework, we can differentiate between three types of waste times:

Speed losses: idle time, reduced speed
Downtime losses: breakdowns, changeovers
Quality losses: defects, restarts, operator faults

2.5 Takt time, target manpower and line balancing

The expression of **takt time** is derived from the German word 'Takt' for the pace of a piece of music. In Operations Management, the takt time is defined as the maximum time it can take to produce a unit in order to keep up with demand. If, for example, the demand is at 6 units per minute, the takt time would be at 10 seconds per unit.

Once the takt time has been determined, the target manpower can be calculated as the ratio of total labor content (the time sum of all worksteps needed to produce one flow unit) and takt time. The **target manpower** is the number of employees (or other resources) needed to operate the process at takt time if the work could be divided evenly between them. Since this is an idealized calculation, the outcome will most likely not correspond exactly with reality, one reason being that it is not always possible to evenly split work between employees if different skills are needed. Still, calculating the target manpower provides a good starting point for line balancing.

Line balancing is the process of dividing an existing workload as evenly as possible between all of the available resources in order to increase overall productivity. If, for example, one workstation has a lot of idle time, maybe it can take over some of the workload from the workstation at the bottleneck. The basic line balancing procedure consists of four steps:

(1) Calculate the takt time
(2) Assign tasks in a way that keeps all processing times below the takt time
(3) Make sure that all tasks are assigned
(4) Minimize the number of workers needed

The general ability of an organisation to adjust its capacity and scale it up and down in order to adjust to changes in demand (the so-called staffing to demand) is an important form of flexibility, which can be achieved e.g. with temp workers or overtime work. A special form of line balancing is the so-called **dynamic line balancing**, which includes walking around the workstations during production, looking for pileups and rearranging resources in order to keep up with demand.

2.6 Advantages of standardization

One common mistake in process optimisation is to just look at average processing times as well as at idle times. Instead, it is also important to evaluate the time differences between employees in order to find examples for **best practice approaches**. Those approaches can then be looked at while composing standards. The goal here is to figure out how exactly the top-performing employees achieve their results - and to find a way in which other employees learn something from them.

3 Week three

3.1 Basic forms of product variety

There are three different **basic forms of product variety**:

(1) Fit variety: Customers need to be able to buy different versions (sizes, shapes etc.) of a product if the product is to be of use for them (personal utility maximization). The more the characteristics of a product move away from the customer specifications, the less use it has for that customer. This kind of variability is also known as horizontal differentiation. Examples are different sizes of shoes or t-shirts, different locations of shops or airports and different departure times of trains or planes.

(2) Performance-based variety: Companies might sometimes offer products of more or less quality (e.g. a 'high end' product and a 'standard' product), so that customers can buy according to their quality needs and / or their financial abilities (price discrimination). This kind of variability is also known as vertical differentiation. Examples are computers with different processor speeds, mobile phones with different weights and diamond earrings with different diamond sizes.

(3) Taste-based variety: Customers also want products to come in different versions appealing to their personal taste in colour, design, sapidity etc. This kind of variability is the outcome of rather 'rugged' individualistic utility functions with local optima and no clear common thread.

A company may therefore choose to aim for more variety for one of the following reasons:

- Their heterogeneous customer base does not accept 'one size fits all'-products (taste-based variety)
- They want to make use of price discrimination mechanisms and offer different quality versions for different income groups (performance-based variety, market segmentation)
- Their customers actively demand more variety (e.g. by wanting to be offered a broad range of foods in a restaurant as opposed to being offered the same food every day)
- Saturating a niche and thus preventing competitors from being active in that niche
- Avoiding a direct price competition with competitors by product differentiation

3.2 The effect of set-up times

Set-up times often have a significant effect on the performance of a process and can even determine a process bottleneck. The most important definition here is that of the batch: A batch is the number of flow units, which are produced between two set-ups. To calculate the capacity of a process with respect to the batch size, the following formula is needed:

*capacity = batch size / (set-up time + batch size * time per unit)*

Note, that this is the capacity of the process for producing a batch. For example, if the batch size happens to be 10 flow unites per minute, this calculation answers the question of how long it will take to produce one complete batch. This is a deviation from the previous definition of capacity, which did not take the batch size into account and was simply calculated as:

capacity = number of resources / processing time

The capacity can by this basic definition be calculated for every station in a process. It is always m / processing time with m being the number of resources (e.g. workers) being devoted to this process step. If, for example, one worker needs 40 seconds to put together a sandwich, the capacity of this station is 1/40 per second or 1,5 sandwiches per minute. If there are two workers on the same station, the capacity increases to 2/40 per second or 3 sandwiches per minute.

Usually, the larger the batch, the more efficient the whole production process becomes (**economics of scale**). Companies with custom-made batches are therefore trying to get their customers to order large batches (sometimes even forcing them). The bigger a batch grows, the more irrelevant the set-up time becomes with the process capacity getting closer and closer to the original definition of m / processing time. This is because the processing time is less and less determined by the set-up time with larger batch sizes. Thus, set-ups reduce capacity - and therefore, companies have an incentive to aim for such large batches. However, large batches also increase inventory - with all of the negative consequences (e.g. storage costs, ageing of shelved products etc.).

3.3 Advantages of mixed-model strategies

Since larger batch sizes lead to (the need for) more inventory, the batch size has to be balanced and chosen with both the positive (greater capacity) and negative (larger inventories) results in mind. A strategy, in which smaller batch sizes (down to a size of just one flow unit) are chosen, is called a **mixed-model strategy**. Since smaller batch sizes have a negative impact on capacity because of the set-up time, reducing the set-up time is an important enabler for running a mixed-model strategy.

3.4 Re-definition of the batch size in accordance with demand

The batch size was previously defined as the number of flow units that are produced between two set-ups. While this definition is correct, it does not take into account the actual demand for the flow units. If a process is able to produce multiple flow units (e.g. cheeseburgers and veggie sandwiches) with one set-up time in between, a batch in a mixed-model production is re-defined as a number of mixed flow units produced during a certain amount of time (before the pattern of production is repeated). The additional set-up times for switching between the flow units during the production of the batch have, of course, to be recognized.

This brings us to the following formula:

*target flow = batch size / (set-up time + batch size * processing rate)*

Here, the **target flow** is defined as the number of flow units needed per time frame in order to stay on top of the demand (e.g. 100 units per hour). The **processing rate** is determined by the bottleneck of the process or by the demand while set-up time and batch size have previously been defined.

If the goal is **determining the ideal batch size**, the formula can be resolved for the batch size. The result has to be set in ratio to the demand for the various flow units within the batch in order to find out how many flow units of each type are produced within the ideal batch size. Note, that the set-up time needed to start the production pattern at the beginning is part of the overall set-up time and thus needs to be included in the total sum of set-up times needed for this calculation.

Obviously, the batches will become larger and the inventory will become bigger the more set-ups are necessary as long as the overall demand does not change (but is simply spread out over more product choices). Variety thus leads to more set-ups and thus to more inventory, which is one of the biggest problems associated with offering more variety.

3.5 How to pick the optimal batch size?

If the batch size is so small, that the process step with the set-up time (assuming, that there is only one) becomes the bottleneck of the process, the process looses on overall efficiency. Thus, the batch size needs to be chosen in a way that assures, that it will not generate a new bottleneck.

If, however, the batch size is too big, any increase in capacity at the station with the set-up time (assuming, again, that there is only one) is ultimately of no use, because it will only lead to inventory piling up somewhere else in the process - wherever the bottleneck may be.

This goes to show, that the **ideal batch size** is one, in which the station with the set-up time has a **processing time which is just identical to the process bottleneck**. Only then will the batch size not lead to the creation of a new bottleneck or additional inventory pile-up. This is calculated as:

capacity determined by the batch size = capacity of the bottleneck

$b \,/\, (s + b * p) = m \,/\, p$

with:

b = batch size
s = set-up time
p = processing time
m = number of resources

3.6 How does product variety affect distribution systems?

The more the demand is divided into smaller and smaller segments, the harder it is to predict. Once there is more product variety, demand is going to become variable as well. Thus, statistical indicators such as **mean** and **standard deviation** are needed to cope with the so-called **variability of demand**. If demand streams are combined, the standard deviation of the combined demand goes up slower than the standard deviation of the previously uncombined demands. Such an aggregation of demand is called **demand pooling** and is an important method for reducing statistical uncertainty. Reductions in uncertainty can also be achieved through variance reduction or stock building.

3.7 Shortening set-up times in order to increase flexibility

Every set-up process can be broken up in **external and internal activities**. External activities can be completed while the station in need of the set-up is still running. Such activities can be moved up front, e.g. preparing patients outside of the operating room while another patient is still being operated on inside. If set-up times can also be improved through other activities (standardization, process optimisation etc.), flow unit types can be changed more often and flexibility is increased. This idea lies at the heart of the mixed-model production, which also profits from configuring a network of production facilities for pooling in order to be more flexible in catering to demand.

3.8 Optimizing the design of business processes

Since the design phase of a process largely determines the later production costs, the question of how to reduce the negative effects of variety on process performance by clever process design is becoming more and more important.

One successful method of improving process design is the so-called **delayed differentiation**. This method allots keeping as many process steps identical as possible for all products before splitting the production process up. The decision, which variant of a certain product is actually produced is thus delayed until the very last possible point in the process. This process design is optimal for products that differ only mildly (e.g. t-shirts of different colour). Delayed differentiation is made easy, if variable elements of the product can be isolated (the so-called **moduled product design**).

An interesting example for delayed differentiation is the casing around the (otherwise completely identical) iPhones. The hype over the iPhone also shows, that even hugely successful products do not necessarily need to offer a lot of variation. The reason for this is, that customers can also be overwhelmed by too much choice. To understand this, one has to understand that most customers only care about the characteristics e.g. of a computer that provide utility for them - not about the actual technical specifications (e.g. gaming performance compared to the actual graphic card of a computer). Companies thus might want to think about limiting their product variety in order to not make customers nervous by offering too much choice and keeping them from purchasing a product.

4 Week four

4.1 The concept of responsiveness

Responsiveness is the ability of a system or process to complete tasks within a given time frame. E.g. how quick can a business respond to customer demands? If customers are made to wait, they are turned into inventory, potentially resulting in a unpleasant customer experience. Any customer waiting time is also an indicator of a mismatch between supply and demand.

Concepts for solving waiting time problems can include increasing the capacity of the resource at the bottleneck as well as increasing process flexibility in order to ensure, that capacity is available at the right time. It has, however, to be kept in mind that waiting times are most often not driven by either the capacity or the flexibility of a process but rather by variability. Variability in the process flow (e.g. customers arriving at random) can lead to unwanted waiting times even when the implied utilization is clearly below 100%. If analysis builds solely on averages and fails to consider process variability, it can thus be wrongfully concluded that there is no waiting time, when, in fact, there is.

To solve this problem, new analysis methods are needed when dealing with process variability. It is noteworthy, that those methods are only requisite when a process has more capacity than demand - if demand exceeds capacity, it can be safely concluded that there will be waiting time even without looking at the process variability.

4.2 Variability in demand and processing

Variability in demand

If there is more demand than capacity, the implied utilization rate rises above 100%, which makes waiting time unavoidable. The more interesting cases are those, in which there is waiting time even though the implied utilization rate is below 100%. Such waiting time stems from demand variability generated by the somewhat **random nature of most demand processes**, e.g. many customers showing up at once at some point in time and no customers showing up at all at some other point in time.

In order to calculate with **demand variability**, we need to define **arrival time**, **inter-arrival time** and **average inter-arrival time**. The self-explanatory arrival time is defined as the time, when customers arrive at a business. The inter-arrival time is thus defined as the time between subsequent customer arrivals. If demand is random, both the arrival times and the inter-arrival times will be drawn from an underlying statistical distribution (often a Poisson distribution). The average inter-arrival time is usually denoted with a.

Another important parameter is the **coefficient of variation** of the arrival time, which is calculated as the **standard deviation over the mean** and is denoted as Cv_a. The coefficient of variation is a way to measure the standard deviation against the mean of the distribution. This is useful, because the standard deviation itself is not really a good measure for variability, since it does not express whether a 10 minute deviation is a lot or not.

If the inter-arrival times are drawn from an Poisson distribution, the coefficient of variation is always 1. This knowledge can be used to calculate other parameters considering this formula:

Cv_a = standard deviation (of inter-arrival times) / mean (of inter-arrival times) = 1

Variability in processing

Variability is not limited to the demand process, but also occurs in processing. The calculations are basically the same, with the average processing time being denoted with p and the coefficient of variation being denoted with Cv_p. The coefficient of variation can be seen as a measure of the degree to which a work process has been standardized.

4.3 Calculation of the time in the queue / waiting time

The **time in the queue** is calculated as follows:

$$\textit{time in queue} = \textit{activity time} * (\textit{utilization} / 1 - \textit{utilization}) * ((Cv\_a^2 + Cv\_p^2) / 2)$$

with:

activity time = service time factor (average processing time p)
(utilization / 1 - utilization) = utilization factor
((Cv_a^2 + Cv_p^2) / 2) = variability factor

Since the utilization factor is calculated as (u / 1 - u), this means that as the waiting time gets higher, the utilization gets closer and closer to 1. The formula for the time in the queue always delivers an average value. This so-called waiting time formula can only be used if the demand is lower than the capacity. If the demand is higher than the capacity, the waiting time will ultimately not be driven by variability, but rather by insufficient capacity.

The **total flow time** of a flow object can be calculated as:

total flow time = time in the queue + processing time

4.4 Calculation of the time in the queue with multiple resources

The previous session was based on the idea, that there was only one resource (one machine, one worker, one physician) doing all the work (m = 1). But what happens if the capacity is determined by more than one resource? **Calculating waiting times with multiple resources involved** makes it possible to derive **staffing plans** - or to, in other words, answer the highly important question of how many resources will need to be put to work in order to meet a waiting time requirement.

The time in the queue with multiple resources is calculated as follows:

*time in queue (for multiple m) = (activity time / m) * (utilization$\hat{(}$square ((2(m-1))-1) / 1- utilization)) * ((Cv_a^2 + Cv_p^2) / 2) = p / m * (u $\hat{(}$square ((2*(m-1)-1) / 1-u) * ((Cv_a^2 + Cv_p^2) / 2)*

If the time in the queue is known, Little's law allows the calculation of the inventory:

*inventory in queue = flow rate in queue (= 1 /a) * time in queue*
*inventory in process = utilization (u) * number of resources (m)*
inventory in total = inventory in queue + inventory in process

Devising a staffing plan

How many employees will it take to keep the average waiting time for a certain service under a minute? A simple way of answering such a question and coming up with a staffing plan is doing the calculation of the time in the queue and to then manipulate the number of employees until a certain average waiting time is met. When seasonal demand (**seasonality**) is to be observed, the calculation has to be redone for every time slice of the day, week or month in consideration.

4.5 Seasonal demand

In practice, demand sometimes exhibits seasonal ups and downs with spikes in demand at certain busy hours, days or weeks. It would be misleading to just ignore those spikes and assume that the demand during the spikes is drawn from the same statistical function as the rest of the demand. In those cases, the analysed timeframe has to be sliced up into equal time intervals and every interval has to be taken as the basis for a separate calculation.

4.6 Reducing waiting time by pooling demand

By **pooling demand**, the inter-arrival times are shortened and thus the specific demand goes up (which is intuitive, since pooling demand basically means combining different demand streams). While the utilization rate is not effected by demand pooling, the waiting time is shortened because some inefficiencies (idle time at station A while station B is overwhelmed) are eradicated. However, pooling more and more resources together also decreases the overall efficiency once the demand is met. Therefore, **companies need to find a viable balance between efficiency and responsiveness**.

What main benefits and costs are connected with pooling in the context of waiting time?

- Pooling assumes total flexibility (but Spanish call center agents will not be able to answer to German customers, even if the call center company decided to pool all calls together).
- Pooling increases the complexity of the workflow, since demands needs to be shifted between resources who might be locally apart (e.g. two hospitals or two plants).
- Pooling interrupts the continuity of interaction between the flow unit (customer) and the resource (worker) and can thus hurt the customer experience because customers will not want to see a different physician or a different financial consultant on every separate visit.

4.7 Order of customer service

All previous contemplations were based on the idea that customers would be served in exactly the order in which they arrived. But is that an optimal procedure? The idea behind **sequencing** (which is also called triaging or prioritizing) is to determine, who should be served first, which is not always the customer who came in first (e.g. emergencies in a hospital). There are four basic models:

(1) First come-first-serve

This quite self-explanatory model (which is also known as **FIFO = first in - first out**) is easy to implement, seems fair and has the lowest variance of waiting time.

(2) Sequencing based on customer importance

Examples for this model are the prioritized treatment of emergency cases in hospitals as well as serving long-time or more profitable customers first.

(3) Shortest processing time rule

This model minimizes the average waiting time by putting the shortest job first and the longest job last. This does not impact the total working time but ensures, that those customers with the shortest working times get served more quickly. The problem with this model is, that it gives customers an incentive to claim false processing times ('I need only five minutes.') in order to get served fist. For this reason, this rule is usually only used in manufacturing or in digital processing, where processing times can be accurately predicted (and no human customers are directly involved).

(4) Appointment rules

Some business are trying to solve the problem of customers not showing up in regular intervals by forcing them to show up in regular intervals through appointments. However, this model is quite impractical for many business (think about fast food restaurants or supermarkets) and downright impossible for others (think about emergency room treatment where individual demand is nearly impossible to predict).

4.8 Calculating the flow time efficiency

Instead of simply trying to optimize processes in order to cut down on waiting time, it is also important to look at the **customer experience as a whole**. How does, for example, a patient experience an appointment at a physician? The customer experience in this specific case does not start in the waiting room at the clinic, but also includes driving, parking, checking-in, filling out forms, driving back home etc. We thus have - so far - looked only at a small part of this entire customer experience. Many activities along this line are not adding any value for the customer - for example, getting to the doctors office is important, but does not cure any sickness.

The question of exactly how much time spent in a process (from the customers perspective) adds any actual value for the customer can be answered by calculating the **flow time efficiency**:

flow time efficiency = total value add time of a unit / total time a unit is in the process

This form of analysis is the basic idea behind the so-called **value stream mapping**. Here, every process is split-up between onstage actions (which are visible to the customer), backstage actions (which are invisible to the customer) and customer actions (which are those actions that the customer performs by himself) as well as additional support processes. Possible ideas for optimising the waiting time are moving work off stage (e.g. doing a rental car check-in via the internet), reducing customer actions (e.g. not retake basic medical information at each hospital visit), removing steps that offer no value (if possible) or avoiding too much fragmentation due to specialization.

4.9 Waiting behaviour of customers

The previous models all assumed, that customers would wait as long as it takes in order to get processed. In more realistic models, customers will leave the process before getting served because the waiting time gets too long. Some customers will not even enter the system, if the the demand is visibly too high (long waiting lines). In these cases, the **outflow** (completely served customers) will differ from the **inflow** (customer demand). But what fraction of demand will a business be able to serve?

There are **four basic models of possible customer behaviour** (which can additionally be mixed):

(1) All customers wait in line forever
(2) Some customers leave the line after a while
(3) Some customers do not enter if the line is too long
(4) Waiting time is absolutely impossible (inventory = 0)

Once one knows the probability, with which an incoming customer is not served, one can calculate how much business a company is missing because of waiting times. Instead of working with the rather complex formula for calculating this probability, the Erlang Loss Table can be used. This table denotes all probabilities for combinations of m (number of resources) and r (for the ratio between the processing time p and the arrival time a).

5 Week five

5.1 The two dimensions of quality

There are two basic dimensions of quality: **Performance quality** measures to what extent a product or service meets the expectations of the customer. **Conformance quality** measures if processes are carried out the way they were intended or promised to be carried out.

The root cause for quality problems is process variability. Were it not for process variability, every run through a process would result in the optimal output or in the very same error, which would then be easy to detect. However, due to process variability, some runs through a process result in optimal outcomes while others result in different kinds of errors. With some very basic **statistical probability tools**, we can assess the chances of such errors and defects occurring during a process. To calculate total error probabilities for an assembly line, one has to look at the error rate of each work step and calculate their yields (the percentage of error-free flow units the work step produces).

The **yield** of the process is defined as the percentage of error-free parts which are produced by the process - which of course depends on the yields of the various work steps. The total process yield is thereby simply the product of the individual yields:

*process yield = yield 1 * ... * yield n*

It is noteworthy, that even small defect probabilities can accumulate to a significant error rate, if there are many steps in a process. For example, if a process workflow consists of 10 steps with every step having a low defect probability of only 1%, the chances of an completely error-free product leaving this workflow are only $0{,}99^{10} = 89{,}5\%$.

The **Swiss Cheese model** explains, why defects or procedural errors sometimes do not get noticed, even if there are highly efficient quality checks in place: Since every slice of Swiss Cheese has some holes (defects) in it, there is a small probability that holes will line up in a way that creates a hole through a staple of cheese slices. This is then akin to multiple quality checks failing during the production of the same flow unit - though the chances of this happening might be low, it is bound to happen from time to time. This insight is also the main reason behind redundant checks, which means checking a quality attribute more than once to etch out all errors that might occur. With redundancy, a process hat to fail at multiple stations in order for the process yield to be affected.

5.2 Scrapping or reworking?

Should a damaged unit be dropped from the process or should it be reworked? In order to answer that question it has to be noted, that **reworking defects** can turn a process step into a bottleneck, which has not been the bottleneck before. Reworking defects (and thus, defects themselves) can have a significant impact on the process flow and on the location of the bottleneck. The bottleneck can therefore not longer be determined by just looking at the capacity of the process steps. Instead, one has to take into account the capacity changes in relation to the scrap and reworking rates.

To figure out where the new bottleneck is, we have to assume that the process as a whole will be executed in a way in which the demand is met, so that there is a match between the process output and the demand at the end of the process. The process therefore needs to start with more flow units then actually needed, so that enough flow units will be left over to satisfy demand. By working the process diagram backwards and determining the new demand for each process step, we can then discover where the new bottleneck will be located.

Instead of completely scrapping a flow unit, flow units can also be reworked, meaning that they can be re-introduced to the process and given a work-over to get rid of defects. This must also be taken into account when trying to figure out whether the location of the bottleneck changes, because some of the process steps will now have to process the same flow unit twice in rework, which will have an impact on their implied utilization. The location of the bottleneck can be determined by finding the process step with the highest implied utilization.

If the demand is unknown, the bottleneck can be located through four simple steps:

(1) Assume that the flow rate is an unknown demand D (e.g. 100 flow units).
(2) Figure out the demand D_x for each process step if D is to be reached.
(3) Divide D_x by the capacity of the process step to get the implied utilization.
(4) Identify the process step with the highest implied utilization. This step is the bottleneck.

5.3 Calculating defect costs

When trying to calculate the **costs associated with the production of defective parts** and / or faulty service results, it is pivotal to determine where in the process the problem has arisen. If an error occurs during the very first process step, little more than simple material costs are lost. If it occurs on the very last process step, we have to forfeit the value of an entire flow unit (including profit). The location of the bottleneck is especially important here. This is because defective flow units that are produced before the bottleneck have to be calculated with input prices while defective flow units that are produced after the bottleneck have to be calculated with the **opportunity costs of lost sales**. This insight drives the location of testing points in the process, who have to be arranged in a way which maximizes the chances of identifying and catching defective flow units before bigger losses occur.

By implementing a buffer between steps who can produce defective parts, the process flow can be protected against errors. However, producing too much inventory through so-called **error buffering** might conceal the necessity for improvements of the error rate.

5.4 The Kanban cards concept

A **Kanban system** is a visual way to implement a **pull system**. The basis of the system are the so-called Kanban cards, which basically are work authorization forms. Only if such a Kanban card is issued (which happens due to demand), work on the next set of units is begun. Since the inventory kept can never grow bigger than the sum of Kanban cards in circulation, the system allows users to keep a definitive cap on the inventory.

By setting the right number of Kanban cards, one can adjust the inventory to the needs of the current demand structure. This corresponds with the idea of a pull system: Rather then having everybody in the process work as hard as possible and pushing the flow units forward, the demand pulls the flow units through the process via the Kanban cards. Action is only taken if demand is there, not because there is idle time to spend or flow units are arriving from some other station in the process. Kanban cards thus support a just-in-time production.

5.5 Measuring quality with the capability score

To calculate the **capability score** (e.g. of a machine used in production or of a supplier), only three variables are needed: The **lower** and the **upper specification level** as well as the **standard deviation**:

LSL = Lower Specification Level
= flow unit needs to have at least XXX (measurable size)

USL = Upper Specification Level
= flow unit needs to have no more than XXX (measurable size)

*capability score = width of the specification level (USL - LSL) / 6 * standard deviation*

Because of the sixfold multiplication of the standard deviation (sigma), this is also known as the **six sigma method**. The capability score of a distribution can be translated into the probability of defect through the normal distribution function of standard spreadsheet software such as Excel.

5.6 Statistical process control

There are two **types of variation** in production and service processes - **common cause variation** and **assignable cause variation**. The common cause variation is the 'usual' variation that is caused by statistical anomalies in the production or service processes. The assignable cause variation is variation caused by a specific change in the underlying process structure.

There is a simple way to distinguish between common cause variation and assignable cause variation: Common cause variation falls within the previously mentioned control limits (upper control limit, lower control limit). Assignable cause variation falls outside those control limits. The control limits can be set as three standard deviations in each direction from the mean. A field with a width of six standard deviations will contain 99,7% of all cases. If a sample lies notably outside of these limits, assignable cause variation is likely to be the reason.

The purpose of **so-called statistical** process control is to constantly monitor the process output in order to be alerted to the occurrence of assignable cause variation almost immediately. In the famous **Toyota production system**, this is realized through a detect - stop - alert framework which catches defects quite quickly. This is critical, because defects tend to (a) produce more defects over time and (b) cause higher monetary losses once the defective flow units get through to the process bottleneck. Both problems provide huge incentives for figuring out how to detect defects as soon as possible. Some techniques that can be effectively used here are the drawing of **fishbone diagrams** as well as **laddering**. The idea behind both techniques is to basically ask 'why-questions' over and over again until the actual root cause(s) of defects is (are) identified.