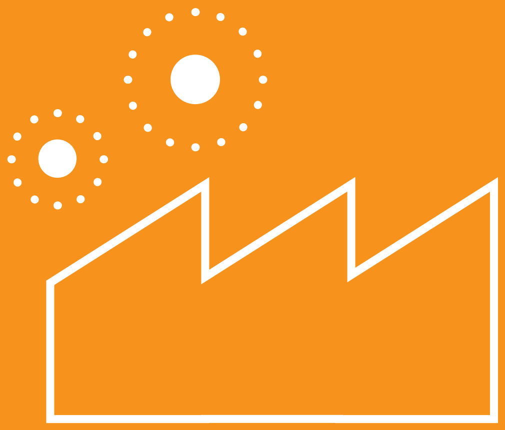
Universitat de Girona

/oberta

Notes on Operations Management

Martí Casadesús Fa
Josep Llach Pagès
Organisation, Business Management
and Product Design Department

Frederic Marimon Viadiu
Faculty of Economics and Social Sciences
(Universitat Internacional de Catalunya)



Notes on Operations Management

Martí Casadesús Fa
Josep Llach Pagès
Organisation, Business Management
and Product Design Department

Frederic Marimon Viadiu
Faculty of Economic and Social Sciences
(Universitat Internacional de Catalunya)

Universitat
de Girona

CIP 658.56 CAS

Casadesús Fa, Martí, autor

Notes on operations management / Martí Casadesús Fa,
Josep Llach Pagès, Frederic Marimon Viadiu. – Girona :
University of Girona - Publications Service, April 2023. –
94 pàgines : il·lustracions, gràfics ; cm. – (Oberta)
ISBN 978-84-8458-643-2

I. Llach Pagès, Josep, autor II. Marimon Viadiu, Frederic, autor
1. Fabricació 2. Automatització 3. Control automàtic

CIP 658.56 CAS

Edited by: University of Girona - Publications Service

ISBN: 978 84 8458 643 2

Girona, April 2023

Index

Presentation

The publication "Notes on Operations Management" is a document that covers different key topics in the management of production systems, organised in 11 chapters. The documents starts with the strategic aspects, and gradually and sequentially develops the more operational issues, covering aspects such as the calculation of the sales forecast, the master production shedule (MPS), the material requirement planning (MRP), stock management, project management, quality management, queue management, just-in-time, etc.

Nonetheless, this is not an extensive and detailed text on the subject like most manuals in the field, but quite the opposite. It is a very concise and synthetic text, whose sole objective is to guide readers on each topic. This is why the document has been specifically called "notes".

It is therefore an initial or complementary text for students who study subjects linked to "Operations Management" and who will undoubtedly need to supplement their knowledge with other resources such as the classes on the subjects, texts with solved problems, specific manuals for each topic, etc.

Chapter 1: Production planning and production process strategies

Objective: Understand the objective of production management and the main planning and production process strategies.

Learning outcomes:

Understand the objectives and challenges of production management.

Production planning strategies.

Production process strategies.

1. What is production management?

The term *production management* or *operations management* refers to the use of methods and techniques to convert raw materials into finished products. This process is based on a set of activities involving raw materials, components, finished products, human resources, working centres, machines, etc., the aim of which is to deliver a final product in the desired quantity and of the sought quality.

Effective production management must encompass ideal planning, suitable organisation and final control of all the stages involved in the preparation and delivery of the final product, always with the aim of obtaining maximum productivity. With these objectives in mind, the main functions of production management are:

- Definition of production processes:
  Choosing the most suitable production process, deciding the type of technology, machines, material handling systems, etc. In this process, it is also necessary to determine the production capacity that will satisfy demand for the product, so that the maximum possible productivity is obtained.

- Production planning:
  Plan production capacity in the short, medium and long terms. To this end, the expected demand for each product or family of products must be predicted, which is the starting point for planning the production to be carried out.
  The objective is to find the best and most economical option to follow in the manufacturing process, determining when each of the productive operations will begin and end, and ensuring a smooth workflow that meets customer demand.

- Control of production processes:
  Process control is established first to ensure that operations are carried out according to plan, and second to continuously monitor and evaluate the production plan to see if changes can be made to it that will better meet the objectives of cost, quality, delivery, flexibility, etc.

- Control and management of product quality:
  Ensure that products meet the quality expected by customers, at the lowest possible production costs. To meet demand, quality management must constantly improve the quality of products.

- Stock management:
  The stock level of all products must be optimised and controlled so that costs are kept to a minimum in relation to the main problems that may arise, including keeping excessive stock in the warehouse, leading to unnecessary storage costs, or having an excessively reduced stock for existing demand, resulting in lost or delayed sales.

- Machine maintenance management:
  Ensure proper maintenance and the replacement of machines and equipment when necessary. Maintenance management plans must therefore be developed that avoid machine breakdowns and the consequent production stoppages.

Consequently, adequate implementation of a production management process can bring multiple benefits:

- Improvement of the comprehensive coordination of all production processes.
- Having reliable and up-to-date information on the production plant for decision making.
- Achievement of greater efficiency and productivity, with savings on costs and production times.
- Development of alternative plans to deal with any emergency or unforeseen event in production planning.

Although since its inception production management has been geared towards the management of physical goods, it is also applicable to services. In fact, this is increasingly the case given the direction the industry is taking towards the future. It is necessary to bear in mind, however, what differentiates the management of a physical asset from that of a service. These differences can be summarised as:

- A service is intangible.
- The provision of the service is heterogeneous. Two providers of the same service will provide different services, and even the same provider is not always able to deliver the same standard service.
- The service is produced and consumed at the same time. Two consequences stem from this fact: the service cannot be stored, and it must be done right the first time as there is no second chance.
- The service allows a high level of interaction with the customer during the process.
- A service is often knowledge-based.
- The way to assess quality is more complex in services.
- The quality of services is highly dependent on people (employee motivation).

2. Production planning strategies

In general, two main strategies can be established in production planning. A company's chosen strategy will depend on whether it starts producing once a certain order has been received, or whether the system produces constantly, irrespective of the orders received. Specifically, these strategies are the following:

- MTO (*Make-to-order*). Strategy of production against demand.
  This strategy corresponds to when the company only has a stock of raw materials and produces each item as it receives an order. It can be said to be the make-to-order strategy. To this effect, production of a unit is prompted by an order being placed. It is therefore produced precisely to meet this order. In this way, customisation is permitted, and the client's specific requirements are met. In this strategy, there is never a stock of the finished product. It has the disadvantage that the customer must wait from the time the order is placed until the product is delivered: this is the time required to manufacture the product or to complete the service. It is the strategy that tends to use used when a high degree of customisation of the items by the customer is allowed, even though it considerably increases the delivery time of the product.

- MTS (*Make-to-stock*). Production strategy against stock.
  This is the exact opposite of the previous strategy, since the production of the good is carried out before there is a specific order. In other words, the good is produced and the finished product is stored, which ensures that it can be delivered to the customer at the time of placing the order. To this end, the product must be fully standardised, since when a particular unit is manufactured, the specific customer that will receive it is not yet known. Orders will be served from the finished product inventory because the company will have all types of stock: raw materials, components, semi-finished and finished products. The customer does not intervene at all in the design of the product. The product is always the same, for all customers. Production is generally planned in large batches to minimise manufacturing costs. It requires a large investment in equipment to increase efficiency and reduce manufacturing costs.

Obviously, in practice there are other half-way strategies between the two extremes, wherein part of the process can be against stock and the other part against demand. In fact, two additional production strategies can be defined:

- ATO (*Assembly-to-order*): Here, there is a stock of semi-finished products, and assembly occurs when each customer's personalised order is received. Standardised components or subsystems are thereby used to make customised products. The first part of the process can be against stock to manufacture some standardised components that are used to make the final product; and the second part, where the product is customised, is made to order. This production strategy is also called "modular".

- ETO (*Engineering-to-order*): The ETO strategy is applied when each product requires a new design and it therefore makes no sense to have a stock of raw materials. Deciding on the design of the product is also part of the order. This would be the case, for example, with the manufacture of a yacht, where the customer starts the purchase with the design itself.

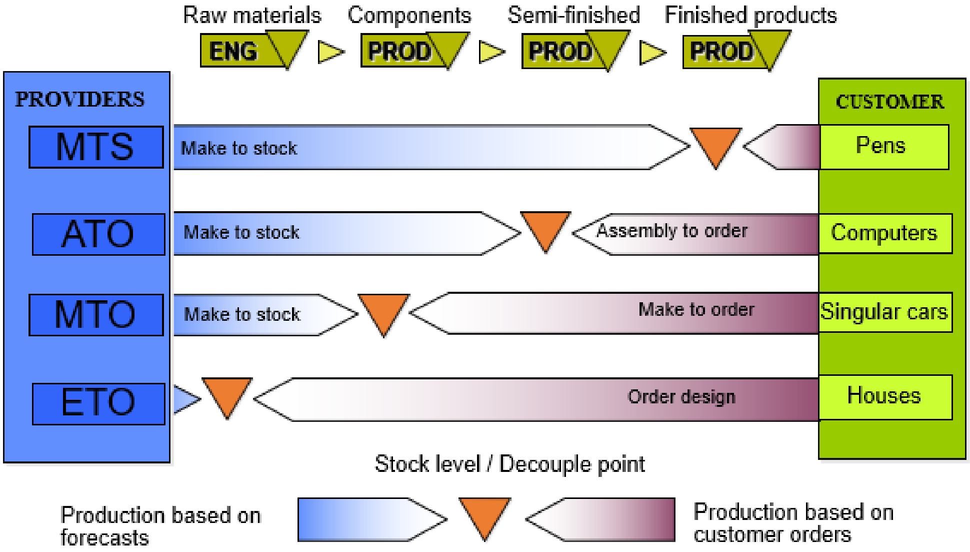


Figure. Production strategies.

To analyse the implications of each strategy, aspects such as those set out below must be considered:

- Customer participation in the design process: Adaptation to the customer's particular needs or requirements is higher in ETO and MTO strategies.
- Quality: The fit between the specifications required by customers and the delivered product is greater in the ETO and MTO strategies.
- Waiting time: In production against stock, the waiting time is zero since delivery is immediate.
- Stock: In MTS, there is an accumulation of products at the end of the process.
- Efficiency: The MTS strategy is more efficient in terms of unit costs, because it involves specialised equipment and tools that reduce production costs.
- Investment in equipment: This is higher in MTS since this strategy seeks to lower the unitary cost of production, meaning that the equipment is very specialised and unique to gain efficiency in the process and lower costs. In contrast, the MTO and ETO strategies require generalist tools and equipment that require less investment.
- Skill level of the worker: ETO and MTO generally require a more skilled workforce. The employee must deal with orders, each on of which is different, requiring them to have a wide range of skills.
- Product range: There is more range in MTO since the purchaser can customise each product, which is not possible when manufacturing according to MTS, where the customer each good is for is not yet known.

3. Process strategies

Regarding production processes, four different types can also be defined. Notably, it is the process itself that is analysed and not the strategy used to plan production. To this effect, there are different types of processes depending on the approach adopted.

- Process approach: Most of the world's production is dedicated to making low-volume, high-variety products in *job shops* or workshops. These facilities are organised around specific activities or processes. In a factory, these processes can be departments dedicated to welding, painting or any other activity. In an office, these processes can be accounts payable, sales, payroll, etc.

- Repetitive (modular) process approach: This type of process is the classic assembly line widely used in the manufacture of automobiles and household appliances. It is more structured than a process-focused facility, and it is therefore less flexible.

- Product Focus: High-volume, low-variety processes focus on the product. The facilities where this type of process takes place are organised around the product being manufactured. They are also called continuous processes, because they have very long, continuous production batches. Products such as glass, paper, sheet metal, light bulbs, beer and microchips are made by means of a continuous process. The equipment is highly specialised to increase efficiency and decrease production costs per unit, and so represents a major investment.

- Mass customisation: This is the rapid, low-cost production of goods and services that meet the requirements of increasingly unique customers. It is about producing exactly what the customer wants, when the customer wants it, and doing so economically. It can occur in the manufacture of computers and some household appliances where the components are manufactured in a repetitive process approach and the product is customised in the last assembly phase.

This table explains the characteristics of each of the typologies depending on the different aspects analysed.

Process focus (low volume, high variety)	**Repetitive (modular) approach**	**Product focus (high volume, low variety)**	**Mass customisation (high volume, high variety)**
Work workshops	Assembly line	Continuous process	Standardised sub-components, and final assembly is made to order
1. Small quantity and a wide variety of products	1. Long runs, a standardised product from modules	1. Large quantity and a small variety of products	1. Great quantity and a great variety of products
2. Highly qualified operators	2. Employees with moderate training	2. Less qualified employees	2. Flexible employees
3. Instructions for each job	3. Few changes in the work instructions	3. Standardised work instructions	3. Custom orders that require a lot of work instructions

4. High inventory	4. Low inventory	4. Low inventory	4. Low inventory in relation to the value of the product
5. Finished products are made to order and not stored	5. Finished products are made according to frequent forecasts	5. Finished products are made according to a forecast and stored	5. Finished products are manufactured on demand
6. Programming is complex	6. Programming is routine	6. Programming is routine	6. Sophisticated programming accommodates custom orders
7. Fixed costs are low and variable costs are high	7. Fixed costs depend on the flexibility of the installation	7. Fixed costs are high and variable costs are low	7. Fixed costs are usually high and variable costs are low

Table. Comparison of the characteristics of the four types of production processes.

Source: Jay Heizer, Barry Render, Chuck Munson (2020). *Operations Management: Sustainability and Supply Chain Management* (13th ed.) GE. Pearson Education.

Chapter 2: Analysis and design of production processes

Objective: Understand the main concepts and the associated metrics to analyse processes and make decisions to optimise them.

Learning outcomes:

Definition of production processes.

Variables of measurement and analysis of a process and laws that relate these variables.

Process programming to optimise performance.

1. What is a process?

A process transforms inputs or *outputs* into *outputs.* Inputs are items such as raw materials, labour, equipment, information and other miscellaneous resources needed to produce the *output.* The result can be either a physical product or a service. It is also important to analyse the level of satisfaction of the client or user of this *output.*

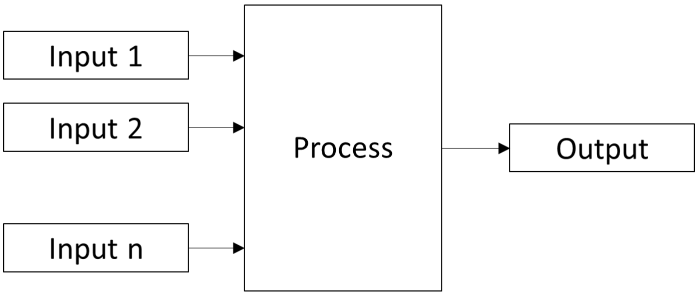


Figure. Flow of a process.

To this effect, a process is a set of activities carried out by an organisation that takes inputs *and* transforms them into a product or service. For the process to be useful, the result or *output* must have a value for the customer that is greater than the sum of the *inputs.* The added value must be such that someone (a customer) is willing to pay a higher price for the product than the total resources invested in the process.

Productivity: Ratio of *outputs* (goods and services) to *inputs* used (resources such as labour and capital). It is the coefficient between the output and the input, meaning that it is always a relational measure.

- Productivity = Output / Input

Productivity therefore measures the amount of *output* that can be obtained for each unit of *input* considered.

- Productivity of a factor: the relationship between the *output* and an analysed factor or resource. Using a single *input* to measure productivity is known as single factor productivity.

- Total productivity (or multifactorial productivity): the relationship between the *output* and the value of all the resources (factors or *inputs)* used. To be able to add all the *inputs,* they must be expressed in the same unit (often the monetary unit).

If the unit used to add the different *inputs* is the monetary unit and the monetary unit is also used to measure the *output,* the total productivity is expressed in a dimensionless way. It can be read as the amount of euros (or another monetary unit) obtained by the system from the market for each euro that enters the system through the considered *inputs.* This will therefore be the performance, indicating the amount of euros obtained for each euro that enters the system.

Here are some observations relating to measuring productivity:

- The quality may change, while the quantity of inputs and outputs remains constant.
- There are external elements that can affect productivity and are variations that cannot be attributed to the process management. The person (or team) in charge of the process has no control over them.
- In some cases, the appropriate units of measurement may not be found.
- Measuring productivity is particularly difficult in the service sector, where the final product can be difficult to define.

Three key factors have traditionally been considered to analyse processes: labour, capital and management. Depending on the type of process we are analysing, we can find other critical *inputs,* such as energy, technology, information, etc.

2. Process measurement

The first thing that must be known is the *flow unit*. This refers to the product or service that is manufactured or produced. It can be a car, a bottle of water, a hairdressing service, a smartphone... anything that is the *output* of the system or process. Therefore, in the manufacturing environment it will be the physical good produced, which is generally stored as a final product to go directly into the market. In the service environment, the *output* is the service provided, such as the doctor's visit, the consulting report received by a company, a financial service that a client receives from their bank, etc.

Second, the process needs to be understand, which can be helped by a flowchart composed of activities (which consume time and resources), stocks (raw materials, work in process and finished products), decisions, flows, etc.

Therefore, the main and basic elements of a flowchart are:

- Rectangles to represent activities (which consume time and resources).
- Inverted triangle to represent a stock. Three types must be considered: raw material, work in process (WIP) and finished products.

- Arrows for a flow (can be material or information).
- Rhombus-diamonds for a decision.

Different parameters must then be measured to analyse and learn about the process in question. Some of them are related in some way.

2.1. Processing time

Processing time is the time it takes an activity to process a unit of flow.

- Processing time = p (in the manufacturing environment)
- Processing time (or service time) = s (in the service environment)

It can also be considered for the entire process or system (system processing time). In this case, this concept is also called the system *lead time.*

2.2. *Lead time*

Lead time is the time it takes a unit of flow to go through the entire process; in other words, the time elapsed between when a unit enters the process (system) and when it leaves it.

2.3. Capacity

The capacity of a resource/activity is the maximum number of flow units that an activity (or resource) can process per unit of time. Therefore, when measuring a capacity there is always a unit of time in the denominator. The capacity of the resource (or activity) is the inverse of the processing time:

- Capacity of the resource (or activity) = 1/p

If an operation takes 15 minutes to perform (p = 15 minutes), its capacity is 4 units/hour.

If there are *m* resources, then:

- Resource capacity (or activity) = m/p

Process capacity is the maximum number of flow units able to be completed by the entire process per unit time. The activity with the lowest capacity is called the bottleneck. The bottleneck limits the total flow of the system.

- Process capacity = bottleneck capacity

This is why it is so important to rigorously analyse the bottleneck.

2.4. Production rate

The production rate of the system is also called *flow rate* or *throughput.* We also often call it simply *production.* This is the units of flow produced per unit of time. It is measured in the same way as the capacity, but the concept is different. The flow rate measures the actual production at a particular point in time, while capacity refers to a characteristic of the system.

When talking about the *rate* of something, we are referring to a measure where the numerator is a certain unit and there is always a unit of time in the denominator.

The actual production rate depends not only on the process capacity but also on short-term decisions based on such things as *input* availability or the actual demand rate.

Therefore, the maximum production rate is the capacity of the process.

2.5. Usage or performance

Process usage is the percentage of process capacity used at a given time. This percentage is also known as the process yield.

- Process utilisation = Current production rate / Process capacity

Analogously,

- Resource usage = Resource usage rate / Resource capacity
- Activity usage = Activity usage rate / Activity capacity

The usage of the activity can also be assessed as the percentage of its process time (p) related to the system cycle time.

- Activity usage = Process time of this activity / System cycle time

2.6. Cycle time

Cycle time is how often the process is “processing” or producing a unit of output or flow. It is measured as the time elapsed between two consecutive outputs of the process. It is verified that:

- Cycle time = 1 / Flow rate

2.7. Work in process (WIP)

Work in process (WIP) is the number of flow units in process at a particular time. It is not the stock of raw materials, because in WIP some operations have already been carried out on this raw material (value has already been added), although the product is not yet ready to go in the market.

2.8. Productivity

One of the most important parameters when analysing a process is productivity.

Productivity is the relationship between output and input (relational measure). If a unit of time features in the denominator, productivity and *flow rate* are the same. It is necessary to determine which units feature in the numerator (the main *output* to be analysed) and what *inputs* need to be considered.

2.9. Inventories or stocks

Three types of inventories are identified in any process:

- Raw material (from suppliers)
- WIP (some value has been added to the raw material, but the good is not yet ready to be delivered to the customer)
- Finished products (ready for shipment)

2.10. Bottle neck

As already mentioned, the bottleneck is the limiting factor in a system restricting system performance. It is the operation with the lowest capacity. Therefore, if the limiting factor is an activity, it is the operation with the longest cycle time.

To detect the bottleneck, it is necessary to express all the capacities of the activities or operations in the same units to be able to compare them. To illustrate the point, in the figure below the shortest capacity is operation B, with a capacity of 4 units per hour. In other words, B has the longest cycle time (15 minutes). Therefore, the bottleneck is B. The capacity of the entire process is 4 units/hour.

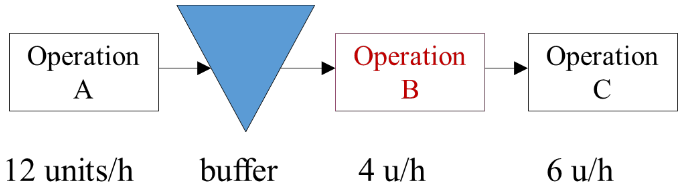


Figure. Example of a process in which a *buffer is inserted* to protect the bottleneck.

2.11. Buffer

A buffer is an inventory: the number of units stored. A buffer previous to the bottleneck protects the flow in the event of a flow interruption from upstream.

What size should this buffer be? The answer depends on the expected duration of the flow interruption. In the above case, if a worst-case failure of A (disruption of flow from A to the bottleneck) is expected to be one hour, a buffer of 4 units will allow B to work for the hour (fed from this buffer) and the system will not lose productivity. After one hour, when A returns to normal operation, the flow from A will be re-established.

3. Planning a linear process

If the purpose of a process is to achieve maximum throughput (that is, work at full capacity), the bottleneck must be operating at 100%. This is the most usual goal, but there may be situations where the optimisation of another parameter is sought.

A greater flow can often be obtained simply by redesigning the process (using the same resources). It must be kept in mind that the higher the flow rate, the shorter the cycle time, since they are opposites.

Some alternatives to reduce cycle times through process redesign are:

- Remove activities. By carrying out an effective analysis, activities that do not add value can sometimes be found and removed.
- Reduce the waiting time between activities.
- Eliminate rework. This means getting things right the first time.
- Carry out activities in parallel.
- Postpone some activities to optimise the use of some resources.
- Reduce set up times, making it easier to reduce the batch size.

When the flow rate is defined as the units of output per unit of time, then it is equivalent to productivity. Increasing productivity therefore means increasing the flow rate. In other words, increasing the flow rate (throughput*)* is the same as reducing the cycle time (Flow rate = 1 / Cycle time).

However, in most cases, the downside of increasing throughput is that in parallel it increases lead time. If the lead tin increases, this means that WIP also increases in the same proportion (this is a consequence of Little's law, developed in the next subsection).

Therefore, there is a trade-off or compromise between these parameters: improving the flow rate (productivity) is effectively the same as reducing the cycle time, and often involves an increase in the lead time*,* and therefore in WIP, making it necessary to analyse whether shortening the cycle time is worthwhile. The cost of improving cycle time is a higher WIP.

The simplest process is a linear one. This is the case where the activities are performed sequentially: when an operation has been performed, the flow unit enters the next station. Programming these type of process to achieve maximum productivity in terms of units of flow per unit of time requires these steps:

- Find the bottleneck.
- Plan the bottleneck to work at 100%.
- Plan the previous operations (at the bottleneck) to feed material to the bottleneck at the rate at which it can work.
- For downstream operations, those that come after the bottleneck, the material flow just needs to be left as it is.

Some ideas to manage the bottleneck:

- The system must be given work (production) orders at the rate set by the bottleneck capacity.
- Time lost in the bottleneck represents lost production for the entire system.
- Increasing the capacity of an activity that is not a bottleneck serves no purpose.
- Increasing the capacity of the bottleneck increases the capacity of the entire system.

4 . Little's Law

Little's law states that in the long term the flow (throughput*)* of a process is the coefficient between WIP and the *lead time.*

- Flow rate = WIP / lead time

This law is important when the flow rate or throughput is a constant parameter. This is often the case since it is a parameter determined by demand. In this regard, it can be said that it is external to the system. In this case, reducing lead time has a proportional impact on reducing WIP.

It must be kept in mind that reducing WIP is important, not just to reduce inventory costs but also to simplify the visualisation of the flow, which has been shown to ultimately impact on the quality of the product. It has therefore been concluded that reducing the lead time is important. Since this is made up of several activities, on analysing some of them can be found to be able to be shortened, or even eliminated, without losing added value.

The figure below shows the accumulated inputs and outputs in a system. In the long run, the two slopes should be parallel; in other words, the input rate should be the same as the output rate. If the input rate is greater than the output rate, WIP increases uncontrollably (and lead time to the same degree) and the system collapses.

The figure below illustrates Little's law from the point of view of descriptive geometry. The slope of accumulated inputs is the ratio between WIP (vertically) and lead time (horizontally).

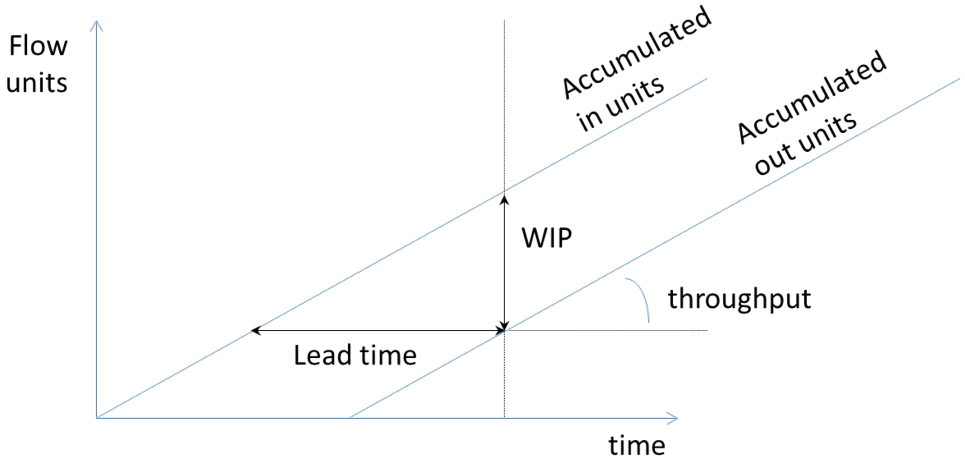


Figure. Useful input-output diagram for expressing Little's law.

Chapter 3: Demand forecast

Objective: Understand the need to estimate the expected demand to subsequently plan the production and understand the main aspects and techniques to consider when making this forecast.

Learning outcomes:

Use of demand forecasting

Types of calculation for demand forecasting

Methods for time series analysis:

Moving averages

Weighted moving averages

Exponential smoothing

Linear regressions with seasonality coefficients

1. Why is it necessary to forecast demand?

Demand forecasting consists in calculating as accurately as possible the sales volume of each product during a predetermined future period. This calculation, which can be carried out with varying degrees of accuracy, and with greater or lesser risk of adjusting to the near future, is essential to subsequently determine the amount of products to be produced in each period.

A correct demand forecast allows the company to improve subsequent production planning, with the consequent reduction in production costs. If the demand forecast is incorrect, the company risks making poor product and target market decisions.

Anticipating demand in fact means anticipating the level of activity from which most of the rest of the company's parameters are defined: fixed costs, variable costs, investment plans, etc. In addition, adequate forecasting allows you to:

- Anticipate demand, knowing when to increase or decrease staff and other resources to keep operations running smoothly.
- Optimise stock levels of products and components, increasing turnover rates and reducing storage costs.
- Provide insight into upcoming cash flow, enabling more accurate budgeting to pay suppliers and other operating costs.

The demand forecast depends on many factors that cause it to vary. In general, the following should be considered:

- Time of year: If you work with seasonal products, there are periods of the year when they are more in demand than others.
- Evolution of sales: Products sales can continuously increase or decrease.
- Company strategy: Opening new markets or introducing new products will affect sales, so it will be more difficult to forecast them based on historical data.
- Market evolution: The emergence of new competitors, economic developments or changes in regulation will also affect forecasts made of orders for a product.

Whatever the case, bearing in mind that to sell a product today it must have been produced in a previous period, it is essential to foresee the demand and, consequently, the production. The product is thereby always produced to cover an expected demand that has somehow been calculated.

2. Types of demand forecast calculation

Different methodologies can be used to carry out a demand forecast. They are generally:

- Time series: This type of calculation is based on the historical sales of a given product and can therefore be used when major changes in the market are not generally expected.
- Qualitative methods: Based on conducting surveys of consumers' or potential customers' opinions of different products or on expert predictions in a certain sector. It is used for the launch of new products in the market.
- Extrapolation of results: For the introduction of new products to the market, pilot tests can also be carried out (in a certain market or sector), or the sales obtained in a mature market taken as a basis from which to extrapolate the results to the market.

In either case, 4 fundamental concepts must be kept in mind in any type of demand forecasting, as shown in the figure below.

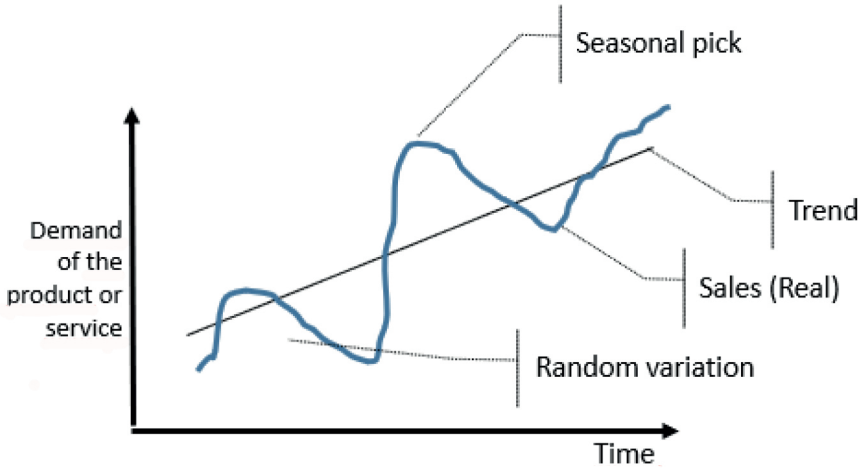


Figure. Concepts to consider in demand forecasting

Specifically, they are:

- Economic cycle: This is determined by the ups and downs of economic activities, marking a pattern that is repeated every certain number of years. It is very difficult to estimate using time series calculation.
- Trend: Pattern that represents an increase or decrease in demand over time. It can be explained by an economic cycle or by the life cycle of a product, among others. It generally lasts a few years.
- Seasonality: Pattern that always repeats itself, such as hours in a day, days in a week, months in a year, etc. It is dictated by changes in the climate, social customs, festivals, etc.
- Randomness: These are the unexplained variations in demand. They may be due to chance or unusual situations not included in the models used. They do not represent any kind of pattern, otherwise they could be used for forecasting. Their effects are not repetitive. In an "ideal" environment they would not exist.

3. Methodologies for calculating demand based on sales and historical data

Different methodologies can be used to carry out a demand forecast. Due to the high level of randomness of the demand to be forecast, it is often not necessary to use highly complex calculation methods given that the probability of making errors in the forecast is much greater than the accuracy of the mathematical model used. This is why very simple and rather intuitive methods are often used, giving us an approximate forecast very quickly.

Among the simplest and most intuitive methods, the following stand out:

3.1 Moving *averages* (MA)

This is a method based solely on the application of arithmetic averages between the last real sales to establish what the forecast for sales for the next period is. It is described like this:

$$F_{t+1} = MA_t = \frac{1}{n}\sum_{i=0}^{n} X_t \quad , \qquad t > n$$

Where:

$$F_{t+1} = Demand\ forecast\ for\ the\ period\ t+1$$

$$t = Subscript\ that\ identifies\ the\ periods$$

$$X_t = Demand\ on\ the\ period\ t$$

$$n\ = Number\ of\ periods\ considered\ in\ the\ calculation$$

It is called the moving average depending on the number of periods, *n*, that are used for the calculation. Consequently, MA(3) will indicate the moving average using the previous three periods for the calculation.

In its simplest version, it is recommended when the trend in sales is practically non-existent. In fact, in this method long periods (high *n*) are used if the demands are considered to be very stable, and short periods if they are not. The method obtains good robust forecasts vis-à-vis random effects not due to the trend.

3.2 Weighted moving averages or *weight moving average* (WMA)

An alternative to solving the trend problem is to use weighted series, or in other words to use different weights on the observations, giving the most recent data more relevance. The weightings can thereby take seasonality into account. This weighting must be determined based on experience, intuition or historical data, among others. It is described mathematically in this way:

$$F_{t+1} = WMA_t = \sum_{i=0}^{n} w_i X_t \quad , \qquad t > n$$

$$\sum_{i=0}^{n} w_i = 1$$

Where:

$$F_{t+1} = Demand\ forecast\ for\ the\ period\ t+1$$

$$t = Subscript\ that\ identifies\ the\ periods$$

$$X_t = Demand\ on\ the\ period\ t$$

$$n\ = Number\ of\ periods\ considered\ in\ the\ calculation$$

$$w_t = Weighting\ of\ the\ i - th\ period$$

It is called the weighted moving average depending on the number of periods, *n*, that are used for the calculation. To this effect, WMA(3) will indicate the moving average using the previous three periods, all of which have different weights, for the calculation.

The weighted method gives better results than without weighting when it is considered that the demand is not very stable and, therefore, for the forecast of the demand for the next periods, the last sales that the company has had will greatly affect the forecast.

3.2 Exponential smoothing

Unlike the previous methods, this one does not require a large volume of data on historical demand since it is the previously calculated forecasts that are used to calculate the next one. In fact, the formulation is very simple, as it only requires the previous forecast, the current demand in the forecast period and the smoothing constant.

Specifically, the formulation of the exponential smoothing is as follows:

$$F_t = F_{t-1} + \alpha\ (A_{t-1} - F_{t-1})$$

Where:

$$F_t = Demand\ forecast\ for\ the\ period\ t$$

$$F_{t-1} = Demand\ forecast\ for\ the\ previous\ period\ \ (t-1)$$

$$\alpha = smoothing\ constant$$

$$A_{t-1} = Real\ demand\ on\ the\ previous\ period\ \ (t-1)$$

The exponential smoothing method uses a smoothing constant alpha (α) with a value between 0 and 1, although in real application it is usually between 0.05 and 0.5. The constant functions as a weighting factor, like in the weighted moving average method, and its variation depends on the need to give more weight to recent data (higher α) or earlier data (lower α). Consequently, if α = 1, the demand forecast for the next period would be the same as that of the current period.

This model allows you to emphasise the most recent or the oldest demand. However, its disadvantage, like the moving average methods, is its response to the trend: it does not give good results for products that are experimenting a high demand increase.

3.3 Additive and multiplicative seasonality

A more complex method than the previous ones, but also simple enough, is the use of linear regressions and coefficients that enable the seasonality undetected by the regressions to be predicted. This method can be used when there is clearly a trend in sales, be it increasing or decreasing, and seasonality is also detected; or in other words specific periods of time when there are more sales than those determined according to the trend.

This method focuses on obtaining the parameters of the regression line that indicates the sales trend, and some seasonality coefficients that indicate by how many units (or what percentage of them) the product increases. When the seasonality coefficients indicate that product units increase or decrease sales with respect to the trend, this is the method of additive seasonality. When the coefficients refer to percentages, it is the method of multiplicative seasonality.

It is common to analyse the data by annual periods and have 12 seasonality coefficients, one for each month. To this effect, the seasonality coefficient for the month of January will tell us how many units have been sold, or it will give us a percentage, compared to the trend.

An iterative process must be carried out to calculate the trend line and seasonality coefficients.

Chapter 4: Aggregate plan and master production plan

Objective: Understand the different existing strategies and techniques for developing aggregate plans and master production plans, based on sales or their forecasts.

Learning outcomes:

Understand the objective of producing production plans, based on the required *inputs* and *outputs* of the process.

Aggregate production plan.

Sales and operations planning (S&OP) process.

Capacity and demand decisions to be made in the design of a production plan.

Methodologies and strategies for the realisation of production master plans.

1. What is the production planning process?

Managers define and deploy a strategy for each company. They make decisions about markets, the launch of new products, the incorporation of technologies, where to locate factories, what investments need to be made to vary capacity, etc. All these decisions affect the long term.

There are other decisions or policies that affect the medium term, which is usually a year but depending on the nature of the business this could be between 3 and 18 months. In this medium term, decisions are made that will affect the ability to produce and may also impact on the demand.

An example of a mid-term decision that affects capacity is the choice between producing overtime or subcontracting to meet a peak in demand. Another such example is the choice between hiring workers during high demand months and laying them off in the low season, or trying to schedule workers' holiday allowance to keep employment levels stable. These two examples have an impact on production capacity.

Decisions can also be made in the medium term to modify demand. A special price can be negotiated with customers to deliver an order with a certain delay. This action is called back ordering. It is a decision that affects the medium-term sales schedule and the delivery of orders. They are decisions of a very different nature from the two described in the previous paragraph.

For their part, capacity options are internal to the organisation. There is no need to negotiate them with anyone from outside the company. In contrast, demand options are more proactive and impact the demand profile. The first belong to the company's operations departments, and the second are of a commercial and logistical nature.

Therefore, to meet demand some decisions may have to be made that modify the rate of production and others that may affect the rate of sales. This is why when planning what needs to be produced in the medium term it is called sales and operations planning (*S* &OP). Both the pace of production and the pace of deliveries must be decided at this stage. The S&OP must respond to both aspects, and the result of the S&OP process is the aggregate production plan.

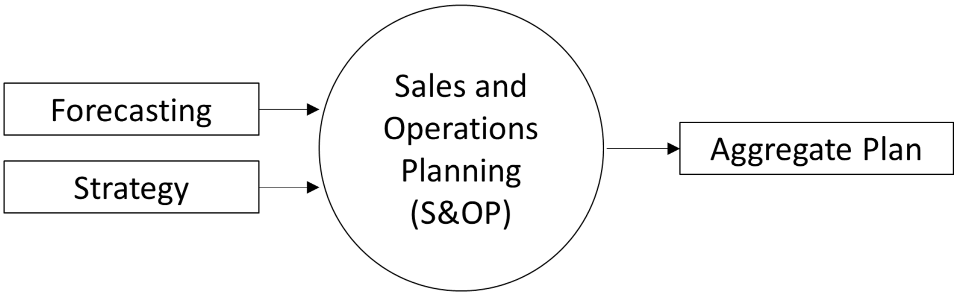


Figure. Sales and Operations Planning (S&OP) Process.

S&OP is, therefore, the medium-term planning process. As stated, the outcome of S&OP, the aggregate production plan, must satisfy the sales forecasts in the medium-term time horizon.

It must be kept in mind that production needs generally equate to the sales forecasts. However, there is a conceptual nuance between the two terms. *Production needs* are forecasts corrected by some strategic decisions, while *sales forecasts* are calculated based on the past and can set a trend. Nonetheless, a strategic bet can modify these forecasts and transform them into production needs, which is why some authors prefer to talk about production needs rather than forecasts.

This aggregate production plan is *the input* to the shorter-term scheduling process. This short-term programming needs more specificity. The process of breaking down the aggregate plan into greater detail is called disaggregation, since S&OP uses the aggregated unit and not the final product that goes to market. The aggregate unit facilitates an estimation of the amount of product to be manufactured, without specifying the details of the exact products. Once all the products forming part of the schedule are considered, this disaggregation, as shown in the figure, gives rise to a master production schedule (MPS). The MPS is applied over a short period (one month or six weeks), providing information for material requirement planning (MRP).

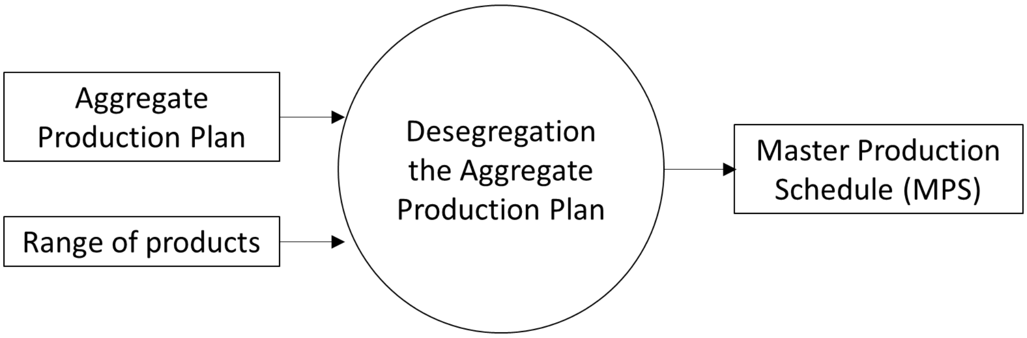

Aggregate Production Plan
Range of products
Desegregation the Aggregate Production Plan
Master Production Schedule (MPS)

Figure. Disaggregation process of the aggregate production plan.

2. Aggregate Production Plan

Medium-term production needs are therefore sales forecasts modified to consider strategic decisions that affect the long term. These production needs are one of the inputs to the S&OP process. The other input is decisions that affect the medium term. The output of the S&OP is the aggregate production plan, which determines the amount of the product and the time to produce it during this period (medium term).

S&OP uses product information in an aggregate unit, also called a logical unit. Each company must establish how the different product families are "translated" into this logical unit. To this effect, the aggregate plan is also expressed in the same aggregate or logical unit. For a car manufacturer, for example, the logical unit will be "cars", with no reference to the different models. In this scenario, a high-standard car with several functionalities can be equivalent to two "standard cars". All models need to be measured in relation to the "standard car". A milk producer will use "litres of milk" processed, with no reference to different products delivered to the market. Different families of products will be produced in the same factory, including regular milk, skimmed milk, different flavoured yogurts and milkshakes, and so on, and each family of products is presented in different packaging and sizes. However, at the aggregate level, only "litres of milk" will be considered. Everything must be measured in "litres of milk", which in this case is the aggregate planning unit.

To elaborate a feasible and optimal aggregate plan, the costs associated with the production process (inventory costs, labour costs, subcontracting costs, etc.) also need to be known. In this regard, it must be decided whether to produce the product by employees working overtime or by buying it already made from another producer or competitor, in which case the cost of the extra hours and the price of the product when bought from the competitor must be known.

Last, it is necessary to find a function that considers all the production and logistics costs that facilitates the search for the optimal planning in terms of costs.

In summary, to produce the sales and operations planning (S&OP), the following points need to be clear:

- A logical unit for measuring sales and production: the *aggregate unit.*
- A demand forecast in aggregate units.
- The main costs and other parameters that affect production and distribution (cost of inventory, cost of delaying a delivery, cost of hiring a worker, cost of subcontracting, etc.).
- An analytical model to estimate and minimise total costs.

With all this information known, the amount of aggregate units to be manufactured and when they need to be used in the medium term (the aggregate production plan) will need to be made. If the medium term is one year, it is reasonable for the plan to indicate the monthly quantity to be manufactured and delivered.

3. Aggregate planning strategies

When generating an aggregate production plan, some questions arise:

- Will the accumulated inventory be used to absorb changes in demand? And particularly, will it be necessary to manufacture in advance what will be delivered during periods of

high demand? Will the aggregate plan depend on the holding inventory cost that will need to be shouldered from the moment the unit is manufactured until its delivery?
- Is it better to pay overtime during a period of high demand or to outsource? In this case, each country's labour legislation must be considered.
- Can the number of workers be varied on a regular basis by hiring and firing workers? If so, is it feasible to provide training in a short time so that new workers are operational immediately? The implications of this system in terms of working climate must also be considered. There are sectors where there is no choice but to implement this practice.
- Can you consider changing the demand curve by offering discounts or special prices during periods of low demand? This decision has an impact on the behaviour of the market. It is a measure of a different nature than the previous ones, since it does not affect the ability to produce but it does affect the demand pattern.

3.1. Options for the aggregate planning process

Therefore, the necessary decisions (choices) can be classified into those affecting capacity and those affecting demand. The first do not aim to change demand but to absorb fluctuations by changing the rate of production. For their part, *demand options* try to smooth out fluctuations in demand, so their goal is therefore to stabilise the labour force.

Capacity options:

- Change in inventory levels. In this case, the costs associated with the inventory must be considered. What must be known is how much it costs to have a unit of product in stock for a month. Also important to remember is that stocking food products is different from stocking hardware products.
- Changes in the number of workers in the different months, meaning periodic hiring and dismissals and involving not only the direct cost of hiring and firing, but also aspects related to the training of new workers, the impact on the working climate, the impact on the knowledge that can be lost with workers leaving the company, and so on.
- Outsourcing. Again, it should be kept in mind that in addition to the direct cost of purchasing the product from a supplier, there are other effects to be analysed. Outsourcing involves some risks such as customers possibly ending up "migrating" and connecting directly with the alternative supplier. This practice also means that the company loses control over the quality of the product.
- Use the same workers, but extend their capacity by means of overtime. Note that these hours are more expensive than normal or regular hours. It is also necessary to see what obstacles are posed by the labour legislation of the country.

Demand options:

- Change demand by offering discounts, advertising or special campaigns, with complementary products, etc.
- Delay deliveries (*back-ordering*) that should be served during high demand months.

3.2. Strategies for the aggregate planning process

There are basically two strategies to find the optimal aggregate plan: levelling the production rate and chasing demand. Among these two pure strategies, there are many intermediate options (mixed strategies).

- Levelling production strategy

  This consists in producing the same amount of logical units of product in each period, irrespective of the demand, which is variable for each period of the year (generally by months). The total production required annually (or during the period considered) will be distributed evenly among the months, resulting in some level of inventory in months of low demand and having orders pending delivery in times of high demand. This is the constant workforce plan.

- Chase strategy

  This is the direct opposite of the previous strategy. The production will be the exact amount needed in each period. There will therefore never be an inventory or pending orders (no back-ordering). Conversely, capacity must be adjusted to the seasonality of demand. Great flexibility is required.

  This is the only valid strategy for service companies since one of the characteristics of a service is that its production is simultaneous with its consumption.

- Mixed strategy

  A mixed strategy is any intermediate situation. The two pure strategies described above provide a framework and a point of reference, but other more advantageous scenarios in terms of costs and other criteria can usually be found. From a practical point of view, pure strategies may not be desirable, but they are always a good starting point to find a better solution.

4. Aggregate production planning methods

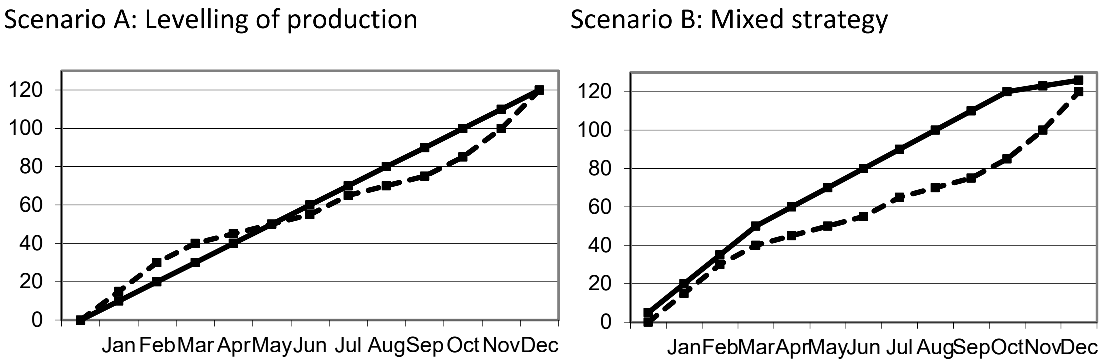
Different methods can be used. The goal is always to minimise the cost of the function associated with the aggregate production plan.

4.1. Iterative/graphical method

The iterative method is the most applied because it can be easily programmed into a spreadsheet. It also allows different scenarios to be generated. The two pure strategies (production levelling and demand monitoring) are generally modelled first, and then different mixed strategies are tested to minimise costs.

The figure below shows two strategies for a particular case. The dashed line shows the accumulated production needs, which are equal in both graphs. The continuous line represents two different planning strategies. In this case, scenario A is the result of a production levelling strategy. The total production requirement at the end of the year is 120 units, and will be produced at a constant rate of 10 units per month irrespective of the sales forecast, which varies each month. In the first months, production needs are above the planned production, so there will be orders that cannot be delivered on time. During the second part of the year, it is the other way around. Stock will accumulate because the rate of production will be higher than the need for it.

In scenario B, the expected demand pattern is the same, but the strategy to meet production needs (the proposed production and sales plan) does not allow for delays in serving orders. At any given time, there is an inventory. The continuous line, which is the plan, is always above the dashed line (production needs). At the end of the year, there is still a small stock left. The production slope is generally constant, with only two changes in slope: one in March, which is hardly perceptible in the graph, and another in October.


Scenario A: Levelling of production

Scenario B: Mixed strategy

Figure. Examples of aggregate production plans using graphic technique (Note: The dashed line represents the production needs, and the continuous line the aggregate plan).

The demand chase scenario would be represented by a graph with the two curves entirely superimposed, meaning we would only see one line.

At this point, the total costs associated with each scenario must be calculated. Additional scenarios can be represented, also with their associated total costs calculated. Last, you must decide on and stick to a plan, or in other words, execute it.

4.2. Heuristics

In some specific cases, and under certain hypotheses, there are heuristics or algorithms that can provide at least a good solution, if not an optimal one. It must be kept in mind that the possible solutions are practically infinite in many cases.

As an example, one of the heuristics that can be used is the Bowman method, which is a suitable method when the aim is to find a solution for a master production plan and the following assumptions are met:

- There is only one product or family of products to make.
- There are different ways to obtain the product, with a different variable cost of production.
- There is no fixed cost of production, and no change in cost across production levels.
- The product can be stored.
- Stock outages do not occur.
- Units produced in one period can be used to meet demand in the next period.

In this particular case, and under these restrictions, the Bowman method provides the optimal solution to the problem. To apply it, and in general, you must follow the following steps:

- Establish the demand for each period (usually months) that needs to be served.
- Establish production capacity for each period, considering all sources of production (for example, production in normal hours, production in overtime, outsourced production, etc.).
- Calculate production costs for each source in each period, bearing in mind the holding costs of units in stock when planning to produce them in periods prior to when demand is expected.
- To meet the units of demand in each period, the units of production will be allocated to the most economical source (usually producing in normal hours) for that same period.
- Once the production units for the same period are exhausted, the production will be allocated to the other alternative sources (overtime, during normal hours in previous periods, outsourced production, etc.) that guarantee the minimum cost.
- Once all the production necessary to cover the demand has been assigned, the total production in each period for each source and the expected cost of the entire plan must be defined.

This method can also be used in cases where the demand can be deferred. However, although this may be a good solution, it cannot be guaranteed to be the optimal one.

4. 3. Mathematical optimization methods

Alternatively, an analytical procedure, such as linear programming, can be used to provide the optimum solution. Specifically, the *simplex* method is used, which is able to optimise a function subject to some restrictions. The downside is that this optimum solution may not be feasible from a practical point of view, but whatever the case it is a good starting point from which to develop a feasible plan.

Excel software, one of the most widely used spreadsheets, offers a macro called Solver, which optimises a function located in a cell using the simplex method. The function must take into account all the costs that need to be considered (inventory costs, labour costs, outsourcing, etc.). This function needs some parameters specific to the case (holding cost of a unit in inventory for a month, hiring cost per employee, etc.). Other model constraints must be expressed as equations.

Specifically, it is first necessary to establish what the variables of the model are and what you want to have control over and decide about. The most important variables are the quantities to be produced each month: these variables are precisely the aggregate plan, which is what we are looking for. It will also be necessary to decide on the level of stock, and on hiring and firing of employees. Therefore, some of the variables will be:

- Production in month *i* : P_i, i = 1 to *n* (number of months)
- Stock in month *i* : S_i, i = 1 to *n* (number of months)
- Contracts in month *i* : C_i, i = 1 to *n* (number of months)
- Firings in month *i* : F_i, i = 1 to *n* (number of months)
- Workers in month *i* : T_i, i = 1 to *n* (number of months)
- ...

You must know parameters on which we cannot decide, such as the price of keeping the product in stock, and the cost of firing, subcontract, etc. Among these parameters are the production needs for each month (or in simpler terms, the monthly sales forecast).

- D_i: Aggregate sales forecast D_i, i = 1 to n (number of months)
- CInv: Inventory Cost of a unit for a month
- CCon: Cost of hiring a worker
- Prod: Quantity of units that a worker can produce per month
- ...

The function to be minimised will depend on these variables and parameters:

- Total cost = inventory cost + hiring costs +...
- Total cost = Σ CInv * S_i + Σ CCon * C_i + ...

There will also be some restrictions, such as:

- $S_i = S_{i-1} + P_i – D_i$ for i = 1 to *n* (number of months)
- $T_i = T_{i-1} + C_i – F_i$ for i = 1 to *n* (number of months)
- P_i = Prod * T_i for i = 1 to *n* (number of months)
- ...

Last, using the simplex method, the minimum " total cost", subject to the necessary restrictions, must be found. This will give some values of the variables, among them the values of P_i, which is precisely what is being sought.

Chapter 5. Material requirements planning

Objective: Understand that it is essential to plan material needs to meet production demand, ensuring the availability of capacity.

Learning outcomes:

Dependent and independent demand.

Material requirements planning (MRP I).

Capacity requirements planning (CRP).

Manufacturing resources planning (MRPII).

Enterprise resource planning (ERP).

1. Introduction: Why is it important for companies to plan their material requirements effectively?

Material requirements are understood as all the resources companies need to implement their production system. These resources can be human, economic and material.

Companies' production systems are becoming increasingly complex to be able to respond to customer demand. A company’s success in the market depends to some extent on their ability to plan materials, production and stock management. Therefore, companies must not only equip themselves with resources, but they must also know how to combine them in an optimal way.

It is precisely from the need to manage all resources that MRP (material requirements planning) systems arise, representing a significant advance for the company management because MRP enables the company's entire production system to be integrated into an information system.

The use of MRP systems dates back to the Second World War, when the United States government used programmes to organise the logistics of its military resources. Later, these systems were transferred to industry, and in the 1960s and 1970s the first MRP systems as we know them were created.

In short, an MRP system is an information system designed to plan manufacturing production. It identifies the materials needed, estimates their quantities, determines when they will be needed to meet the production schedule and manages the lead time, with the goal of meeting demand and improving overall productivity.

Material requirements planning can be relatively straightforward when production volumes are small and products are simple. Conversely, it can be very complicated when production volumes are larger and products are more complex.

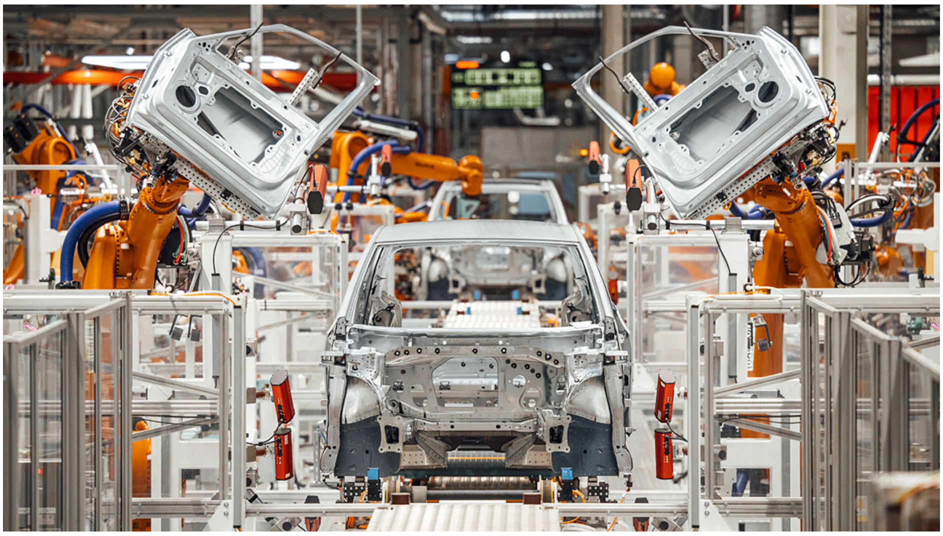


Figure. Volkswagen production line.

Source. https://www.volkswagenag.com/en/news/stories/2020/09/pioneers-in-climate-friendly-vehicle-manufacturing.html#

Without planning for material requirements, it is impossible to manage inventory effectively to have the right amount of the right items at the right time. Having too much inventory is costly, but not having enough can lead to stockouts, which are often the root cause of production interruptions, delayed shipments, additional costs and poor customer service.

2. Dependent and independent demand

As introduced above, MRP systems are materials planning and management systems that schedule production and control inventories. In short, they respond to the question of how much and when the necessary materials should be procured (produced or purchased) to meet the demand.

An important aspects for the correct functioning of MRP is the differentiation between independent and dependent demand.

- Independent demand:
  Independent demand is the demand for products that does not depend on the company, but corresponds to the demand for finished products by customers. This demand is usually external to the company because customers' decisions are not controllable by the company. Demand for spare parts is also considered independent demand, even though it is not a final product.

- Dependent Demand:
  Dependent demand is the demand generated by the decisions made by the company. Based on independent customer demand, the company calculates which components are needed to manufacture the final product, considering its structure.

For example, if a car manufacturer forecasts a demand for a certain number of units in the coming weeks, the company's planning department must consider how many steering wheels, engines, wheels, etc. must be available to meet the demand.

There are three key questions that the company must answer when planning for dependent demand:

- What components are needed to manufacture the final product?
- How many units of each component are needed?
- When are the components needed?

3. Calculation of material requirements planning (MRP)

The MRP calculation requires basic information from the production system, and in return generates two types of outputs: production orders and procurement or purchase orders (PO) within a given time horizon.

The basic information for the MRP calculation is as follows:

- Bill of materials (BOM):
  Hierarchical representation of all the components that make up the final product. It is organised by levels, the highest level showing the finished product and the lowest level the raw materials. It also includes the quantity relationship between two consecutive components.

- Master Production Schedule (MPS):
  Quantities of each final product to be produced in a given time interval. The time horizon referred to by the MPS is generally short. To define the quantities to be produced, both the external demand forecasts and the firm orders already received for a specific product must be considered.

- Inventory record:
  Amount of available stock of both the final product and the components that make it up. For each product and component it is also necessary to know the manufacturing or purchase delivery time.

Using the basic information, the OFs of both the final product and the components, and the PO of the raw materials, will be calculated for each component of the BOM, starting from the upper level and going down to the lower level.

To calculate the FO and the PO it is necessary to know the following previous concepts:

- Gross requirements (GR): Amount of product that must be available for external supply (needs of external demand) and for use in the company’s other production processes or in other manufacturing phases (needs of internal demand).

- Scheduled receptions (SR): Receipts of material corresponding to orders placed in the past and that must arrive within the planning horizon.

- Stock: Quantity of product remaining in stock at the end of the corresponding period.

- Net requirements (NR): Need for an article that cannot be supplied with the planned stock and which will therefore force the generation of a production or purchase order.

- Incoming orders: Quantities of product that will be received in this period from OC or OF that have been issued in previous periods. The batch policy established by the company will be considered.

- Outgoing orders: Quantities of product to request at any given time, considering the delivery time and the period in which it must be received.

The process to carry out the calculation of the FO and PO is iterative for all the components of the product and follows the following stages, always starting with the final product:

1. In the case of the final product, gross needs are calculated from external demand, and in the case of components and raw materials from the previous component, the amount of relationship specified in the BOM is considered.
2. Scheduled receptions on the horizon are included.
3. The net requirements are calculated to be able to meet the gross needs, the stock of the previous period and the scheduled receptions of the same period.
4. Incoming orders are calculated according to the batch policy of the company for manufacturing products, and depending on the supplier for purchase products.
5. The time when the FOs and POs should be launched is calculated, considering the delivery time (outgoing orders).
6. The process goes back to the beginning for the next component, following the structure of the BOM.

4. Capacity planning requirements (CRP)

The MRP calculation does not verify whether the company has sufficient capacity to produce the internally generated production orders to meet external demand. The calculation of capacity requirements planning (CRP) arises from this need.

In short, CRP is a process that consists of calculating the capacities required at the different work centres to satisfy the OFs, and comparing them with the available capacity at each centre over the periods of the horizon.

The CRP calculation process is carried out in four steps:

1. The charge (time) generated by all FOs is determined.
2. The charges calculated in the previous step will be proportionally distributed, where appropriate, throughout the periods of the horizon.
3. The required capacity per period in each work centre is determined.

4. The available capacity of each work centre is compared with the necessary capacity, and its deviations are determined.

By way of example, the following figure shows how in periods 3 and 4 the capacity required by the FOs calculated in the MRP is higher than the available capacity.

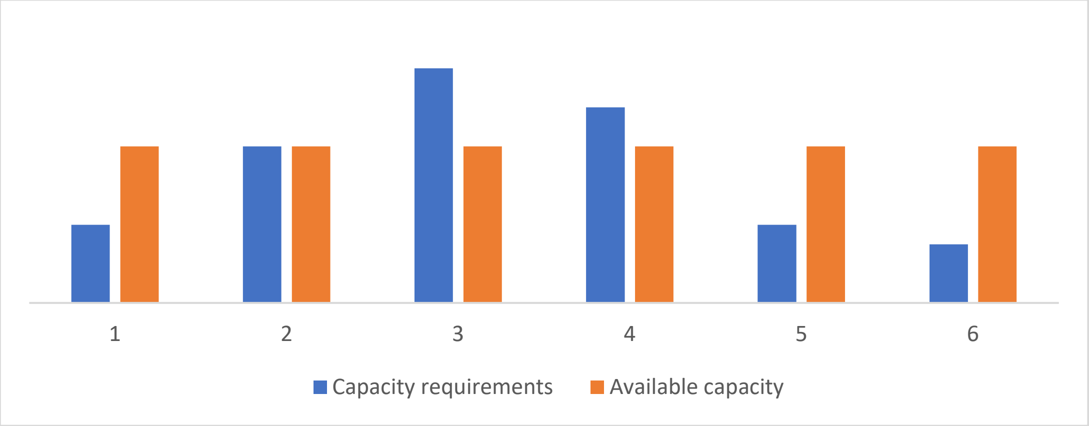


Figure. Capacity requirements *vs.* available capacity. Source: own elaboration.

Different actions can be taken to address this situation of lack of capacity. In any case, the contextual situation of the company and the availability of other resources, such as economic resources, will dictate the best course of action:

- Temporary readjustment of production orders.
- Readjustment of production batch sizes.
- Increasing employee flexibility.
- etc.

5. MRP II

When referring to MRP systems, a distinction must be made between MRP I (material requirements planning) and MRP II (manufacturing resource planning). In fact, MRP II systems are an improved version of MRP I systems. While the acronym of the latter stands for material requirements planning, MRP II stands for manufacturing resource planning. However, the terms are often used interchangeably.

As discussed above, MRP I systems appeared in the 1960s to address the need to effectively manage inventories of demand-dependent products. However, in some respects they quickly became obsolete. They therefore had to evolve until a more complete planning model was found that considered the entire business organisation and included capacity calculation, which was carried out a posteriori by means of the CRP. MRP II systems emerged from this need in the 1980s.

In contrast to MRP I, MRP II also includes financial and sales information to consider the availability of resources. In addition to answering the what, how many and when to buy materials, MRP II answers the following questions:

- What resources are necessary?
- How many resources are necessary?

- When will these resources be needed?

In short, MRP II systems identify capacity problems in the master production plan, resolve discrepancies between resource availability, and calculate forecast consumption.

The following table details the main differences between the MRP I and MRP II systems:

Characteristics	MRP II	MRP II
Stock control	x	x
Bill of Materials (BOM)	x	x
Master Production Plan (MPS)	x	x
Equipment maintenance schedule		x
Accounting and financial planning		x
Control of the capacity of work centres		x

Table. Main differences between MRP I and MRP II.

6. ERP (Enterprise resource planning)

In the 1990s, with the evolution of computers and the development of new technologies, MRP systems evolved towards an integration not only of the production aspects of the company but of all areas of business management, from financial management to sales management. ERP systems thereby came into existence.

The aim of ERP systems is to integrate and automate the company's main processes and data in a single technological platform, facilitating the management of data and the organisation of the company in real time.

To this effect, when a firm order arrives at the company and the seller registers it in the system, all the departments involved are informed in real time. They include the production system, which has to produce the demand; the warehouse, which has to supply the material to produce the order; the financial department, which has to activate the process to collect the order; and the logistics department, which will be informed when and how the order must be delivered.

The fact of integrating different tools in the same programme facilitates the cross-referencing of data and the generation of reports on the different areas. This information makes the company more flexible and competitive organisationally. With an integrated management system such as an ERP, the company can use the information it generates to optimise processes, reduce costs and facilitate decision making.

In recent years, due to the rapid advance of technology, ERP systems are no longer implemented on company premises but are managed from the cloud in a decentralised manner.

Chapter 6: Short-term scheduling

Objective: Understand the main challenges of short-term scheduling (allocation and sequencing of manufacturing orders) and techniques for optimising it.

Learning outcomes:

Definition and objectives of short-term scheduling.

Priority rules.

Assignment of production orders.

Sequencing of production orders.

Techniques for sequencing production orders: Johnson and Jackson algorithms.

1. What is short-term programming?

Short-term programming can be considered the last stage in production planning. It essentially consists of matching established tasks and operations, often by previous MRP execution, to specific people and machines or workplaces.

The degree of influence of short-term scheduling on the company's results is usually significant, since the fulfilment of delivery deadlines may be at stake, which is a highly valued aspect for customers.

The horizon with which they work is very short: days, hours and minutes, which is why trained professionals, often production managers, are needed to carry out this task using different methodologies or software to support the scheduling process. These methodologies or techniques can be classified, in a very general way, as follows:

- Forward scheduling: Forward scheduling starts as soon as production orders are received. Production starts as soon as orders are received, irrespective of delivery dates. It is used in make-to-order (MTO) processes where delivery is usually as soon as possible.
- Backward scheduling: This scheduling starts from the delivery date of the order. Its main objective is to meet the established delivery deadlines. Therefore, the last operations are scheduled first, and those that precede them are scheduled successively in reverse order.

To carry out this scheduling, there will generally be two distinct phases:

- Assignment of production orders.
- Sequencing of production orders.

However, before embarking on this scheduling stage, it is essential to know your objectives. Different objectives will require different methodologies and, as a result, different programming of operations.

2. Objectives for the short-term scheduling

The choice of the appropriate scheduling technique depends on many factors such as the characteristics of the process, the flexibility of the workplaces, the volume of production orders and the consideration of one or more objectives aligned with the company's strategy.

These criteria normally use one of the following indicators:

- Fmax: Dwell time (or length stay) in the system of the part that is the most
- Fmin: Dwell time in the system of the part that is less than the most
- Fmig: Average dwell time of the parts in the system.
- Tmax: Delay of the part that is most delayed
- Tmean: Average delay of all the parts produced

The most commonly used criterion in a production system is the minimisation of Fmax, i.e., scheduling production in such a way as to minimise the time on the production line of the part or piece that takes the longest time to produce. However, this is not the only one, as the chosen criterion often varies depending on each specific situation. Some criteria that could be used are:

- Finishing the last piece as soon as possible, i.e. to minimise Fmax.
- Reducing the average dwell time of all pieces in the system.
- Reducing as far as possible the unproductive time of the work centres between two consecutive operations.
- Reducing as far as possible the number of pieces in process (work-in-progress or WIP).
- Obtaining a regular "output stock" in batch production (balancing of production lines).
- Maximising the use of work centres.
- Minimising the number of orders that do not meet the dates committed to with customers.
- ...

It is clearly impossible to obtain a production schedule that optimises all the criteria that can be established, and improving one criterion is often linked to the worsening of others. For example, maximising the utilisation of work centres will involve working with larger batches, and will therefore probably increase the average delay of the parts produced in the case of a balanced demand. The objective to be met must thereby be unique, and if different objectives are to be met, one of them must be the priority.

Notably, these criteria are not scheduling techniques but simply indicators that will measure the effectiveness of the scheduling and sequencing rules applied.

3 . PHASE 1: Allocation of production orders

The allocation or loading - also called scheduling - of production orders is the first phase of short-term production planning. It is usually known after the launch of the MRP:

- Production and purchase orders with delivery deadlines.
- The expected workload for the planned production orders.
- One or several manufacturing routes for these orders.
- Work centres with different restrictions (schedules, shifts, capacities, etc.).

- Operation times for each garment in each work centre.
- Set-up times for the work centres, which may or may not depend on the garments that have been produced previously.

From here, these production orders must be assigned to the company's different work centres to optimise the objectives the company has set, be they to minimise operating costs, reduce the completion time of the last production order, reduce unproductive time, etc.

This allocation is very simple if all production orders can only be dealt with in a certain work centre, but this is generally not the case. Production orders can often be processed in more than one work centre, or all production orders can be processed in any of the company's work centres. If we also consider that part of an order (a certain number of units) could be allocated to one work centre and part to another, the number of possibilities is immense. To this effect, the allocation problem can be very simple or very complex.

There are specialised techniques and applications using artificial intelligence that apply different heuristics or algorithms to try to optimise this allocation, or at least simplify it and make the task more flexible. Their use will depend on the complexity of the problem and the added value they bring to each company.

In any case, once the work centre that will process each work order has been assigned, it is usual to then develop the corresponding Gantt chart. This is the best visual tool to determine many of the parameters necessary for scheduling between work centres: workloads, processing time, flow time, idle time, availability of work centres, etc.

4. PHASE 2: Sequencing of production orders

Once the work orders have been assigned to a certain work centre, the second stage of scheduling, the sequencing of the production orders to be carried out, begins. In other words, the following questions are answered: How will the production orders be organised in a given work centre? Which ones will be carried out first? In what sequence will the parts to be produced be processed?

Very simple priority rules, or dispatching rules, can be used for this purpose, although there may be as many different priority rules as there are programmers or production managers. The most common are the following:

- FIFO (first in first out): The first part to arrive is the first to be processed.
- SPT (shortest process time): The part with the shortest operation time is processed first.
- LPT (longest process time): The part with the longest process time is processed first.
- EDD (earliest due date): The part with the earliest delivery date is processed first. The most pressing.
- SFT (shortest float time): The order with the least float is processed first, bearing in mind that the float is the difference between the delivery deadline and the processing time. In other words, this is the garment with the shortest time margin for production.

These priority rules should not be confused with scheduling indicators. They will provide a production sequence that will have certain indicators. It will then be necessary to determine which priority rule in each given case gives the best indicators (waiting time, utilisation time, delays, etc.), depending on the company's objectives (which may not always be the same).

Any sequencing problem can be described with the following notation: n/m/A/B

Where:

n: Number of parts or production orders to be processed

m: Number of machines, resources

A: Type of flow, where:

- P or "permutation": All parts have the same route, and all machines have the same sequence.
- F or "flow-shop": All parts have the same routing
- G or "general": All the parts have different routes.

B: Criteria (Fmax, Tmax, Fmig, Tmig, etc.)

In very specific cases, and under certain hypotheses, heuristics have been established to calculate the optimal option. Furthermore, the application of techniques such as artificial intelligence or the simulation of industrial processes is necessary to obtain good results.

The following are some of the simplest heuristics to apply, provided the following assumptions are met:

- Each production order may only be carried out at one known work centre.
- Once a production order has been started, it must be completed before starting another at the same work centre, thereby allowing for no interruptions.
- Two operations may not overlap on the same garment.
- Each work centre can only process one operation.
- The operation time is known and constant.
- The operation time is independent of the sequence.

4 .1. Sequence of n jobs on one machine (n/1/A/B)

Bearing in mind the above hypotheses, in the case of a set of production orders or parts to be processed on the same machine, and depending on the priority rule chosen, it has been detected that:

- SPT minimises the Fmig (average dwell time in the system) and the average stock of finished products.
- EDD minimises the minimum delay (Tmin) and the maximum delay (Tmax).
- SFT maximises the maximum slack.

In any case, and due to their ease of application, the most commonly used rules are EDD and FIFO (although they are not optimal), especially in factories where work is done to order. Overall, the SPT priority rule is the one that obtains the best results, although it is impossible to use in factories where new orders arrive continuously and with very different processing times. Orders with a very long work centre occupancy would take a long time to be produced.

4.2. Sequence of n jobs on n machines. Johnson's algorithm (n/2/F/Fmax)

For the n/m/F cases, i.e. when all parts use the same manufacturing route, there are two theorems to consider:

- Theorem 1: In an n/m/F problem, with all parts and machines available simultaneously, there is an optimal programme where the parts have the same sequence in the first two machines; therefore, the programmes considered in the search for the optimal one can be limited to those of this type.

- Theorem 2: In an n/m/F/F/Fmax problem, with all the parts and machines available simultaneously, there is an optimal programme where the parts have the same sequence in the last two machines; therefore, the programmes considered in the search for the optimal programme can be limited to those of this type.

The Johnson algorithm is established based on these two theorems. It is a heuristic used to solve process sequencing situations where two or more jobs (or production orders) must be carried out, passing through two machines. The order is always the same: some orders are carried out on the first machine, followed by some orders on the second one.

When the objective is to minimise Fmax (dwelling time in the system of the part that spend more in the system), the algorithm that gives the optimal option is the following:

- List all production orders together with the operation time on each machine.
- Select the shortest processing time. If it is a time corresponding to the first machine, it is scheduled first, while if the time corresponds to the second machine, it is scheduled at the end of the sequence. Any ties are resolved arbitrarily.
- Once the production order is scheduled, either at the beginning or at the end of the sequence, it is removed from the initial list.
- The above steps are repeated for all remaining orders, towards the middle of the sequence, until the entire list is eliminated. The final result is the optimal sequence under the set conditions.

4.3. Sequence of n jobs on three machines. Johnson's algorithm. (n/3/F/Fmax)

Johnson extended his algorithm for three machines, in cases where the second machine is of little relevance to the operation time. It is based on the creation of two dummy machines, and the algorithm follows these steps:

- Based on three work centres or machines: M1, M2 and M3.
- Two fictitious machines are created, A and B, where:
  - In "A" the work execution time will be equal to the sum of the times of M1 and M2.
  - In "B" the work execution time will be equal to the sum of the times of M2 and M3.
- From these times calculated for A and B, the procedure developed by Johnson for two machines is applied, considering that the jobs first pass through A and then through B.

- The optimum obtained for A and B will also be the optimum for the M1, M2 and M3 only if the shortest times of M1 or M3 are not less than the maximum time on the M2 (intermediate) machine.

4.4. Sequence of n jobs on n machines. Jackson's algorithm (n/2/G/Fmax)

For cases with two machines where the flow is general and, therefore, production processes can start on either machine or only operate on only one, Jackson's algorithm gives the optimal solution when the objective is to optimise Fmax (dwelling time of the part that stay longer in the system).

In this case, the sequence proposed by the algorithm is different for each machine, M1 and M2, and corresponds to the following criteria:

- For M1:
    - First, the orders that go from M1 to M2, sorted according to Johnson's algorithm.
    - Second, the orders that only operate in M1.
    - Last, the orders from M2 to M1, sorted according to Johnson's algorithm.
- For M2:
    - First, orders from M2 to M1, sorted according to Johnson's algorithm.
    - Second, orders that only operate in M2.
    - Last, orders from M1 to M2, sorted according to Johnson's algorithm.

Chapter 7: Stock management

Objective: Understand the need to manage stocks in the most efficient way and learn how to optimise stock management by using simple but robust models.

Learning outcomes:

Definition and types of stocks.

Costs involved in stock management.

ABC analysis.

Stock management policies.

Economy lot: Harris-Wilson formula or EOQ.

Economic lot with parallel production and consumption (EPQ).

Economic lot with deferred demand.

Batch management with non-homogeneous demand: Silver and Meal 's method

1. Introduction: Why do we need to manage stocks?

For a company to be able to produce, it must inevitably have stocks of all kinds of materials (raw materials, semi-finished products, spare parts, finished products, etc.). These stocks must be managed as efficiently as possible, minimising the costs involved.

However, it must be kept in mind that these costs are not only due to the acquisition of each product, but also to other aspects such as storing the products in the warehouse and the launch of each order. To this effect, placing few orders per year will decrease the costs associated with them (reception of orders, quality control, etc.), but at the same time it will involve high storage costs since the orders will be larger. It is therefore necessary to manage this situation in the most economical way possible. To do so, the ideal model to buy or manufacture each product and the quantities involved must be established. As the number of references to the warehouse increases, this task becomes more complex.

Stock management therefore includes the set of actions carried out by companies to monitor and activate purchase or production orders for the products required to be stored in the warehouse, ensuring that demand can be met at the lowest cost. Comsequently, the supply of products relies on a set of indicators that help warehouse managers to know when an order needs to be made and in what quantity. Balanced stock levels can thereby be maintained to meet the needs of each product.

Figure. Warehouse. Source: https://www.mecalux.es/blog/almacen-de-transito

Consequently, stock management is the methodology that allows you to optimise the inputs and outputs of the products involved in a company's production (raw materials, semi-finished products, finished products, spare parts, etc.). Its main objectives are:

- To always know the situation of the inventory.
- To classify the products based on their relevance.
- To optimising investment in products and their management, bearing in mind all the costs involved.
- To ensure customer service, both internal and external, by maintaining adequate stocks.

The benefits of efficient inventory management are reduced inventory management costs (storage and release of orders) and an increased return on capital invested in the warehouse. Stock management is currently a dynamic activity, which takes advantage of existing technologies and allows the best policy at each specific moment to be determined.

2. Types of stocks and costs involved

Stocks can be classified in different ways, but the main types are as follows:

- Available stock: Quantity of product in the warehouse to meet demand.
- Minimum stock: Minimum number of stocks of a product.
- Maximum stock: Maximum number of stocks of a product.
- Cycle stock: Products that serve to meet a cyclical demand, such as when in certain periods of the year the demand for a product increases considerably compared to the rest of the year.
- Safety stock (ss): Minimum number of stocks of a product, established during the planning process to counter the variability in the demand forecast or in the expected delivery time. The aim is to avoid stock outages.
- Perishable stock: Stock of products with an expiration date.
- Dead stock: Stock of products without demand.

Optimal stock: Most efficient stock level with respect to the expected demand for each product.

The existence of these stocks means the emergence of costs linked to their management. Bearing in mind the unit costs and the annual costs, they are basically the following:

- Unit acquisition cost (Ca): Cost of purchasing or manufacturing a unit of product.
- Launch cost (Cl) or *set-up*: Cost due to the launch of a purchase or manufacturing order. In the case of purchases, this cost includes those associated with receiving each order (receiving it, quality inspection, placing it in the warehouse, handling payment, etc.). In the case of manufacturing, it includes the costs of preparing each machine before launching a batch: cleaning, calibration, changing moulds, and so on.
- Possession cost (Cp): Cost due to storing a product unit for one year in the warehouse. It includes the costs of handling, deterioration of the product, conservation, rental cost of the warehouse, etc. It is generally calculated as a percentage of the acquisition cost, on the understanding that after one year the product loses this percentage of its value. The Cp will be at least equal to the opportunity cost of the money invested.
- Shortage cost (Cr): Cost due to stock breakage of a unit over the period of a year. It considers the costs afforded by the discount that must be given to the customer for delivering the product later than expected, and the costs involved in the loss of dissatisfied customers due to the shortage.
- Annual acquisition cost (KA): Annual costs due to the purchase or manufacture of the products involved. It is calculated as the unit acquisition cost of each product (Ca) by the annual demand for it (D).
- Annual Launch cost (KL): Costs due to all purchase or manufacturing order releases made at the end of the year. It is calculated as the launch cost (Cl) times the number of launches performed each year.
- Annual cost of possession (KP): Costs due to the maintenance of all products in the warehouse for one year. It is calculated as the cost of possession of each product (Cp) by the average stock at the end of the year.
- Annual breakage cost (KR): Costs due to all stock breakages during the year. It is calculated as the breakdown cost of each product (Cr) by the average breakdown volume at the end of the year.

3. ABC stock classification

The ABC inventory classification method is used to organise the distribution of products in a warehouse according to their relevance to the company, their value and their turnover. To this effect, products are placed in the warehouse based not on the required volume or quantity, but on their economic importance to the company.

This is where the Pareto principle or the 80/20 rule, according to which approximately 80% of the results are due to 20% of the causes, comes into play. Applying this principle to stock management, it means that 20% of the products generate 80% of the movements in a stock, and the remaining 80% account for just 20% of this movement. In this regard, the products can be classified into three types:

- Items A:
  These generally take up 20% of the space in the warehouse, but they are the ones with the highest turnover. They are the products in which the company invests the largest proportion of the budget and they generate approximately 80% of company income, and therefore it is essential that they are never out of stock. They are the most critical items, which is why a thorough or permanent stock control is necessary. All items A are located in low areas with direct access to the loading dock.
- Items B:
  These represent approximately 30% of the references in the warehouse and are the products with an average rotation. They have a lower turnover and value than items A. Their stock management can be carried out by establishing minimum and maximum stocks to be met. In the warehouse, they are stored at medium height with medium difficulty access.
- Items C:
  These often represent over 50% of the stored products, even though they are the ones in least demand and with the lowest turnover. Therefore, it is not necessary to allocate a large amount of resources to their control and management, and a significant safety stock is often present. Items C are in the less accessible areas of the warehouse and furthest away from the loading dock.

Different indicators can be used to classify products according to the ABC system, depending on what the company's objective is. The three main indicators are:

- Unit cost of each product:
  Sorting the products according to the unit cost of each makes sense when the cost differences between them are very high. In this way, more attention can be paid to items A, the products that have a higher cost.
- Total inventory cost:
  Products are ordered according to inventory cost, that is the unit cost of each product by the number of products in the warehouse. This way a much more accurate approximation of the actual costs of the inventory is achieved, although the inventory level of each product requires constant updating.
- Inventory usage and value:
  This is the most common method, since it considers not only the total cost of the inventory, but also the rotation of the products. Items A items will be the greatest quantity products in the warehouse and the ones with the highest turnover.

Other characteristics can also be present, such as the profit margin of each product and the costs in the event of a stockout.

Whichever indicator is used, the data obtained for all products in a warehouse are organised from largest to smallest, and in general, the 20% most relevant products will be items A, the next 30% items Bs, and the last 50% items Cs.

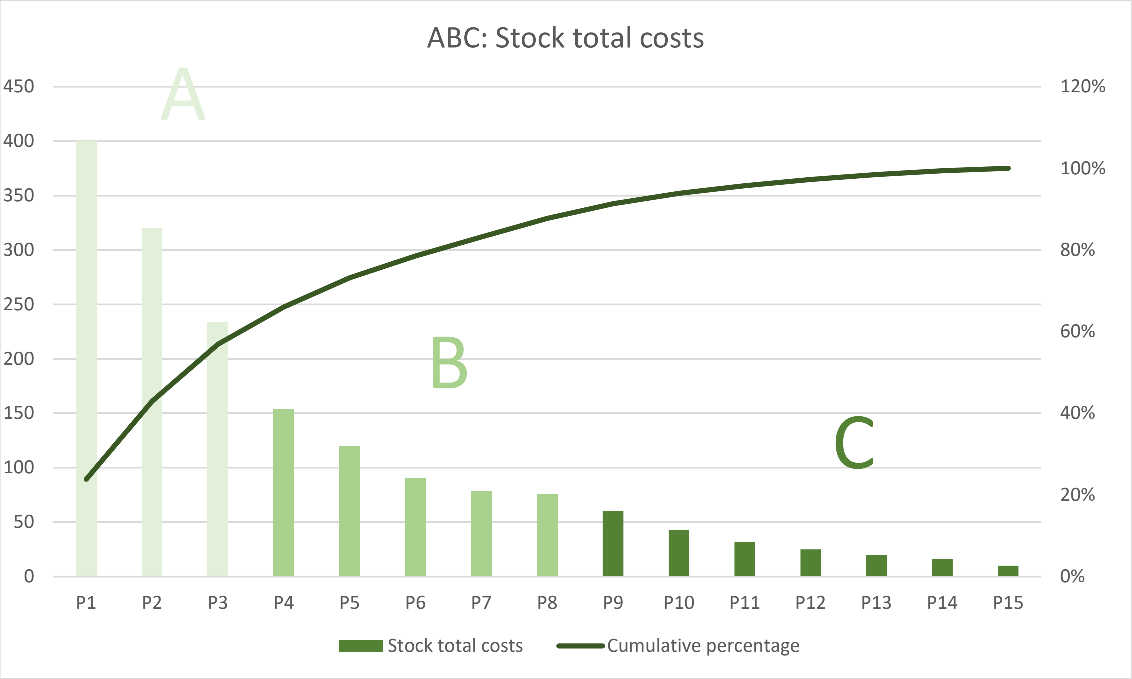


Figure. Classification of products according to the ABC analysis (total inventory costs).

4. Stock management policies

The most basic stock management is based on observing supply rules that determine the level of stocks of a product or the period in which an order must be placed. Determining these rules correctly is essential to avoid undesirable accumulations of products.

- Coverage level (S): Maximum number of products that a warehouse can handle, usually limited by its maximum capacity.
- Purchase/Manufacturing Lot (Q): Quantity of products that are purchased or manufactured in each order.
- Review period (T): Period in which the status of available stock is reviewed to decide whether to purchase or manufacture more of the product.
- Delivery time (t) or lead-time: Time between when a purchase order is placed and when the product arrives at the warehouse or, in the case of manufacturing, between when production starts and when the product is available for the next productive task.
- Order point(s): Reference level used to place an order, purchase or production, when the actual stock is less than this level. The order point will be directly proportional to the delivery or manufacturing time of the product and the expected demand during this period.

Order point (s) = demand (D) x lead time (t)

Keeping the above variables in mind, different general strategies for inventory management can be determined. To this effect, depending on whether the stock review is continuous or periodic and whether the purchase or manufacturing quantity is fixed or limited to a maximum coverage, the four most common options will be established.

	Fixed Quantity (Q)	Refill to coverage (S)
Continuous review	(s,Q) When the level is less than s, Q is requested	(s,S) When the level is less than s, up to S is requested
Periodic Review (T)	(T,Q) No information system	(T,S) For each period T, up to S is requested

Table: General strategies for stock management.

On analysing the previous strategies, the following questions can be considered: Does it make sense to produce in batches? What size should the lots be? The frequent application of batch production among companies leads us to assume that it must have clear advantages.

Processes are usually scheduled in batches. Manufacturing in larger batches does absorb launch or set up costs across a larger number of units. Therefore, manufacturing in large batches results in a decrease in cost per unit.

The figure below shows an example with a launch cost (Cl) of €100 and a unit acquisition (or production) cost (Ca) of €20. The total unit cost therefore includes a Cl-dependent part and a Ca-dependent part. Specifically:

$$\text{Total unit cost} = \frac{\text{Cl} \cdot (\text{Q} \cdot \text{Ca})}{\text{Q}}$$

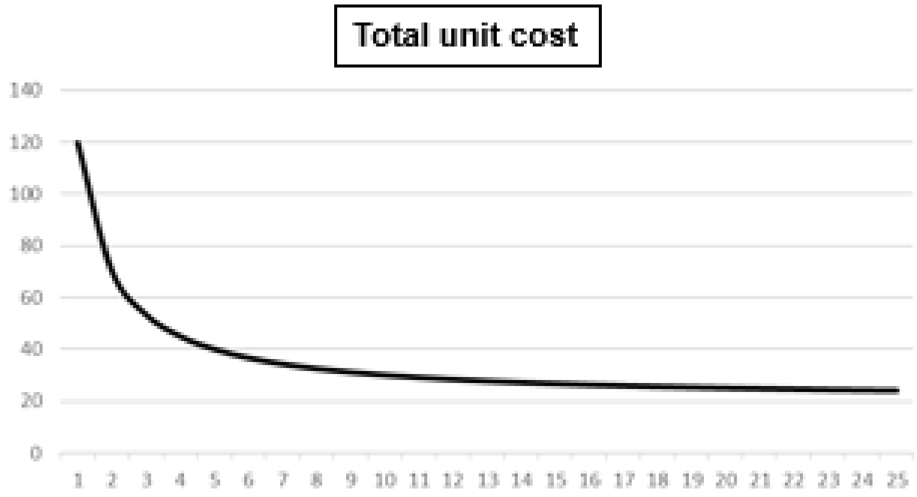


Figure. Total unit cost based on lot size.

Looking at the figure, if the minimum lot (one unit) is manufactured, the cost per unit is €120. If the lot is two units, the cost per unit is €80. How the unit cost function is asymptotic in the variable cost (€20) is thereby verified. If the batch were infinite, the unit cost would be just the variable cost. The launch cost (Cl) would be distributed among infinite units and, therefore, it would disappear.

In any scenario, looking closely there are two possible strategies to reduce the total unit cost:

- Increase batch size:
  The first strategy leads to manufacturing in large batches, and therefore to having longer lead times and high WIP (work in process). All in all, this strategy negatively affects product quality. This is what is known as the mass production paradigm (first implemented by Ford in the USA at the beginning of the 20th century).
- Decrease the launch cost (Cl) or set-up cost:
  The second strategy is better in terms of flexibility, quick access to the customer and quality. It facilitates changing the process if the customer requires a variation or customisation of the product. Effort in terms of engineering is required to reduce, and ideally eliminate, set-up costs. This strategy was first introduced by Toyota in Japan in the 1960s, leading to what is known as the lean production paradigm (lean production).

5. Economic lot: Harris-Wilson formula or EOQ

If a certain annual demand for a product is predicted, it will not be optimal to place a multitude of orders, since the annual launch costs will be very high. Neither will it be optimal to place excessively high orders, given that the annual cost of possession would be very high. Between these two extremes, there is surely a purchase lot size that allows costs to be optimised, without excessive order releases or excessive stock in the warehouse.

The economic batch of production (economic order quantity – EOQ) is a mathematical model defined by Harris and Wilson that allows the size of the purchase or production lot that minimises the costs of stock management to be established. It can be applied if the following assumptions are met:

- The horizon of study is unlimited, meaning that the process continues indefinitely.
- Demand is continuous and homogeneous over time. Companies generally work with the annual demand (D).
- The delivery time (lead time) of the supplier is constant and known.
- The products are produced and sold simultaneously.
- Out of stock is not accepted.
- The lot to be ordered (Q) is constant and enters the system in a block and instantaneously.
- The total costs of inventory management are calculated from the unit holding costs (Cp) and the costs of each release or order (Cl). These are constant over time.
- There are no discounts on the unit acquisition or production costs (Ca) depending on the volume.

The model starts from the analysis of the total annual costs of stock management, bearing in mind the following variables:

KT = Annual inventory management costs

KL = Annual costs due to releases of each batch

KP = Annual cost of possession

KA = Annual acquisition costs

D = Annual demand for the product

Ca = Unit cost of acquisition or production

Cp = Unit cost of annual possession; generally calculated as a percentage of Ca

Cl = Cost of each launch

Q = Purchase or manufacturing batch (* indicates optimal)

N = Number of annual releases

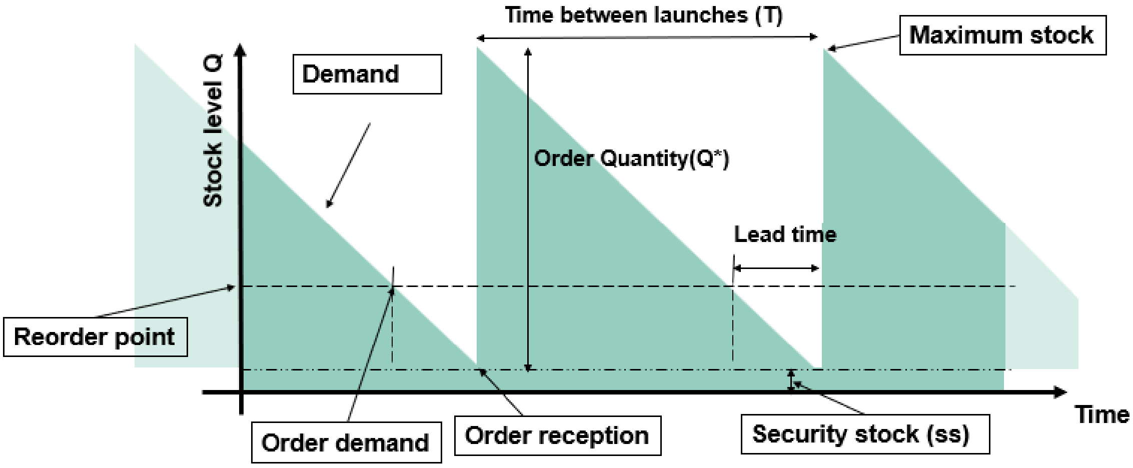


Figure: Economic lot model (EOQ).

In consequence, the annual costs of stock management can be defined as:

$$\mathrm{KT} = \mathrm{KA} + \mathrm{KL} + \mathrm{KP} = \ \mathrm{Ca} \cdot \mathrm{D} \ + \ \mathrm{Cl} \cdot \mathrm{N} \ + \ \mathrm{Cp} \cdot \text{Average stock} \ = \ \mathrm{Ca} \cdot \mathrm{D} + \mathrm{Cl} \cdot \frac{\mathrm{D}}{\mathrm{Q}} + \mathrm{Cp} \cdot \frac{\mathrm{Q}}{2}$$

To optimize costs with respect to the lot to be chosen, the equation is derived with respect to Q and set to zero.

$$\frac{d\ KT}{d\ Q} = -Cl \cdot \frac{D}{Q^2} + \frac{Cp}{2} = 0$$

It turns out that the optimal economic lot Q* is equal to:

$$Q^* = \sqrt{\frac{2 \cdot \mathrm{D} \cdot \mathrm{Cl}}{\mathrm{Cp}}}$$

And in this way:

$$\text{Number of annual launches (N)} = \frac{\mathrm{D}}{\mathrm{Q}}$$

$$\text{Time between launches (T)} = \frac{\mathrm{Q} \cdot \text{Anual operative days}}{\mathrm{D}}$$

It should be kept in mind that this is a very "robust" model, which means that even if there are significant changes in the variables that are assumed to be constant (for example, the annual demand), the increase in cost totals with respect to the optimal point is small. Therefore, even though it is very difficult for the assumptions to be fulfilled in their entirety, the model can give us a lot size that is close to the optimal one for managing the orders of each product.

An extension of this model is the “EOQ with quantity discount” model, adjusted to different discount scenarios in the purchase of products. This model considers an initial Q obtained by applying the same formula as the EOQ model for each discounted quantity range. If the initial Q falls within this range (of units to be bought at a discount) it is maintained; otherwise, it must be replaced by the amount closest to obtaining the discount in that range. Once the Q has been calculated for each discount range, the total annual inventory management costs for each case must be calculated. The discount range with the lowest total cost will be selected.

6. Economic lot with parallel production and consumption (EPQ)

A variation of the previous model, suitable for application in the case of manufacturing, is when product consumption and manufacturing occur simultaneously over a period. In other words, there is a certain period in which the product is being manufactured and consumed in parallel. In this case, it is also possible to calculate the optimal manufacturing lot that optimises total management costs. This model is called economic production quantity (EPQ).

First, you will need to define the following two rates:

d = Daily demand, equivalent to the annual demand (D) divided by the working days

p = Daily production

The rate *p* must be greater than *d*, otherwise the expected annual demand could not be met with production.

Two distinct periods are established:

T1 = Time when it is produced and consumed at the same time, and therefore the resulting rate of manufacture is *p-d*.

T2 = Time when it is only consumed, and therefore the resulting rate of consumption is *d*.

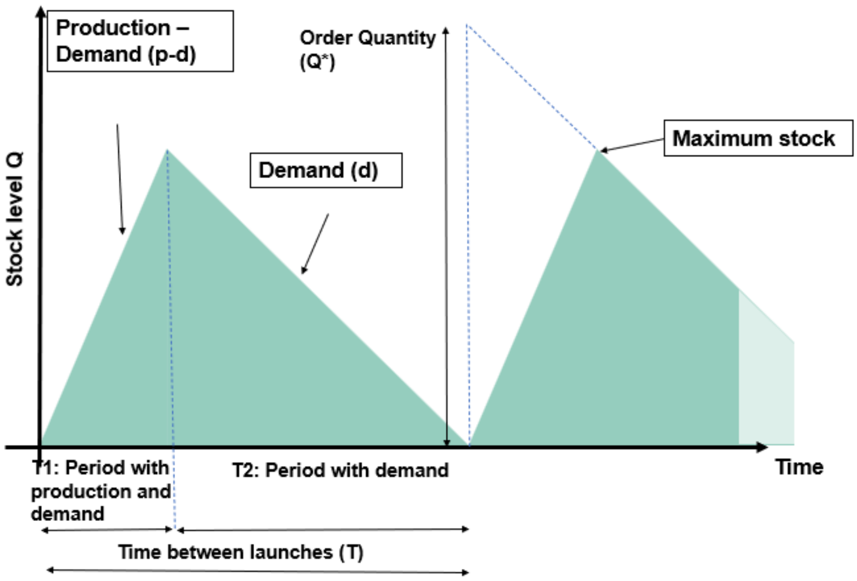

Production – Demand (p-d)
Order Quantity (Q*)
Maximum stock
Demand (d)
Stock level Q
T1: Period with production and demand
T2: Period with demand
Time
Time between launches (T)

Figure. Economic batch model with parallel production and consumption (EPQ).

Following the same procedures as above, you would have:

$$\mathrm{KT} = \mathrm{KA} + \mathrm{KL} + \mathrm{KP} = \mathrm{Ca} \cdot \mathrm{D} + \mathrm{Cl} \cdot \mathrm{N} + \mathrm{Cp} \cdot \text{Average stock} =$$

$$= \mathrm{Ca} \cdot \mathrm{D} + \mathrm{Cl} \cdot \frac{D}{Q} + \mathrm{Cp} \cdot \frac{Q}{2} \cdot \frac{(p-d)}{p}$$

And operating in the same way, the resulting optimal lot would be:

$$\mathrm{Q}^* = \sqrt{\frac{2 \cdot \mathrm{D} \cdot \mathrm{Cl}}{\mathrm{Cp} \cdot \frac{p-d}{p}}}$$

The maximum inventory reached would be:

$$\mathrm{I\ max} = \mathrm{Q} \cdot \frac{(\mathrm{p-d})}{\mathrm{p}}$$

The different times can be determined from this maximum inventory, in addition to the production and consumption rates, as follows:

$$\mathrm{T1} = \frac{I\ max}{p-d} = \frac{\mathrm{Q}}{\mathrm{p}}$$

$$\mathrm{T2} = \frac{I\ max}{d} = \frac{\mathrm{Q}\ (\mathrm{p}-\mathrm{d})}{\mathrm{p} \cdot \mathrm{d}}$$

As in the previous case, because this is a “robust” model it is used to determine the approximate size of the manufacturing batches of each product to minimise management costs.

7 . Economic lot with deferred demand

When it is feasible to defer the demand – that is, to deliver it with a certain delay compared to the expected date – it is also possible to model this and optimise the purchase or manufacturing lot.

In this case, it must be kept in mind that deferral will also entail a cost linked to the discounts that must be given to customers for late delivery or, as the case may be, to the loss of customers due to this situation. Therefore, it will be necessary to establish some additional variables:

KR = Annual stockout costs

Cr = Annual break unit cost

Following the same procedures as above, you would have:

$$\mathrm{KT} = \mathrm{KA} + \mathrm{KL} + \mathrm{KP} + \mathrm{KR} = \mathrm{Ca} \cdot \mathrm{D} + \mathrm{Cl} \cdot \mathrm{N} + \mathrm{Cp} \cdot \text{Average stock} + \mathrm{Cr} \cdot \mathrm{VR}$$

Where VR = Breakdown Volume

Bearing in mind the periods for which stock will be available or will be a stock shortness, we would get:

$$\mathrm{KT} = \mathrm{KA} + \mathrm{KL} + \mathrm{KP} + \mathrm{KR} = \mathrm{Ca} \cdot \mathrm{D} + \mathrm{Cl} \cdot \frac{D}{Q} + \mathrm{Cp} \cdot \frac{(Q - VR)^2}{2 \cdot Q} + \mathrm{Cr} \cdot \frac{VR^2}{2 \cdot Q}$$

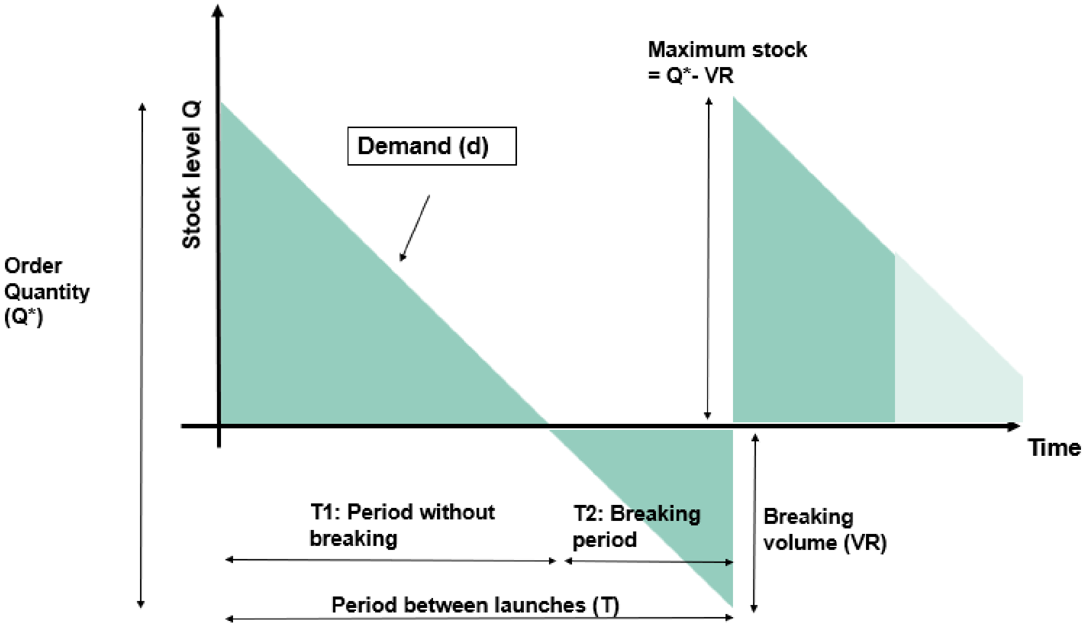


Figure: Economic batch model with deferred demand production.

Optimising the previous costs solves:

$$\mathrm{Q}^{*} = \sqrt{\frac{2 \cdot \mathrm{D} \cdot \mathrm{Cl} \cdot (\mathrm{Cp} + \mathrm{Cr})}{\mathrm{Cp} \cdot \mathrm{Cr}}}$$

$$\mathrm{VR}^{*} = \sqrt{\frac{2 \cdot \mathrm{D} \cdot \mathrm{Cl} \cdot \mathrm{Cp}}{\mathrm{Cr} \cdot (\mathrm{Cp} + \mathrm{Cr})}}$$

It is interesting to note that not only is an optimal manufacturing lot (Q*) calculated but also an optimal rupture volume (VR*). Therefore, the established break is not directly due to a one-off lack of stocks but is planned a priori. In other words, inventory management establishes a priori that there will be a set of customers whose orders cannot be delivered on time. This is because in these cases annual breakdown costs are lower than annual holding costs, i.e., it would be more expensive to have stock for all customers than to not anticipate that some deliveries of the product will be delayed.

8. Batch management with inhomogeneous demand: Silver and Meal method

One of the most common problems in inventory management is the high variability in demand forecasting. The previous models are developed considering a constant demand during the year, which is often not so given that many products have a high seasonality.

In this case, it is useful to calculate the coefficient of variability, which will serve to determine whether the previous models can be used or not. Specifically, it refers to:

$$VC = \frac{H \cdot \sum_{t=1}^{H} D_t^2}{(\sum_{t=1}^{H} {D_t}^{2})} - 1$$

Where:

H= Periods (weeks, months, etc.) in which the demand is known

D_{t} =Demand in each period t

To this effect, it is established that when VC ≤ 0.25 the demand can be considered homogeneous, and the Harris-Wilson model (EOQ) is recommended to calculate the optimal lot. Otherwise, when VC > 0.25, the Silver-Meal heuristic is recommended.

The Silver-Meal heuristic will not give us the optimal lot, but it will give us a good solution that could come close to this optimum. In addition and bearing in mind that it is used when the demand is very variable between periods, it will provide us with different batches depending on what this demand is. In other words, it does not provide a single optimal lot to be used throughout the period, as in the previous cases, but several.

The Silver-Meal heuristic is based on minimising stock management costs per period bearing in mind that they can be defined as follows:

$$KT_t = \frac{KL+KP}{t} = \frac{1}{t} \cdot (Cl + Cp \cdot \sum_{j=1}^{t} (j-1) \cdot D_j)$$

where:

$\mathrm{KT_t}$ = Average cost per period

T = Number of periods covered by Q

$\mathrm{D_t}$ = Expected demand during period j

Specifically, the heuristic establishes the need to calculate the average cost per period for each possible lot (Q). In other words, it is necessary to calculate this cost when Q only covers the demand of the first period, when it covers that of the first 2, when it covers that of the first 3, etc.

Once all the possibilities have been calculated, the one with the lowest cost per period is chosen and it is established that the first batch Q will cover the demand for all the affected periods. It should be kept in mind that once detected that the average cost per period is beginning to grow, it will continue to grow in all subsequent periods.

From this point, the heuristic is restarted for the following periods.

Chapter 8. Just in time

Objective: Understand the objective of the just-in-time production philosophy and some of the support tools that facilitate its implementation.

Learning outcomes:

Definition and origin of just-in-time.

Pull system versus push system.

Support tools for the implementation of just-in-time.

1. Introduction: definition and origin of the organisation system *just in time*

The just-in-time (JiT) organisation system is a demand-driven production philosophy. Its main objective is to increase profits by eliminating all non-essential costs in the production process while still satisfying customer needs.

As the expression itself indicates, JiT consists of ensuring that not only do the supplies reach the production system at the right time, but also that the final products reach the customer at the right time. In other words, the raw materials and components must be available for production in the quantity and at the time they are needed, and the final product must be delivered to the customer in the quantity and at the time previously agreed, at the minimum cost.

The origin of the JiT concept is attributed to the founder of the Japanese Toyota brand, Sakichi Toyoda, his son Kiichiro and the industrial engineer Taiichi Ohno, who implemented it at the carmaker brand as responsible of the machine workshop. For this reason, the JiT system is considered one of the pillars of the TPS (Toyota production system).

After Japan's defeat in the Second World War and with the consequent increase in production costs, Toyota decided to redefine its production system vis-á-vis strong external competition in terms of high production volumes and low costs due to economies of scale, especially from the North American industry.

Toyota's commitment was based on optimising the production process by reducing costs that did not add value to the product. Specifically, the engineer Ohno made a list of seven costs or waste (*muda*, in Japanese), which has subsequently been widely disseminated:

1. Overproduction
2. Waiting
3. Transporting
4. Processing
5. Inventory
6. Motion
7. Defects/reworks

Among all the costs, the one considered the most dangerous is overstock, as its presence can mask and make undetectable other costs or waste in the company.

Overstock is often caused by poor materials management, for example buying more than necessary to take advantage of an offer or discount. Excess stock entails costs associated with its maintenance, representing an immobilised capital that does not add value to the product and is difficult to manage. This excess can also lead to a drop in the quality levels of the products since it can "hide" production errors, defects that are detected later in the production process. In Ohno's own words, *“the more inventory a company has…the less likely they will have what they need*.”

The analogy of the "river of inventory" is often used to represent the hidden waste due to excess inventory, where the company is represented by a ship sailing through water, i.e. through the level of inventory. As the water level goes down, more and more rocks appear; i.e., potential problems (and costs), which had previously been hidden under the water, and include lack of quality, changeover times, batch size, and so on.

The JiT philosophy proposes the removal of these rocks or problems, as opposed to increasing the water level or stock to hide them.

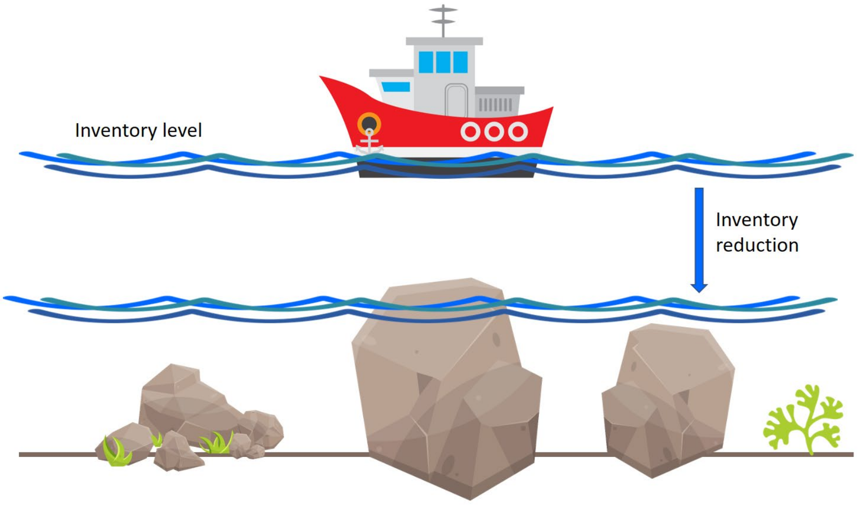


Figure. River of inventory. Source: own elaboration.

2. Pull system versus push system

Mass production systems have traditionally been based on the push philosophy, where work orders for the different stages of the production process are generated in response to a forecast of demand and firm orders. It is often said that the product is pushed from the demand for materials to the customer.

Contrarily, the pull manufacturing system plans from the analysis of demand and stock management. Each stage of the production process collects the elements or parts from the previous one at the time and in the quantity needed, thus eliminating stock and overproduction costs. It is the customer's order that initiates the production process and stretches the entire process in a concatenated manner.

Unlike the push system, where planning is centralised and, therefore, the operational production areas have no information on the real customer demand, in the pull system planning is focused on one of the production processes, which will set the pace of work of the entire production chain. This reference process is called *takt* and is usually located near the end of the chain, when the product is awaiting final assembly.

The processes prior to the *takt* process can be regulated by means of a system of cards, called *kanban*, which indicate orders on demand between a customer process and a supplier process up to the start of the production process. These cards are usually for production, transport of materials or urgent tasks.

This card system involves small regulating warehouses between the client process and the supplier process that generate the necessary information and visibility for the latter to know what it must manufacture at any given time. The way it works is simple. While the transport *kanbans* generate orders for the transport of material stock from the supplier process to the customer process, the production *kanban* generates manufacturing orders based on the demand of the customer process.

The use of *kanban* cards has various benefits:

- Reduction in the amount of material involved in the process
- Savings in the space required to store the material
- Reduction in replenishment and consumption runs
- Savings in the time required to supply a material
- Elimination of stock breakage
- Facilitating the visual management of warehouses and the visualisation of bottlenecks.

3. Support tools

The implementation of the *kanban* card system within the JiT is not a simple task since the material flow within the production line must first be constant and even. Large fluctuations make it difficult to integrate the system. Various tools are available to facilitate the implementation of *kanban*, some of which are described below.

3.1. Heijunka

The *heijunka* system departs from the traditional production model, which is based on the production of large batch sizes. Batch processing has been a widely used method for organising the manufacturing process since the invention of mass production. This type of processing is based on the production of large batches of products without considering fluctuating customer demand.

The aim of *heijunka* is precisely the opposite: to obtain a flexible production adapted to the needs of demand that avoids the creation of intermediate stocks between two manufacturing stages, thereby reducing storage and maintenance costs. In other words, it adjusts batch sizes so that they match as accurately as possible the real demand for the product.

Heijunka is obviously not applicable if there is no or little product variation. Therefore, the practical management of *heijunka* requires a good understanding of customer demand and the effects of this demand on processes.

3.2. Layout

The plant layout is the physical arrangement of the elements that make up a production line. It includes the spaces necessary for movements, storage and all the activities carried out along the production line.

The objective of an optimal plant layout is to optimise both space and working time. It should be kept in mind that there is no single model of plant layout, but in general terms there are three types: product layout, process layout and fixed position layout.

Distribution by product is applied when production is continuous or repetitive. The workstations are placed next to each other and follow the order of the operations to be carried out in the production process. It is a fully automated process, an example of which would be a self-service restaurant.

Distribution by process is carried out when the size of the batches is variable and there is a wide variety of products. Unlike distribution by product, the machinery and workers are grouped by similarity of operations and the products only circulate through the necessary workshops or departments. An example would be a hospital, where patients only go to the departments where they need to be treated.

The JiT philosophy avoids process distribution. The fact that this philosophy is underpinned by a customer-driven approach means that distribution must be flexible, and the number of employees must be adapted to the demand. According to the JiT philosophy, the most common type is the U-shaped layout. The main characteristic of this typology is that the input and output of the production process are distributed in parallel, allowing the number of workers to be adapted to changes in demand. Therefore, while the U-shaped distribution allows for greater flexibility, it also requires greater operator versatility.

Last, the fixed position distribution is applied when the product cannot be moved because of its size or weight, requiring machinery and workers that are mobile to carry out the production process. An example would be the construction of a ship in a shipyard.

3.3. SMED (Single-Minute Exchange of Die)

SMED is considered a continuous improvement tool that aims to reduce set-up/changeover time within the production process. Improved process efficiency will lead to a reduced risk of material defects and machine breakdowns. The reduced time can also be used to increase productivity and reduce material stock. In the words of the tool's creator, *Shigeo Shingo*, "the fastest way to change a tool is not to change it at all".

The implementation of SMED follows 7 stages:

1. Preliminary preparation: collect as much information as possible on the product and the machinery involved in the production process, as well as historical data on its operation.
2. Analyse the activity on which the SMED will focus: analyse in detail how the change takes place and the current quality of the product being produced.

3. Differentiate between internal and external preparation: internal preparation is when it can be carried out with the machinery stopped, and external preparation is when it can be carried out with the machinery running.
4. Organise external preparations: the aim of this stage is to check that all external preparations will be ready once the machinery is running.
5. Convert internal preparation to external: make improvement proposals to convert the internal preparation time into a structured plan of action.
6. Reduce internal setup times: make improvement proposals to reduce internal setup times in a structured plan of action.
7. Follow-up: once the previous steps have been carried out, check for possible deviations from the improvement planning and, if necessary, carry out corrective actions.

Chapter 9: Queuing systems

Objective: To Identify and model productive situations with wait in line queues, based on knowledge of the variables to be analysed, and optimisation systems that minimise waiting times.

Learning outcomes:

Identification of the queuing phenomenon and its main characteristics.

Modelling waiting processes.

Analysis of costs associated with wait in line queues.

Psychology associated with waiting times.

Operational analysis of waiting systems: waiting time, queue length and server occupancy.

Simulation of queuing models.

1. Introduction

The figure shows a queue of people who are waiting to be served by a server, who is currently serving another customer. It is important to note that these people are waiting because they want to receive a service. They are willing to wait because this service brings value to them. All these people must wait until the server has attended to those in front of them. This wait is very different from waiting for a bus to arrive. In this case there is no queue, but simply one or several people waiting for the bus to arrive at the same time. There is also no queue when some students are waiting in the classroom for the teacher to arrive and start the class.

However, once the bus arrives, they will have to queue to get on it. They cannot all get on it at once. The door has the capacity for one customer to go through it every three seconds. In this case, the service would be "get on the bus", and the server (in this case, the bus door) can only serve an average of one customer every three seconds.

It is necessary to identify exactly what a queue is. It is not just people waiting. It is effectively a line of people waiting, but they are waiting because the server is serving customers individually.

Figure. Queue of people.

In the case of the queue in the figure above, each person will be served when it is their turn. The unit of analysis is the person. More generically, the unit that waits to receive a certain service from a server is called an "arrival".

The same laws can be applied to both processes for providing a service and to processes in the world of manufacturing physical goods. A change in terminology, however, is necessary. For example, in the environment of the physical goods production industry we can talk about stocks (quantity of units waiting in front of a machine or department to be processed), while the people waiting in the hairdressers are defined as people in the queue or in the line, and not as stock.

The following table contains some situations in which queues occur in the industrial and service environments.

Situation	**Queue arrivals**	**Service requested**
Motorway toll	Driver who wants to pay	To pay
Supermarket	people	Pay for the purchase
Physician	Patients	Medical care
Port	Ships	Cargo loading
Mechanisation workshop	Parts to be machined	Machining (welding, etc.)
Call center	Calling customers	Attention to the call
Workshop	Broken down machines	Repair of the machine

Table. Examples of queuing situations.

2. Costs associated with a queuing system

The costs incurred by the server for providing the service can be differentiated from the costs that the client "pays" for waiting. The figure shows how these costs are based on the quality of the service provided. Quality can be measured, for example, as waiting time in the queue. To this effect, if you want to provide an excellent service (little waiting on the part of the customer), you will be on the right side of the figure and the costs to the supplier will be high. In the case of the checkout station in the supermarket, the more cash registers there are (each manned by one person), the higher the cost to serve. If you look at the costs incurred by the customer (which we can think of as proportional to the waiting time), you can see how low the costs are on the right side of the figure.

The behaviour of the two cost functions is therefore very different: while the costs of providing the service increase with service quality, the costs associated with waiting decrease. Therefore, you need to look for the total cost, and its optimum.

You could even go a step further and assume that the provider wants to increase the quality of the service, under the assumption that this will attract even more market. This means moving the optimal point to the right of the figure. In this case, you can choose to either decrease the cost slope of providing the service, which depends directly on the provider, or try to lower the customer's cost curve. The latter will be more complicated because it does not depend directly on the server.

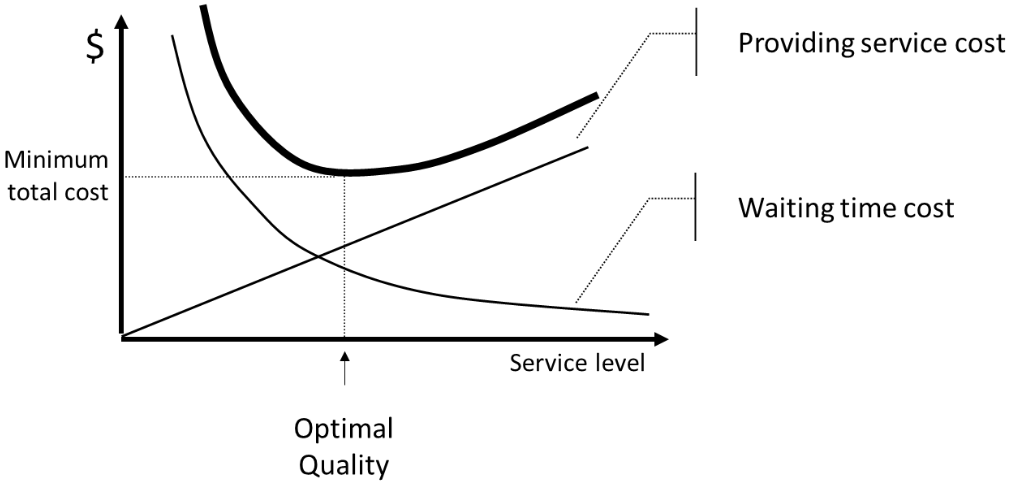


Figure. Costs associated with the phenomena of waiting queues.

3. Psychology of waiting time

It is well known that the perception of waiting is very different depending on different circumstances. It is the difference between kronos and kairos. Kronos is quantitative and is measured in days, or in any segmentation of the day (in a unit of time), while kairos refers to the content of that time. Thus, an hour doing something boring or something fun is always an hour (the same kronos), but the kairos is not the same.

There are some things that affect *kairos* over which the provider has control, and others over which he has none.

3.1. Factors in which the company or supplier can intervene

The perception of time is lengthened in situations such as:

- Queues, in which there is a perception of unfairness (for example, if there are people pushing in).
- The wait is uncomfortable.
- It is not known how long you will have to wait.
- The reason for the wait is not known.

In these cases, the company can intervene to optimise the perception of the waiting time. The company can prevent people from jumping the queue or give an estimate as to how long they will have to wait, or information on what they are waiting for.

3.2. Factors relating to the client

There is nothing the provider can do about these. Here are some examples in which the perception of waiting is long:

- When the perception of the value of the service is low.
- When waiting alone (no one else is in line).
- When the customer is angry.
- When the client is distressed.

3.3. Factors relating to both the server and the client

Situations where the perception of the length of the wait depends on both the customer and the supplier.

4. Characteristics of a waiting system

In every system, there are three parts that need to be analysed.

- What we call the "arrival", which is the unit that must be served by a server.
- The queue itself: the people (generally “arrivals”) who are queuing.
- The server

The figure below shows a queuing system and its three main parts. On the left, there is a set of elements (arrivals) that are the natural market of the system. It is the population that will eventually order the service. At some point, these “arrivals” will be queuing.

The system is composed of the queue, the server and the arrivals that are being attended to or served at a given time.

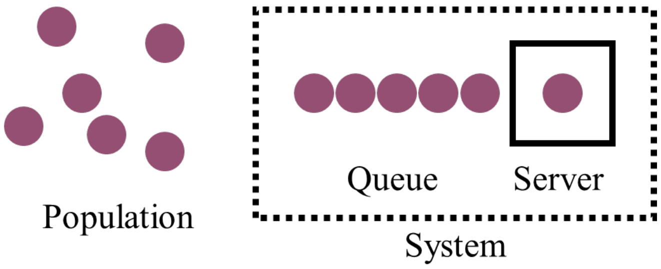


Figure. Parts of a queuing system.

Variability is what explains the formation of queues. The variability is basically in the population or in the server. These are the two sources of variability. It has been verified that there is a certain variability located at the entrance to the system and on observing how customers arrive. The time between two consecutive customers is variable. There is also variability in the service time to each customer.

If there is no variability in the entrances (if the time between arrivals is fixed) or in the service (the serving time is constant), then there is no queue. A customer will arrive, be served at a certain time, and when the next customer arrives there will be no one in line and he will be served at once.

4.1. Characteristics of arrivals

You need to know some basic information such as whether the population that can potentially request the service is limited or unlimited. A neighbourhood hair salon has a small potential market in the neighbourhood in which it operates. An online service can have a virtually unlimited potential market.

You need to see the pattern of arrivals into the system. Each customer decides to queue up to be served. It may be that the pattern is deterministic (which would be the case of some customers who decide to go to the hairdresser every first Friday of the month, for example). Most often the pattern is random, following a law of probability. Among random patterns, the most common is that determined by Poisson's law. This occurs when there are many reasons that explain why someone decides to queue, and none of them explain the behaviour any more than the others.

Poisson's law determines the probability that *x* "arrivals" will arrive during a unit of time.

$$P(x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

- P(X): probability that *x arrivals* will arrive during a unit of time.
- λ : arrival rate. Average speed at which customers arrive. Therefore, it is measured in a ratio of arrivals per unit of time. In the hairdresser's, for example, the arrival rate would be three customers per hour, if you consider that this is the average of customers arriving every hour.

4.2. Characteristics of queues

You need to look at basically two things: the queue size and the queue formation policy.

1. Length of the queue: It is necessary to know if the queue has a maximum or if it can be unlimited. In practice, there are factors that limit the length of the queue. For example, in a hair salon it could be the number of chairs in the waiting area.
2. Another issue is to know the waiting policy when the "arrival" starts queuing. Some of these queuing policies or disciplines are:
   - FIFO (*first in, first out*)
   - LIFO (*last in, first out*)

- Shorter service time
- Severity in an emergency department
- Others

4.3. Server features

Several aspects need to be analysed:

- System configuration: whether there is one channel or if it is multi-channel; if it has one phase or more, and so on.
- Service rate: how fast the server is serving.
- Service pattern: either constant or deterministic. The first would be the case of a coffee vending machine that always takes 35 seconds to serve a coffee. However, the serving time is often random, following a negative exponential distribution. This is so when there are many causes that explain the length of service, but none predominates over the others.[1]

  The negative exponential law is expressed as:

  $$F(t) = 1 - e^{-\mu t}$$

  - F(t): Probability that service duration is *t* or less
  - μ : Service rate

5. Measures to analyse queuing phenomena

Some of the measures to determine if a system is adequately dimensioned or not can:

- Average time a person (or object) waits in the queue.
- Average line length.
- Average time a customer is in the system (includes queue and service time).
- Average number of people in the system.
- Server performance or occupancy.
- Probability that the server is idle.

6. Types of queue models

We will use Kendall's notation, which identifies the models according to three characteristics: A/B/C.

- A: Distribution of "arrivals".
- B: Service time distribution.
- C: Number of channels.

[1] The Poisson distribution and the negative exponential are in fact related. In a process where arrivals follow the Poisson distribution, the inter-arrival time is calculated using the negative exponential distribution.
The Poisson distribution calculates how many arrivals there are in a unit of time, while the negative exponential determines how much time there will be between two arrivals.

When the distribution is Poisson, or negative exponential, it is said to be Markov, and the letter M is used. When a distribution is constant or deterministic, the letter D is used.

For each queue model (determined by the trio A/B/C), there is a set of formulas to find the different parameters of a system. You only need to know λ, µ and the number of channels. This is a straightforward way to analyse a queuing model. There are three ways to solve the queuing theory exercises. The first is to directly apply the formulas once the model is determined.

The system in a hairdresser's with a single hairdresser would be M/M/1. The first M means that the arrivals are distributed according to a Poisson (hence Markov) function. The second M refers to the fact that the service time distribution is a negative exponential (hence, also Markov), and the 1 means that there is only one server.

The model type influences system performance. The figure below shows two different models. The first (the one on the left) is M/M/1, and the second, M/M/s. It must be assumed that the two scenarios have the same entry rate (λ: the demand). It is also assumed that the service capacity is the same: in the first case it is µ, and in the second each server has capacity µ/s (so the capacity to serve is the same). Having said that, the waiting time in the second case is less. Therefore, the model influences the performance.

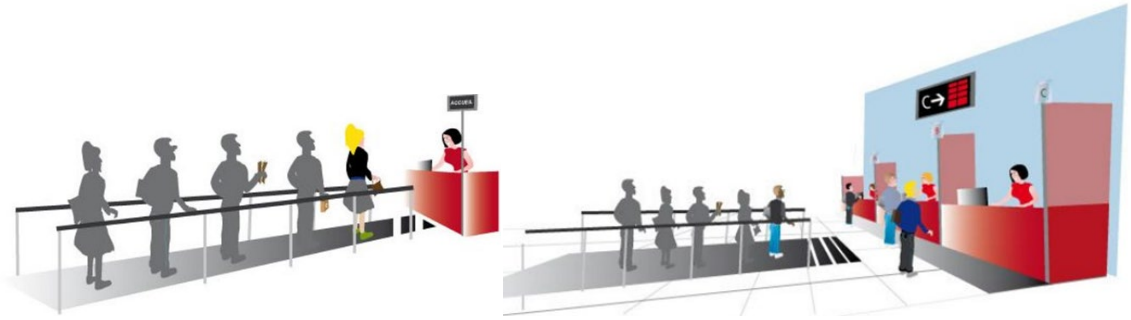


Figure. Two queue models. The one on the left is M/M/1 and the one on the right is M/M/s.

The following table presents different models depending on the characteristics of each.

Model	Example	Technical name	Arrivals	Service	Channels
A	Hair salon	M/M/1	Poisson	Negative exponential	1
B	Billing windows	M/M/s	Poisson	Negative exponential	s
C	Car wash tunnel	M/D/1	Poisson	constant	1
D					

Table. Description of model types.

Using Kendall's notation:

- A: random variable “time between arrivals”
- λ : average rate of arrivals (number of arrivals per unit time)
- Esp [A] = a = 1/ λ : average time between arrivals
- S: random variable “service time”
- µ : average service rate (number of arrivals served per time unit)
- Esp [S] = s = 1/ µ : average service time
- ρ : server performance or occupancy or utilization
- Nq or Lq: average number of people in the queue (*queue*)
- Ns or Ls: average number of people in the system (*system*)
- Wq: average waiting time in the queue
- Ws: average waiting time in the system
- P_0 : probability of finding 0 people in the system
- Pn: probability of finding *n* people in the system

7. *Input-output graphs*

To analyse the process, you need to know the pace of inputs and outputs. To this effect, the number of units in the system is graphically identified. This will be the number of people in the system when we apply it to a service. When applied to the production process environment, it is called WIP (work *in process)*. The time a unit is inside the system can also be identified. This can be the average time in the system in the service environment, or *the lead time* when referring to industrial processes or the manufacture of physical goods.

In the figure below, inputs are shown as a solid line and outputs are shown as a dashed line. The horizontal axis reads the time when the units enter (continuous line) and leave (dashed line). Therefore, the horizontal distance between the continuous and dashed line measures the *lead time* of a specific "arrival". On the other hand, looking at a given moment and measuring the vertical distance between the two lines, the WIP is found for that given moment.

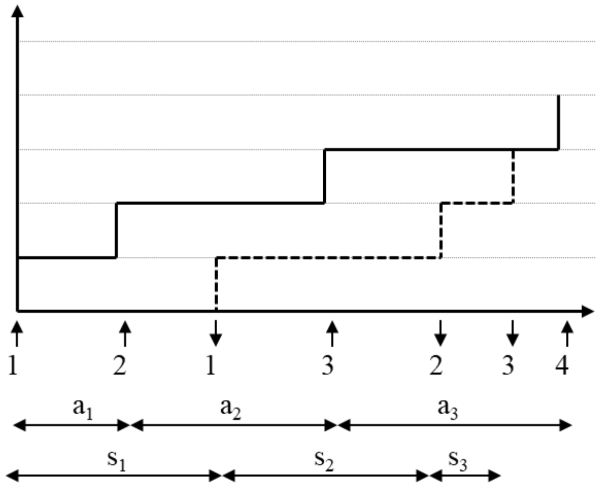


Figure. *Input-output* graph of a queuing phenomenon.

8. Operational analysis

There are four laws that are always fulfilled. The parameters that need to be known are: (i) the arrival rate (λ); (ii) the server rate (μ); and (iii) the number of channels. These laws enable other parameters of the system to be found based on these three.

8.1. Law of conservation

It could be argued that all units (arrivals) that enter the system must leave it sooner or later. Therefore, the rate of arrivals (λ) must be equal to that of the departures. The leaving rate is a weighted average. ρ % of the time (when the server is working), and the output is at the rate of μ. However, when the server is idle, which is (1 – ρ) % of the time, the output rate is zero. Therefore,

$$\lambda = \mu\rho + 0(1 - \rho)$$

From here it is concluded that:

- $\rho = \lambda / \mu$

Server utilisation is the relationship between demand and capacity. Obviously, the demand (λ) must always be less than the service capacity (μ), because otherwise the amount of units (arrivals) in the system would grow indefinitely.

In the case of m servers: $\rho = \lambda/m\mu$

8.2. Little's Law

In a previous chapter on process analysis, the following was defined:

WIP = throughput * lead time,

relating three parameters: the stock in the system, the processing speed and the process time. Its adaptation to the service notation according to Kendall is:

- $E[N_S] = \lambda E[W_S]$

8.3. Law of homogeneity

This can be formulated in two ways. The first refers to time and the second to the number of arrivals.

- $E[W_S] = E[W_Q] + E[S]$
- $E[N_S] = E[N_Q] + E[N_{served}]$

8.4. Law of Hopp and Spearman

This refers to the waiting time in the queue.

$$W_q = \left[\frac{p}{m}\right]\left[\frac{utilization^{\sqrt{2(m+1)}-1}}{1-utilization}\right]\left[\frac{cv^2{}_a + cv^2{}_s}{2}\right]$$

- Usage = λ/mμ
- m: number of channels
- CV: standard deviation/mean (coefficient of variation of a variable)
- CVa: coefficient of variation of the time between arrivals
- CVs: service variation coefficient

It can be seen to be composed of three parts. The first refers to the ability to serve. In this case, *p* is the process time. You can replace this *p* with an *s,* depending on whether you are referring to a process time or a service time. All formulas can be adapted to the manufacturing environment or to the service environment.

The second part refers to the relationship between demand and capacity to serve. The utilisation is λ/mμ.

The third part refers to the variability of the process. This degree of variability is given by both the variability of inputs and the ability to serve. In both cases, it is measured for the coefficient of variation of the two variables (A – time between arrivals – and S – serving time).

Last, a particular case that occurs very frequently in the real world must be mentioned. It is the case when these three conditions are met:

- The arrivals arrive with a time lapse according to a Poisson distribution.
- The server is able to serve according to a negative exponential distribution.
- There is only one channel.

When these three conditions are satisfied, and on applying the four equations (laws) mentioned above, it can be concluded that:

$$E[N_s] = \frac{\rho}{1-\rho}$$

This equation leads to some interesting conclusions. First, it can be observed that if ρ is zero (the server is completely idle), this means that there is no one in the system. However, if the employment level is 1, the system is unstable and tends to have an infinite number of people. It is necessary to see what the optimal point of employment in each case is. If the cost of waiting for the customer is high and the cost of serving is low, then you need to be at low occupancy levels (low ρ) to not make customers wait.

When it is the other way around, when the scarce resource is the server, then you need to keep them busy at all times, even if the client has to wait a long time.

9. If you queue

There are occasions when multiple queue models overlap. For example, in a car repair shop, the vehicle first enters a workstation, then is sent to a second station or a third station based on certain probabilities. To this effect, each car flows through different workstations according to specific requirements. These systems are called *open systems,* because each unit (car) enters the system (workshop) through a door or entrance, and after a while it leaves it. In open systems, each queue can be treated as an isolated queue.

Other times, a *closed queuing system* can be modelled*,* where the processed units never leave the system. The units go from one station to another, but there is no entrance or exit gate. It takes some imagination to model reality in this way, but there are situations where it is a useful system. For example, think of a workshop for mechanising a mechanical piece of some kind and imagine that there are *n* machines working continuously. From time to time, they break down and need to be repaired. It can be considered that there is a mechanic and that when a machine breaks down it asks for the service "to be repaired". The machine will then queue up in front of the mechanic. When it is repaired, it queues up again at a station that we can call "working machines". In this second state, there is no need to queue; it is as if its server capacity is infinite. Again, from this "working" state the machines (which are the units that flow in the system) request the "repair" service with a certain frequency, so each machine goes around in this closed system with just two stations, the "working machines" station and the repair shop station. Therefore, the machines never leave this system.

The analysis of closed systems is not as simple. A specific algorithm must be used, such as Buzen's algorithm. In this case described above, however, resolving it through simulation is simpler.

10. Resolution of cases

There are three ways to approach solving a queuing system:

- Apply the formulas directly to each model
- Apply the four operational laws
- Make simulation

The first two cases apply in simple systems. When they are more complicated, simulations are required. In this case, we will never be able to get exact results, but results close enough to the optimum to be able to make decisions.

There are several ways to do simulations. For example, Monte Carlo simulation can be done using a spreadsheet, while more specialised software is required for increasingly complicated simulations.

Chapter 10. Quality management

Objective: Explore the concept of quality management in the productive system.

Learning outcomes:

Definition and origin of quality management.

Quality assurance: quality and environmental management systems.

Total quality management.

1. The origin of quality management

Companies nowadays work in increasingly unstable environments where competitiveness and internationalisation are key factors for their survival. In this context, management models are an essential tool for improving day-to-day management and company decision-making, and for establishing a strategy to help companies move towards business excellence.

To break away from traditional management models such as Taylorist or Fordist, companies are implementing more flexible and versatile ones. Specifically, the main transformations they are undergoing are:

- Reduction, reorganisation and simplification of the departmentalisation and hierarchisation of the company and of its production process.
- The customer becomes the central figure around which the company's strategy pivots. We no longer speak only of the company's external customer, but also of its internal customer.
- The implementation of the quality paradigm as a model aimed at achieving the satisfaction of the fundamental figure of the company, the customer, through the cultural change that involves the implementation of a new model of participation, motivation and training of the company's personnel, among other models.

According to the International Organisation for Standardisation (ISO), quality is defined as "the set of characteristics of an entity that make it capable of satisfying established and implicit needs".

In this regard, the following question should be asked: whose needs are to be satisfied? The answer is clearly the needs of the user or customer. However, it should be emphasised that the term user or customer does not only refer to the final or external customer, i.e., the person, external to the company, who will purchase the product or service. In fact, the term user or customer also includes the internal customer, i.e., a company employee, who fulfils the roles of supplier and customer throughout the production process.

In short, quality management is a management philosophy focused on quality and based on the participation of all members of the company, which seeks to increase the satisfaction of stakeholders (including external and internal customers), the company's shareholders and society in general, while improving organisational efficiency and obtaining benefits for all members of the organisation and for society in general.

2. Quality assurance

Among the main ways in which a company can achieve its commitment to quality, first we find the use of a number of tools to improve quality. Second, and indicating a greater involvement of the whole organisation in quality, we find quality assurance.

Quality assurance emerged as a natural evolution of quality control, which was limited and ineffective in preventing the appearance of defects. It therefore became necessary to create quality systems that incorporated continuous improvement as a fundamental principle, and which were also useful to anticipate errors before they occurred.

A quality assurance system focuses on ensuring that what an organisation provides (be it a product or a service) meets the specifications previously established by the company and the customer.

Quality assurance is a system that places the emphasis on the whole process from design to delivery of the product or service to the customer, concentrating efforts on defining the processes and activities that enable to obtain products in accordance with the specifications.

This concept represents the need to involve the entire organisation in quality management, to demonstrate that it is capable of offering a product or service with the right characteristics, always controlling the production or service in accordance with the established requirements. Quality assurance could therefore be understood as the organisation, planning and control of all company activities and functions, aimed at achieving quality in accordance with certain requirements.

The implementation of a quality management system in any organisation is motivated by a number of external and internal factors that will to some degree or another influence the definition and application of the appropriate type of quality management system, which must have a number of basic requirements and a defined and organised structure.

Among the management systems for quality assurance, the most widely accepted system implemented by companies worldwide is based on the ISO 9001:2015 standard, which defines the "requirements for a quality management system".

When implementing any quality management system, and especially quality assurance systems, it is necessary to analyse the required resources and means to carry out the project, establish a general planning and define the structure within the company itself that will be responsible for supporting and collaborating in the different tasks involved in the project.

The basic stages that make up a quality assurance and quality management project begin with an analysis of the initial situation, followed by the development of the chosen system, the actual implementation, the subsequent evaluation through an internal audit, and last certification by a duly accredited external company.

The basic structure of a quality management system of this kind usually consists of three main members: the quality committee, the quality management and the improvement teams. It is up to the organisation itself to decide who will make up this committee and teams, and to whom the responsibilities for each of the derived functions will be assigned.

Once the structure of the system has been defined and the different responsibilities have been assigned, the implementation activities must be carried out, among which is the preparation of the system's documentation.

This documentation includes the set of documents generated as a consequence of any activity that affects quality, as well as the documentation that comes from outside and is relevant to the management system established in accordance with ISO 9001:2015. This documentation constitutes the basis of the system, so its drafting requires a huge effort on the part of all those responsible for the company and, in particular, those responsible for quality management.

Management system documentation generally consists of the following documents:

- General documents of the management system (planning, designation of the quality manager, etc.).
- Main documents of the management system (quality manual, quality procedures, quality instructions, formats and quality records).

This documentation will be prepared by the people designated by quality management who are involved in the activities to which each specific document refers, and will be initially approved and periodically updated in accordance with ISO 9001:2015.

Once the management system is in place, it should be subject to an internal review to assess whether it satisfies the requirements that apply to it, called a quality management system audit.

Audits, in general, can be of different types depending on the purpose (process, product or system audit) and scope (internal or external audit).

Internal audits are carried out by the company's own staff. Their objective is to identify the weak points of the system to strengthen them by means of a series of corrective or preventive actions that are implemented in cases where non-conformities have been detected.

Last, certification is achieved through the favourable evaluation of an external audit, which is nothing more than an external demonstration (customers, suppliers, society, etc.) of the company's commitment to quality.

Another type of management system that is currently widely implemented in companies is the environmental management system (EMS). These systems were created in the 1990s for companies to achieve a high level of environmental protection within the framework of sustainable development.

The implementation of an environmental management system can help the company in different aspects:

- Identify and control the environmental aspects, impacts and risks relevant to the company.

- Improve the company’s environmental policy and facilitate the achievement of the company's objectives by complying with environmental legislation.
- Define the basic principles guiding the organisation towards its future environmental responsibilities.
- Establish short-, medium- and long-term objectives for the company's environmental performance, analysing the cost-benefit balance for the organisation and its stakeholders.
- Determine what resources are necessary to successfully achieve the pre-established objectives, assigning responsibilities in each case.
- Define and document the different tasks and operations, responsibilities, authority and procedures to ensure that all employees act daily to minimise or eliminate negative possible negative impacts of the company on the environment.
- Improve the organisation's communication, training people to assume these responsibilities.
- Measuring the company’s environmental performance on a day-to-day basis to monitor whether the pre-established objectives are being met and to modify what is necessary when deemed necessary.

The environmental management system most widely used by companies today is the international standard ISO 1400:2015, which sets out the "requirements for an environmental management system". Certification under this standard is valid for three years, renewable if a new certification audit is passed.

The certification process for achieving ISO 14001:2015 certification is in parallel to obtaining ISO 9001:2015 certification, and in fact integrated audits of the two management systems can be carried out in unison.

3. Total Quality Management

Total Quality Management (TQM), which aims to achieve business excellence, is the term used to describe the assurance of quality management and customer satisfaction over time. TQM means that the culture of the organisation is defined and supports the constant achievement of customer satisfaction through an integrated system of tools, techniques and training. This system involves the continuous improvement of organisational processes, resulting in high quality products and services.

Total quality management is based on eight principles, defined in ISO 9000:2015 as "fundamentals and vocabulary for quality management systems", referring to ISO 9004:2018 for "enhancing an organisation's ability to achieve sustained success".

1. Customer focus

The first and foremost principle of TQM is to focus on customers who buy products or services, and on potential customers. Customers are the ones who justify the quality of products and services. Therefore, the company must ensure that customers feel that they have spent their money on a quality product.

Furthermore, the organisation must be clear that the needs of its customers are not static but dynamic, and they therefore change over time. Customers are not only becoming more demanding, but they are also increasingly well informed. The organisation must not only strive to understand the needs and expectations of its customers, but it must also offer them different solutions through products and services, and manage and try to exceed expectations on a daily basis.

2. Leadership

Leadership is essential to maintain unity among employees to achieve interdependent goals. Although there are mainly three types of leadership in the industry, the democratic leadership style is the best for good results. These leaders can create a suitable environment to work effectively within the organisation, wherein all employees work to achieve the organisation's goal.

Leadership is a chain that affects all managers in a company who have staff under them. If one link in this chain is broken, the leadership of the company is broken.

3. Involvement of people

Employees are the lifeblood of the company, and their total commitment enables their skills to be used for the benefit of the company. Employees’ total commitment enables the company to develop new products and increase sales growth. Therefore, all employees in the organisation must be well trained, committed and dedicated to achieving a common goal. Therefore, it is the company’s obligation to create a responsive environment where every employee is motivated to complete their task in the right way.

4. Process approach

The company needs to constantly improve processes to achieve good results. A desired result is achieved more efficiently when activities and related resources are managed as a process. The main change required is to the conception of the company. The company must stop being a company organised into departments or functional areas and become a company organised by processes with which to create value for customers.

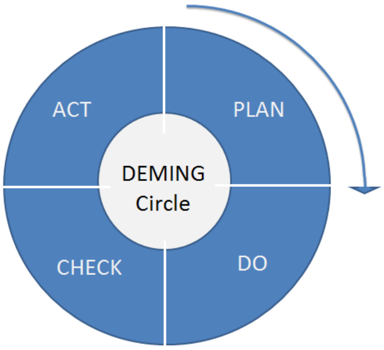
5. System approach to management

Identifying, understanding and managing interrelated processes as a system contributes to the efficiency and effectiveness of a company in achieving its objectives. To do so, it will be necessary for the company to correctly identify and manage all interrelated processes.

6. Continual improvement

One of the most important principles is that of continual improvement, a concept that can be defined as a structured process in which everyone in the company takes part with the aim of progressively increasing quality, competitiveness and productivity in a changing environment. This process is structured according to the continuous improvement cycle, also known as Deming's PDCA circle.

The initials PDCA stand for the different stages of the cycle, corresponding to plan, do, check and act. There is no doubt that the application of the principle of continuous improvement has a positive impact both on the reduction of company costs and improves productivity and profitability.



7. Fact-based decision-making

Decisions should be based on the analysis of the data and information collected. The company must analyse data from different sources to assess the achievement of defined objectives, and to identify areas for improvement, including possible benefits for stakeholders.

8. Mutual beneficial supplier relationship

An organisation and its suppliers are interdependent. A mutually beneficial relationship enhances the ability of both, thereby creating value. It is therefore important to establish relationships with suppliers to promote and facilitate communication, with the objective of mutually improving the effectiveness and efficiency of value-creating processes.

Chapter 11: Project Management

Objective: Understand the main techniques for project management, including the Gantt chart, PERT, CPM and the “critical chain” approach.

Learning outcomes:

Identify what a project is and its characteristics.

Gantt chart.

Deterministic and probabilistic PERT diagram.

Critical Path Method (CPM).

Management of *buffers* according to the principles of "critical chain".

1. What is a project?

A project is defined as a sequence of tasks that must be completed to achieve a goal. According to the Project Management Institute (PMI), the term *project* refers to any temporary endeavour with a defined beginning and end.

Each project is composed of activities, which may have a sequencing relationship among them. Each activity requires a certain amount of time and consumes resources. Resources can be of different kinds, but they can always be translated into a common measure: a monetary unit.

Another characteristic of each project is that it is unique. No two projects are ever the same. A project is always something run for the first time. Since there has been no previous experience of the project, no one knows how to conduct it, although there are obviously close or similar experiences that help give us an idea of how the project will unfold. There is always uncertainty about the time it will take to perform it (How long will it take to complete the project? How long will it last?) and about the cost (How much will it cost to complete the project? How many resources will need to be invested in it?)

To manage these two uncertainties, it is necessary to assign (i) a probability function on the duration time of each activity that makes up the project, and (ii) a budget for each activity.

Project quality can be measured in three dimensions:

- The fit between the specifications the client asks for before starting and the characteristics of the project delivered.
- The difference between the expected cost and the actual cost.
- The difference between the expected time (the agreed term) and the actual time used.

In a project bidding process, suppliers compete based on (i) a price (including internal costs and profit) and (ii) a delivery time. This is something that should be determined based on internal analysis of the project and knowledge of the industry and competitors. Of course, the lower the price and the shorter the term, the better in terms of competitiveness. However, the risk associated with these decisions must be analysed.

Project management includes three phases:

- Planning: setting the objective, defining the characteristics that the product or service to be delivered must have and the team that will execute it.
- Scheduling: determining when each activity will start, and which people and resources will be needed at each moment.
- Control: monitoring the use of resources, costs, quality and budgets. Here you should also review any changes required to keep to the delivery date and total cost established.

Focusing on the programming phase, Gantt charts are very useful for simple projects. The name derives from Henry Gantt, who popularised this chart at the end of the 19th century.

The program evaluation and review technique (PERT) and the critical path method (CPM) are two widely used techniques when the project is more complex. The ideas derived from the "critical chain" help with the use of time buffers to optimise programming.

PERT and CPM were developed in the 1950s to help project managers schedule, monitor and control large, complex projects. The CPM came first as a tool developed to help in the construction and maintenance of chemical plants of the DuPont company. For its part, the United States Navy published the PERT in 1958, which was used to manage the "Polaris" project, relating to the manufacture of high-performance submarines. Critical chain was published in 1997 by Eliyahu M. Goldratt, since when no further relevant contributions in project management have been published.

2. Gantt diagram

This is a type of bar chart that shows the schedule of a project, with the list of tasks or activities to be performed in the vertical axis, and the time in the horizontal axis . You can read when each activity starts and ends, and consequently its duration. It also shows dependence between activities; one activity commonly cannot begin until another has been completed. The figure below shows a project with only three activities, where activities A and B must be completed before C can begin.

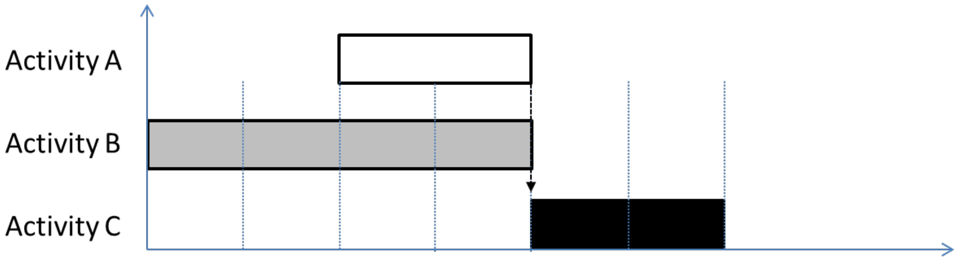


Figure. Gantt chart with activities on the vertical axis.

There is another type of Gantt chart, which shows resources on the vertical axis rather than activities. It is useful to check that there is no overlap in tasks assigned to the same resource.

The following figure shows the same project, but now with regard to resources. In this case, resource X is responsible for activities A and C, while resource Y is responsible for activity B.

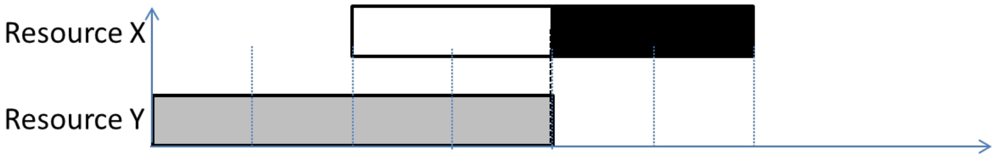


Figure. Gantt chart showing the required resources on the vertical axis.

3. Deterministic PERT diagram

PERT is the acronym for *program evaluation and review technique.* In this diagram, and to facilitate programming, it is assumed that each activity has a specific duration. It must be kept in mind, however, that this circumstance is entirely contrary to the framework of project management since, as mentioned above, the very nature of a project involves the uncertainty of the duration of each activity.

Steps to perform a PERT (deterministic)

A. List the activities. An activity takes time and consumes resources. The list of activities is required to complete the entire project.

B. Estimation of how long each activity lasts. In this simplification, it must be assumed that there is no uncertainty and that the exact duration is known.

C. For each activity, decide which activities should precede it and which should follow it. Care must be taken not to introduce more restrictions than those required by the very nature of the project, since they have a negative impact on the total duration of the project.

D. Draw the network composed of nodes and arrows. A *node* is an event, for example the completion of activity X. It is a moment in time. For their part, the arrows mean an activity, so the time needed to complete it must be allocated, along with the necessary resources. Each node is labelled with a number indicating the sequence in which it is reached. When the network is large, there is an algorithm to perform this numbering of nodes which guarantees that the numbers give the sequence in which each node is reached.

The figure below shows an example network or diagram. Each node shows that a certain state of the project has been reached, for example, that activities B and D have finished. It is therefore an instant in time. Above each arrow there is a letter indicating a specific activity. The relationships of temporal precedence between activities are also clear. It is noted that until B and D are completed, F cannot be started.

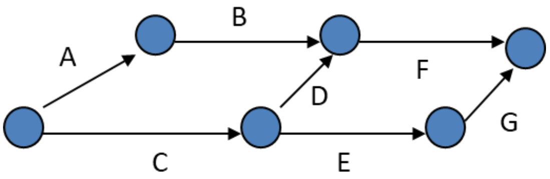


Figure. PERT diagram of a project.

A diagram of this kind has various characteristics:

- There is only one start node and one end node.
- Every activity has at least one predecessor and one successor, with the exception of activities that start from the initial node and those that end at the final node.
- Every activity starts at a node *i* and points to a higher-order node *j,* so that j > i.
- There cannot be two activities with the same start node and the same end node.

Dummy activities, which do not consume any resources and whose duration is zero, can be introduced to meet these diagram characteristics. Dummy activities also helps avoiding to introduce restrictions that are not strictly necessary, and therefore the diagram faithfully will describe the nature of the project.

E. Calculate t_i (earliest time a node can be reached) and T_i (latest time) for each node.

- First, the "earliest" times to reach each node are calculated. The method goes from left to right.
- The later moments for each node are then calculated: here, the method goes from right to left.

In this phase, a special notation is used for each node, which includes: (i) the number of the node, which is like its name; (ii) the earliest time the node can be reached; and (iii) the latest time the node can be reached.

F. Search for “critical activities” and “critical path”.

- Activities are identified by their start and end nodes.
- The margin of an activity is the amount of time by which it can be extended beyond the estimated or expected time, without delaying the entire project.
- A critical activity has no margin or slack.
- The activity margin is calculated as: $\text{Slack}_{i,j} = T_j - t_i - \text{duration}_{ij}$
- The critical path is a continuous path through the project network that (i) starts at the start node of the project; (ii) ends at the last node of the project; and (iii) includes only critical activities (i.e. , activities without margin).
- The critical path is important because it establishes the duration of the project.

G. Gantt associated with the PERT chart. It is useful to draw the Gantt chart, which provides an intuitive view of the activity sequence.

The following figure shows the same example as above, for which the critical activities have already been found to be C, E and G. Therefore, the critical path is formed by these three activities.

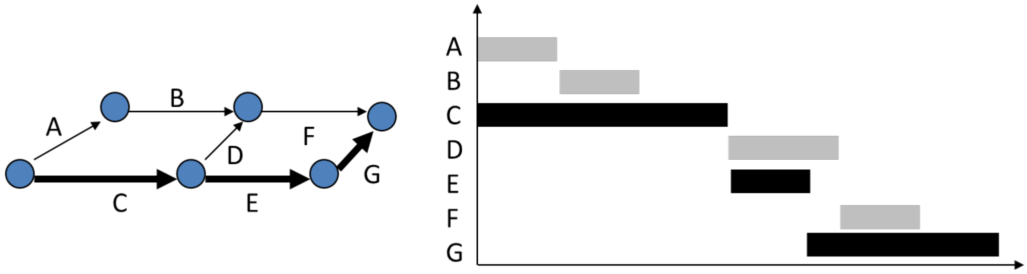


Figure. Network and Gantt diagram of a project, showing the critical path.

H. Resource load profile over time, showing the resource usage. The figure underneath the Gantt diagram below is the load diagram, vertically showing the resources required in the execution of the project, which are often expressed in monetary units.

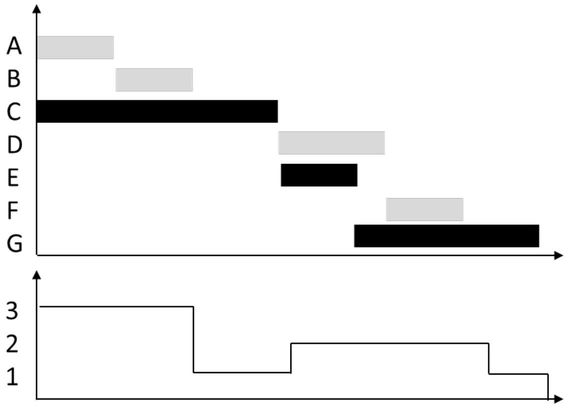


Figure. Gantt chart and resource loading chart.

I. Last, a discussion about the strategy is required "as soon as possible" (ASAP), or a strategy for performing things "as late as possible" (ALAP).

The ASAP strategy reduces the risk of delivering after the deadline, but requires more financial costs. ALAP is the exact opposite, where the risk of late delivery increases, possibly resulting in a fine or a penalty. The advantage is that there are fewer financial costs.

4. Probabilistic PERT diagram

This diagram is used when the constraint of considering each activity as having a predetermined fixed duration is removed. In this regard, it is necessary to consider that the duration of each activity follows a certain probability distribution. To this effect, the uncertainty of the duration is handled by a distribution function.

The duration of an activity is generally distributed according to a beta (β) distribution. Three parameters are required to make an estimate of the mean (expected value) and variance.

- a: Optimist. Time an activity will take if everything goes according to plan. In estimating this value there should only be a small probability (say, once in every 100 times) of the activity time being less than *a.*
- m: Most likely (highest likelihood). More realistic estimate of the time required to complete an activity.
- b: Pessimistic. Only once in every 100 activities will the time be greater than this value.

The expected time and variance for the beta distribution are:

- Average or time expected to complete the activity = (a + 4 m + b) ⁄ 6
- Activity variance = [(b − a) ⁄ 6]^2

Keep in mind that this last formula is for the variance. If the standard deviation is required, it should be calculated as the square root of the variance.

The figure below shows the β distribution. It can be skewed to either the left or the right.

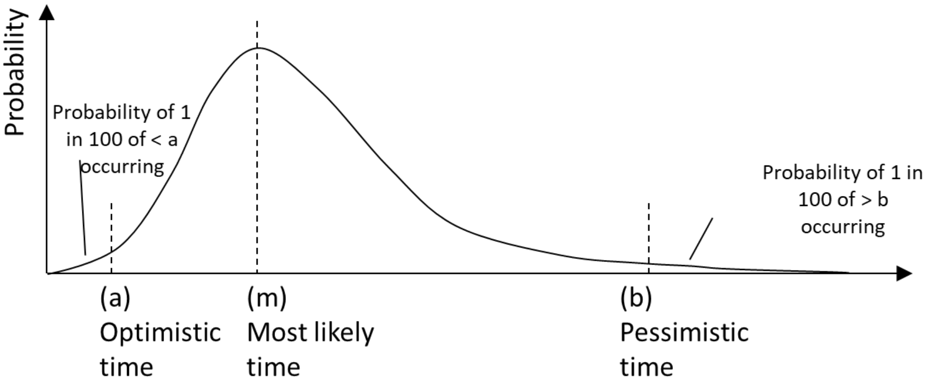


Figure. Density function β.

Once the distribution function of the duration of each activity has been established, the same steps can be followed to schedule the project as for the deterministic PERT diagram. The duration of each activity now follows a beta distribution. You need to take the average for each activity and find the critical path.

The project duration (i.e., the length of the critical path) follows the normal distribution pattern. Given the central limit theorem, the expected duration of the entire project (of the entire critical path) is the sum of all the average durations of the critical activities that make up the critical path. The variance is also the sum of the variances of these critical activities.

Once the expected duration of the project and its standard deviation are known, you can calculate the probability of completing the entire project in less than a certain period, or the probability of completing it between a certain moment and another later time. PERT is a good technique for scheduling a project and evaluating its duration.

5. Critical Path Method (CPM)

CPM, short for critical path method, is a technique wherein each activity has a normal or standard time that is used in the calculations. Associated with this normal time is the normal cost of the activity. Another element is the minimum or record time, which is defined as the shortest duration required to complete an activity. Associated with this record time is the record cost of the activity. An activity can typically be shortened by adding additional resources.

It is an iterative process that aims to reduce the total project duration: if the bonus received for reducing one unit of time is greater than the cost associated with this reduction, then it is worth investing in this reduction.

Here are the steps to follow:

A. Calculate the slope cost or crash cost for each activity.

- Activity slope cost = (Record cost – Normal cost) / (Normal time – Record time).
- This is the marginal cost of shortening the activity by one unit of time.
- The slope between the “normal” point and the “crash” point is assumed to be linear (in a diagram where cost is on the vertical axis and time is on the horizontal).
- There may be different cost reduction time diagrams, depending on the nature of the activity. The figure below shows two cases: in (a) there is a linear variation, while (b) shows a situation where the marginal increases in cost are growing for each unit of reduction in duration.

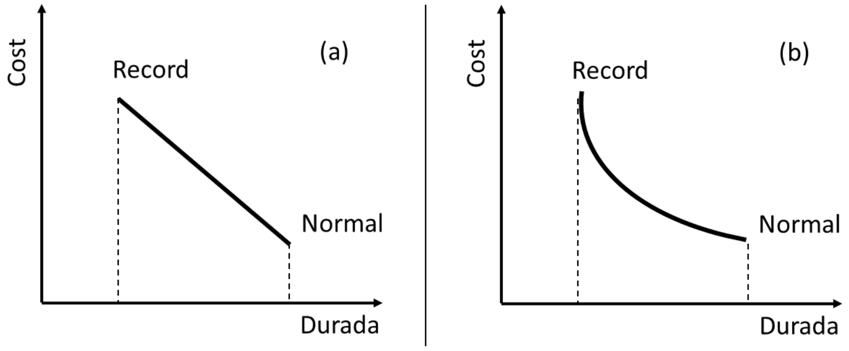


Figure. Examples of activity cost and duration diagrams.

B. Find the project's critical path or paths, using normal activity times. Identify critical activities.

It is only worth investing (reducing) in critical activities as the single way to reduce the time required for the whole project.

This is the first scenario, the starting point of the CPM method.

C. Iterative algorithm

1. Generate alternatives to shorten the project.

2. Choose the cheapest one. As long as an alternative is found where the cost of shortening the project by one day is less than the benefit of shortening the project by one day, it is worth investing in it; otherwise, you need to maintain the existing scenario and finish the CPM process.

3. Exploit the selected alternative until

(i) the chosen activity is in its "record or minimum duration" or

(ii) another critical path appears

4. Calculate the duration of the project and its cost and return to the first step.

Following this algorithm, the cost-time diagram for the entire project will be convex. This is because the first alternatives chosen to shorten the project are the cheapest. When these options are not available in later steps, more expensive alternatives are used, leading to higher marginal costs.

Thus, CPM is a good technique to evaluate the use of resources and their optimisation.

6. Critical chain

This contribution was made by Dr Eliyahu Goldratt in his book *The Goal,* and consists in a way to manage *buffers* (time).

The previous techniques do not manage *buffers* as a way of avoiding deviations in the deadline. PERT/CPM assume that the duration of each activity is really the estimate of the person responsible for that activity. However, following Goldratt, the estimated duration is comprised of the estimate of the actual duration to perform the activity and the safety margin added by each person in charge of the activity, which could be 50%. These security times are added to each activity by each person in charge to avoid a deviation in the performance of this activity. This implies a longer duration for the entire project, as partial margins are inserted throughout the project.

This approach (critical chain) removes partial safety margins and manages them together, with a holistic view of the entire project. The goal is now to optimise the entire project. It uses a shared *buffer,* rather than a set of private *buffers.*

The figure below shows the comparison of a simple project (composed of just three activities) managed first using a classical approach and then using the critical chain method. The critical chain method shortens the project deadline by one week and still maintains a two-week project *buffer,* which reduces the risk of late delivery.

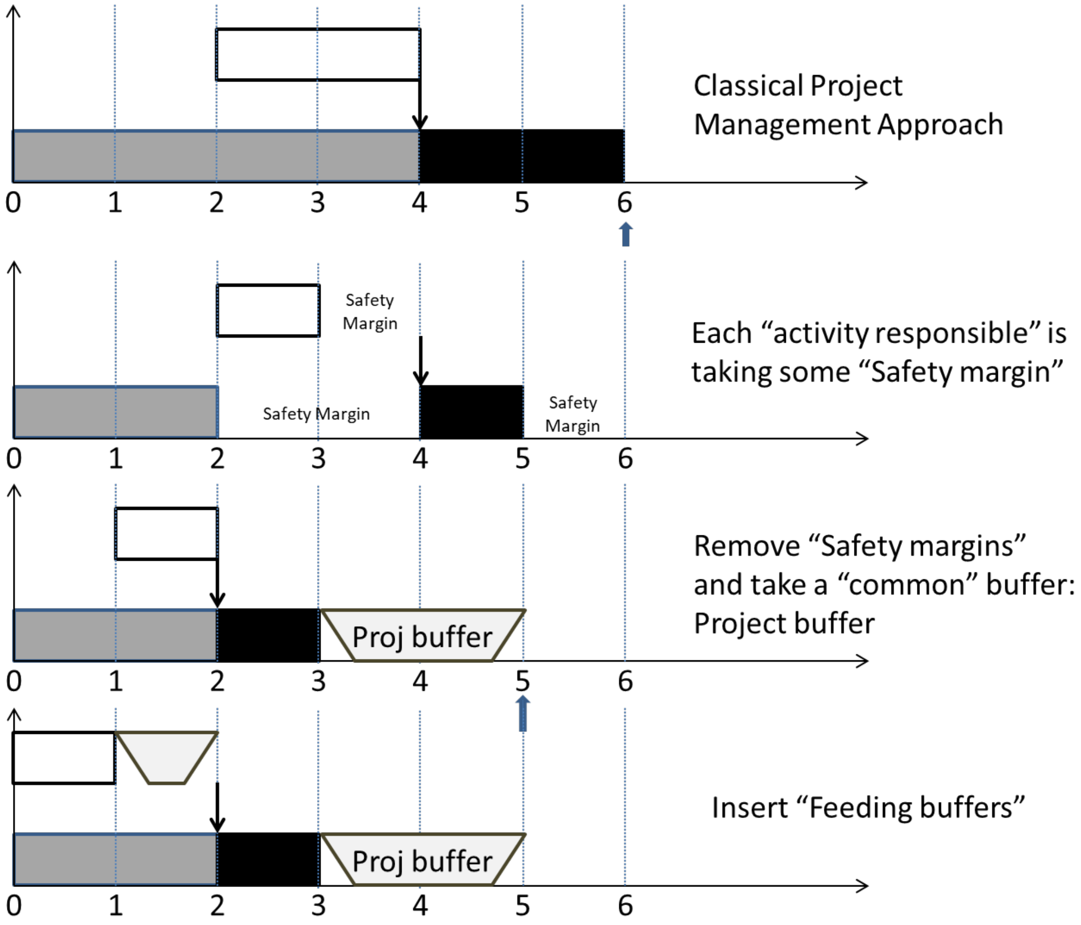

Classical Project Management Approach
0
1
2
3
4
5
6
Safety Margin
Safety Margin
Safety Margin
Each "activity responsible" is taking some "Safety margin"
Proj buffer
Remove "Safety margins" and take a "common" buffer: Project buffer
Proj buffer
Insert "Feeding buffers"

Figure. Comparison of a simple project managed first with a classical approach and then using the critical chain method.

The critical path in the classical management approach. The second line is obtained by removing 50% of the margin of safety in each activity. By rearranging the critical path (third graph), a project *buffer* can be inserted (*project buffer)* to protect the delivery date and be more "competitive" than with the classic management approach. When a non-critical activity is feeding the critical path, a feeding buffer can also be inserted *to* protect the regular progress of the critical path (last graph).

Buffers and their functions are:

Buffer type	**What does the buffer protect?**	**Where is the *buffer inserted* ?**
Project buffer	• Protect the project deadline delivery • Balances critical chain fluctuations	At the end of the project
Feeding buffer	• It protects the critical chain from negative fluctuations in the activities that feed the critical chain • It allows to take advantage of the positive fluctuations of the critical chain	Where a non-critical activity feeds into the critical chain

Resource *buffers*, which are *pseudo-buffers*, can also be introduced: they are not slot-times, but simply a reminder that a resource will soon intervene in a critical activity.

The main steps to apply critical chain principles in project planning are:

1. Find the critical path using the ALAP strategy and the durations without safety margins for each activity.
2. Resource levelling.
3. Determination of the critical chain.
   - This consists of critical activities.
   - A critical activity cannot be pushed to the left without also moving the starting point of the entire project to the left.
   - It should be kept in mind that it is called a critical chain rather than a critical path, a term that emphasises that chain is a more demanding term than path. The chain considers not only the sequence order of activities, but also the availability of resources.
4. *Buffer* points: where to insert *buffers.*
5. Decide the size of the *buffers.*
   - While other steps in this method are fairly automatic, deciding on the size of the *buffers* is strategic. It depends on the competitiveness of the sector. There are no rules, although some academic papers provide some, based on particular assumptions.
6. Insert the *buffers.*

Universitat
de Girona