Chapter 4
Manufacturing Systems Analysis

Jack C Chaplin and Giovanna Martinez-Arellano

4.1 Introduction

Analysis (noun): Detailed examination of the elements or structure of something.

- Oxford University Press

Effective decision-making is critical in manufacturing enterprises, and selecting the correct course of action can be the difference between a successful and competitive enterprise, and falling behind the competition. Decision-making is the process of the selection of a course of action to best achieve your goals, given available options and the available information. Decisions are rarely simple, and the decision maker must select the best choice to maximise one or more criteria, and must often do so with incomplete information.

Within the manufacturing domain, effective decision-making is crucial to remaining competitive. For example, making large investment decisions carries significant risk and significant reward, and this is especially true for smaller companies. Effective decision making is best achieved with access to accurate and high-quality data which is then converted to and presented in more usable forms. Data can be converted into more usable forms to achieve maximum value and to inform the manufacturing enterprise. This is often described as the difference between data and information, with the former being the raw numbers or measurements, and the latter being actionable insight into the manufacturing process.

J.C. Chaplin (✉), G. Martinez-Arellano
Institute for Advanced Manufacturing, University of Nottingham, Nottingham, UK
e-mail: jack.chaplin@nottingham.ac.uk, e-mail: giovanna.martinez@nottingham.ac.uk

J.C. Chaplin et al (eds), *Digital Manufacturing for SMEs*

A common representation of this is the Data, Information, Knowledge, Wisdom (DIKW) pyramid, shown in Figure 4.1-1. Each step in the pyramid can be analysed and compressed to the next stage up, slowly reducing the raw volume of information, and increasing the insight and value. Within this framework, the steps are defined as:

- *Data:* Signals or numbers representing physical phenomena from sensors, but without context or metadata.
- *Information:* Contextualised and inferred from data, information has meaning and context, and can be used to answer questions.
- *Knowledge:* Processed information that has been compared with previous experiences to enable the user to determine why things have happened.
- *Wisdom:* Sometimes omitted from this model or combined with knowledge, wisdom is using knowledge to intuitively understand a process and to determine the best future actions.

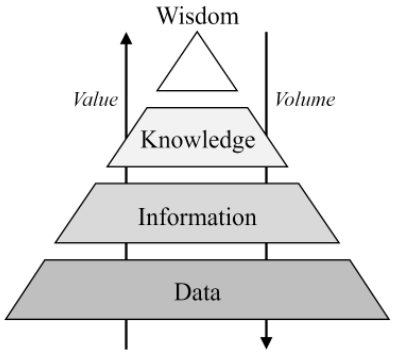

Figure 4.1-1 The DIKW Pyramid shows how data, information, knowledge, and wisdom are interrelated and how they compare in terms of storage volume required, and inherent value.

There are many sources of data in manufacturing enterprises, but one of the most important is data generated by the manufacturing lines and systems themselves. Manufacturing systems analysis, modelling and simulation deals with methods to better understand manufacturing processes and manufacturing lines. By understanding the current situation, and by testing potential changes, it is possible to ensure that the chances of success are maximised before making potentially costly changes (in terms of equipment or line downtime) to the manufacturing enterprise. What's more, these methods will enable tracking the performance of production systems over time, allowing the tracking of improvements as they are implemented or to spot issues before they become critical. The techniques here can be applied from the smallest manufacturing lines (which could just be a machining centre with a worker loading and unloading parts) to larger, more complex manufacturing lines.

However, simulation and modelling can often be overkill for many manufacturing systems. Standard mathematical evaluation can reveal insights into manufacturing systems that are not immediately obvious. Understanding utilisation, reliability, throughputs and capacities is an essential first step to maximising productivity by identifying areas for improvement. These areas will enable targeting interventions to the areas that will give the greatest return.

This book chapter discusses the pre-requisites for manufacturing analysis; understanding the type of system to be evaluated and understanding the question that needs to be answered. It then moves on to discuss two key methods of offline mathematical manufacturing analysis – conventional production analysis and queueing theory.

4.1.1 Manufacturing Systems Morphologies

Morphology (noun): The study of the forms of things.

- Oxford University Press

To understand how to analyse a manufacturing system, it is first important to understand the form and method of operation of the system, as this will influence the methods and formulae used. Knowing the terminology for your manufacturing system also simplifies searching for applicable resources and advice. A diagrammatic summary of these manufacturing systems morphologies can be found in Figures 4.1-2 to 4.1-5.

Dedicated Manufacturing Systems (DMS) use fixed automation to produce core products at high-volume with maximum cost effectiveness. When a single product is likely to be manufactured in large quantities without major alterations for the foreseeable future, a DMS is almost always the best choice, while simultaneously often being simpler to implement. Equipment is typically arranged in a linear manner – the stereotypical *production line* – and linked with a material handling system to move the parts along.

Where volumes are lower and multiple parts are to be manufactured, *batch manufacturing* is one of the most common manufacturing strategies, enabling mid-volume manufacturing by batching production together. The time required to change a production line between product types is significant, so batching ensures this changeover time happens as infrequently as possible.

Group technology is the strategy of gaining efficiency when many similar products are to be manufactured by grouping products into part families with similar features and manufacturing processes. For example, a company making bearings may make the same bearing in multiple different sizes, with multiple finishes and multiple different lubrication methods. Though this would constitute a very large number of potential part variants, the commonality between them allows them to be produced with the same machinery (albeit with different settings and tooling) and therefore comprise a part family. By comparison, though similar in function, a ball

bearing and a cylindrical roller bearing require sufficiently different processes that they would not constitute a single part family.

Utilisation of group technology enables a production line to be set up and configured to produce any members of the part family quickly. This makes changing between products in the same family much quicker and simpler, reducing batch change over times and making smaller batch sizes more cost effective.

Cellular manufacturing groups machines into cells, where each cell is specialised in completing a step or closely related group of steps required to manufacture a product or family of products. Manufacturing time is reduced by bringing the machines into close proximity; the characteristic U shape as shown in Figure 4.1-3 is a common cell layout and enables a single operator who specialises in the cell's task to oversee all machines, and the order in which machines are used in the cell can be changed without significant disruption. Cellular manufacturing is a great choice for lower volumes of highly variable products. Multiples of the same machine in a cell might be grouped together into co-located stations to take advantage of parallelism, and machines which are almost always used together might also be grouped into stations.

Flexible Manufacturing Systems (FMSs) are highly automated manufacturing cells, able to automatically route parts between the constituent machines, enabling flexibility in terms of the parts and part families produced. Shown in Figure 4.1-4, one of the defining differences from a cellular system is the use of an automated material handling system to route products around the system. The stations in an FMS are themselves general and flexible, predominantly computer numerical control (CNC) machines, often with automatic tool changers. This allows for the stations to be utilised for a wider range of tasks. FMSs are ideal for lower volumes of highly variable products similar to cellular manufacturing, but the high levels of automation increases labour productivity and allows for unattended production.

Reconfigurable Manufacturing Systems (RMSs) utilise both a structural design and a digital control system that allows the constituent stations to easily be changed, so the function and capacity of the system can be rapidly altered. Stations can be plugged in and out of the system to change the cell's function. Additional functional modules can also be added or removed from machine and/or stations to change their functionality or increase capacity. The system can scale up or down capacity by adding or removing modules, machines, or stations. Individual stations are often less flexible than would be used in an FMS – system flexibility is given through system structure, not individual flexibility.

Traditional mathematical analysis techniques generally focus on dedicated manufacturing lines and cellular manufacturing morphologies, with variants for when batch manufacturing is implemented on either. FMS and RMS are comparatively recent developments, and their evolving and changing nature makes them less suited to these styles of analysis. Using digital tools and particularly automatically updating digital twins may be more appropriate here to help manage the changes, and digital twins will be discussed in Chapter 6.

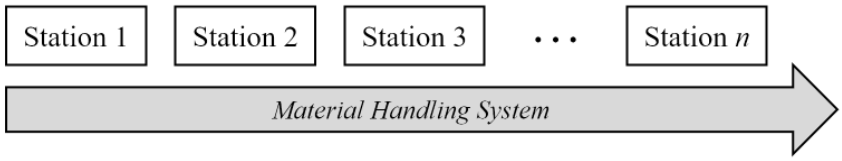

Figure 4.1-2 An example dedicated manufacturing system, showing the typical production line morphology. A single linear material handling system transfers parts or products between processes in sequence. Processes could be single pieces of equipment, manual process stations, or multiple similar pieces of equipment to form a processing station with parallel processing capability.

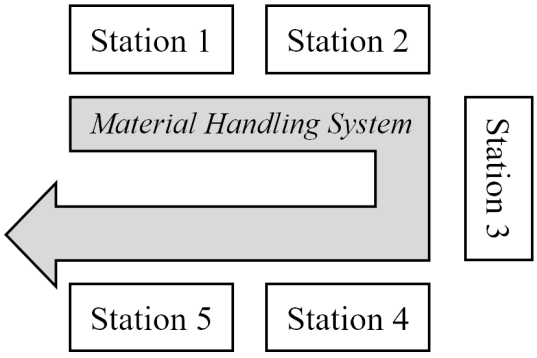

Figure 4.1-3 An example manufacturing cell layout. The arrow represents the most common flow of parts through the cell, but cellular manufacturing can alter the path to make a wider variety of products without significant disruption.

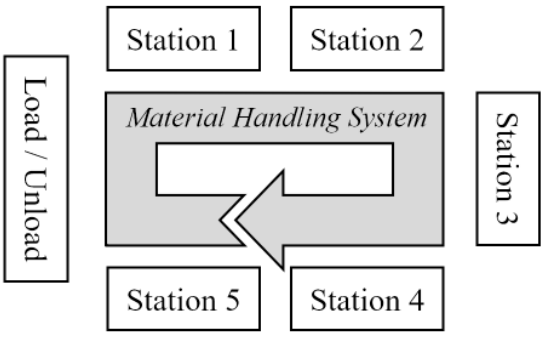

Figure 4.1-4 Example of a Flexible Manufacturing System layout. Parts can be routed between any stations in any order using the Material Handling System.

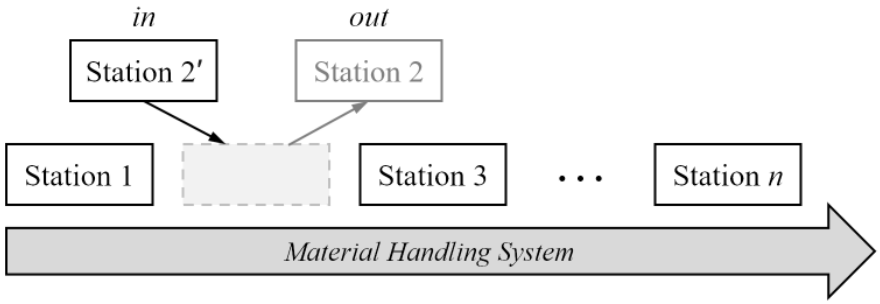

Figure 4.1-5 Reconfigurable Manufacturing Systems allow for rapid structural changes to the cell to change its functionality.

4.1.2 Decision Making

Before beginning analysis of a manufacturing system, it is important to understand what question you are actually trying to answer. Analysis usually serves to inform decision making by measuring key performance indicators (KPIs). By discovering values for KPIs you can better understand the strengths and weaknesses of the current system, and make decisions on how to improve it.

Decisions in manufacturing can mean a wide variety of different choices, but Hayes and Wheelwright [1] separate this into ten possible categories of decisions. These are detailed below together with non-exhaustive lists of what sort of choices these categories include:

- *Capacity:* How much capacity flexibility should be made available, what shift patterns should be utilised, and what strategies are available for subcontracting for temporary over or under-capacity situations.
- *Facilities:* The size, maximum capacity, physical location, and primary task assignment of physical manufacturing facilities.
- *Human Resources:* The policies around recruitment of new employees, the training and development of existing employees, and the culture and management style that the business adopts.
- *New Product Introduction:* How new products are selected and developed, as well as procedures for design (including design for manufacture), and how products are introduced and ramped up on the shop floor.
- *Organisation:* The structure of the manufacturing enterprise, as well as accountabilities, roles, and responsibilities.
- *Performance Measurement:* How the processes and people in a manufacturing enterprise are evaluated and monitored for productivity and other performance measures, as well as any recognition and reward schemes for employees.
- *Production Equipment:* The equipment and technologies chosen for manufacturing products, the physical layout of this equipment into cells or lines,

and the level of automation within these lines. This also covers maintenance approaches and policies, and how much in-house development of new or updated processes is possible.

- *Production Planning and Control:* How production is controlled (either via automated systems or by manual processes), how orders should be assigned and scheduled, and how materials are stored and moved around the manufacturing operation.
- *Quality:* The quality goals adopted by the business, as well as the quality assurance and quality control methods and policies used to reach these goals.
- *Vertical Integration:* High-level strategic decisions such as make versus buy, policies on supplier selection and continued relationships, reliance on single or multiple suppliers to spread risk.

To make effective decisions, it is important to understand exactly what the decision-making process is. The decision-making process [2] is a series of phases, and follows the flow shown in Figure 4.1-6:

1. *Intelligence – Problem Discovery:* First, it is necessary to identify a problem and to determine who the decision-makers and stakeholders in the decision process are. Once the problem has been recognized, the problem should be defined more formally, determining the requirements from the stakeholders to obtain a list of considerations and goals. It is important to take time with this phase, as the group of stakeholders may all think they share a common view of the problem, but there are often details that are assumed and may not be shared between all parties.
2. *Design – Solution Discovery:* Finding possible alternatives that could be implemented, and assessing their possible contributions to the problem. This could be as simple as a discussion amongst the stakeholders to brainstorm ideas, but for critical decision making it is recommended that a more rigorous process is followed. Though commonly overlooked, it is recommended that "doing nothing" is always one of the possible choices. A common recommendation is for the problem to be modelled, as this will enable potential solutions to be tested to enable the selection process in the next phase. This is the phase where the model is created. Models could be numerical models developed in a spreadsheet, full simulations of productions lines, or anything in between. The possible approaches for modelling here depend on the nature of the decision being made.
3. *Choice – Solution Selection:* The possible options developed are then evaluated for their contribution to the problem. Depending on the approach taken in the Design phase, this could either be a continuation of the discussion, gathering data to inform the decision, or running the possibilities on the model to see what impact they have on the defined goals. Even with a fully developed model, the choice of the solution is rarely simple. Many different criteria need to be evaluated, including the time and cost required to implement the solution, and risk and reward of doing so, the possible disruption while the solution is implemented, and the availability of skills to enact it. Remember to evaluate

doing nothing, as the disruption and cost of implementing a solution may be too high.

4. *Implementation – Solution Deployment and Testing:* Lastly, the solution should be deployed and tested. The solution needs to meet the goals defined in the first phase of this process, and as many choices are not instantly implemented or their effects instantly felt it is important to keep monitoring and testing. If the solution is failing to meet expectations or goals, the entire decision making process is an iterative one. It is always possible to return, rethink, and re-plan. It is very common for companies to fall into the trap of continuing with a bad decision when all evidence suggests that things will not improve (the "sunk-cost" fallacy). However, a properly executed decision making process will reduce the chances of this occurring.

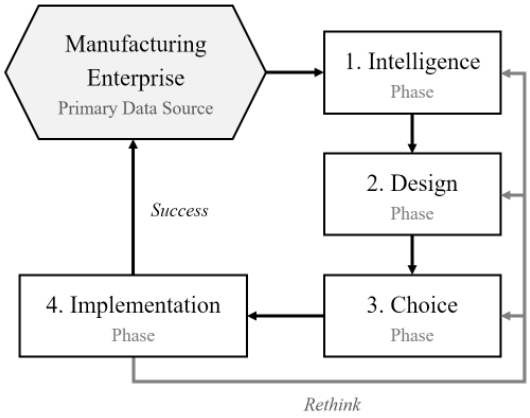

Figure 4.1-6 Phases of the decision-making process. It is never too late to go back and rethink a decision, especially as new information becomes available.

A key aspect of this process is the ability to monitor and check how successful a decision has been, and whether the changes to the manufacturing or business processes are yielding the expected and intended results. Key Performance Indicators (KPIs) are the tool with which the success of a decision can be measured.

4.1.3 Key Performance Indicators (KPIs)

To achieve an optimal manufacturing process, you must first define what optimal is. No manufacturing process can simultaneously maximise yield, productivity, uptime etc. while also minimising material wastage, energy usage, and downtime. Careful selection of KPIs is required to understand what is important and what you are trying to achieve. The indicators through which the performance of a manufacturing process is characterised have evolved and diversified. Traditional KPIs include:

- *Productivity:* The efficiency of production, or the ratio of output to input. Productivity is a KPI you want to maximise, and there are many ways to calculate it, though Overall Equipment Effectiveness (OEE) is a common and useful tool for manufacturing equipment.

$$OEE = AUYr_{os} \tag{4.1.1}$$

  Where:
  - *A: Availability.* The uptime of the equipment, which is reduced by maintenance, breakdowns etc.
  - *U: Utilisation.* What percentage of possible usable time is actually being utilised. Poor scheduling or a lack of available upstream parts (also called starvation) will reduce this.
  - *Y: Yield.* The percentage first-pass yield of the process, referring to the quality of produced components.
  - r_{os}: Operating Capability. The percentage of the maximum throughput the equipment works at.

  An optimal process will never break down or require maintenance, will be active 100% of the time, will produce consistently high quality parts, and will operate at its maximum processing speed, giving an OEE of 100%. Though an OEE of 100% is unrealistic, it remains a useful tool to identify areas of concern. The percentage of first-pass yield of a process is often overlooked, and is an important part of lean production.

- *Cost:* Reduction of processing costs is another common goal and can be an effective KPI. The cost of a process is typically obtained by combining overhead and labour rates with material costs, processing costs, energy costs, and the cost of waste.
- *Quality:* Maximising the quality and yield of processes is another very common goal. How this is measured will depend on the process and application, but improving quality reduces the need for rework or customer returns.

However, these are far from the only KPIs available to manufacturing enterprises. As part of the most recent metrics survey conducted by the Manufacturing Enterprise Solutions Association (MESA), 28 manufacturing metrics were identified as being the most utilised by discrete, process, and hybrid/batch manufacturers [3]. These are organised in categories based on what aspect of the business they represent and serve as a starting point for creating KPIs to represent them.

Overall equipment effectiveness may be a common metric for *productivity*, but it is far from the only one. Productivity and efficiency is key to a company's profitability and ability to compete, so productivity goals are extremely common. Some KPIs include:

- *Throughput:* How much product is produced at a machine, line, unit or plant over a given period. Formulas and methods for calculating this are discussed in section 4.2 Conventional Manufacturing Systems Analysis.
- *Capacity utilisation:* How much of the total manufacturing output capacity is actually being utilised over a given period. Again, see section 4.2 Conventional Manufacturing Systems Analysis.
- *Overall equipment effectiveness:* Shown in equation (4.1.1), this is a common measure of the overall effectiveness of a piece of equipment based on the availability, performance and quality.
- *Production attainment:* Percentage of time a target level of production is achieved, enabling an enterprise to meet the schedule agreed with its customers.

Improving *the quality of products* is another common aim for companies, with improved quality resulting in less waste and rework, and improved customer satisfaction. Quality, and improving it, is a vast topic in its own right, but some common KPIs are:

- *Yield:* Percentage of products correctly manufactured without a need for rework or scrap.
- *Customer rejects/returns:* How many times the customer rejects a product. Correct quality assurance policies should reduce this figure.
- *Supplier quality incoming:* Percentage of good quality materials coming from the suppliers is not as uncontrollable as you might believe. Working closely with suppliers can improve the quality of supply, but you may also have to consider alternative suppliers.

Improving *customer experience and responsiveness* is a common goal for companies but measuring it can be a tricky task. Three common KPIs for this are:

- *On-time delivery to commit:* The percentage of time a completed product is delivered to the customer on the schedule agreed.
- *Perfect order percentage:* The percentage of times customers have received a complete correct order on time.
- *Manufacturing lead time:* The time it takes to manufacture a product, from when the order is accepted to the finished product(s) being dispatched.
- *Time to make changeovers:* The time it takes to change a production line from producing one product to a different one, to meet customer demands in a changing market.

Regulatory compliance is conforming to relevant policies, laws and standards in areas such as health and safety, environmental protection, and data security. This is a clearly important area with non-compliance leading to fines and penalties, combined with the risks the regulations are designed to protect you from. Some KPIs in this area include:

- *Reportable Health and Safety incidents:* Measure of the number of reported health and safety incidents over a period of time, including both injuries and near misses that require action to stop them happening again.
- *Reportable environmental incidents:* Number of reported incidents over a period of time, including chemical spills, waste issues, air contaminants, etc.
- *Number of non-compliance events:* Number of times the plant was operating under non-compliant conditions over a period of time.

Profitability and reducing costs are a broad set of KPIs that can include aspects beyond the manufacturing process and into the wider business. The KPIs listed here are generally the KPIs which deal with the manufacturing processes rather than the business as a whole:

- *Total manufacturing cost per unit:* Typically represented excluding materials, this how much the production process alone costs to manufacture a single product.
- *Manufacturing costs as a percentage of revenue:* Related to the previous KPI, what is the ratio of manufacturing costs to the overall revenue of the enterprise?
- *Revenue generated per employee:* Typically a comparison between multiple manufacturing sites, what is the revenue divided by the number of employees?
- *Average unit contribution margin:* Profit made per manufactured product.
- *Return on assets:* Profit made divided by the value of the assets and deployed capital equipment required to generate that profit.
- *Energy cost per unit:* Energy costs incurred per produced unit or volume.
- *Cash-to-cash cycle time:* Time between the purchase of a product by a customer and the collection of payments from the sale of the product.
- *EBITDA:* Earnings before interest, taxes, depreciation and amortisation – a common metric for the profitability of a business.
- *Net operating profit:* One of the purest measures of cost effectiveness, what is the profitability of the enterprise?

Other KPIs worth considering which do not fit into broader categories include:

- *Work in Progress (WIP) Inventory:* Measurement of the efficient use of inventory materials. WIP represents unattained value, and is often a risk to the business if it cannot be rapidly converted into products.
- *Planned vs emergency maintenance:* Ratio of how often scheduled maintenance occurs versus the need for disruptive and unplanned maintenance.
- *Downtime vs operating time:* Asset availability and reliability.
- *Rate of new product introduction:* How quickly new products can be introduced to the market, including the product design, process planning, ramp up, and manufacture.
- *Engineering-change order cycle time:* How quickly modifications to existing products and process plans can be processed and implemented.

At their core, KPIs are a relatively simple concept – they are the measurable goals by which a production process or changes to that process can be monitored and evaluated. However, there are some considerations to make when implementing KPIs to maximise their usefulness.

4.1.4 Implementing Key Performance Indicators

Effectively measuring, analysing, and improving KPIs is not as simple as it may appear. While certain metrics work well for specific processes, it is often the case that there are multiple combinations of metric indicators needed to ensure that a larger business objective is being met. Implementation of KPIs follows a cyclic pattern with five stages:

1. *Definition:* The first phase of the life cycle is defining the KPIs to be used. Although there are thousands of KPIs already defined and used by manufacturers KPIs might sometimes need to be redefined as a more focused or aggregate KPI depending on the exact business objectives.
2. *Collection:* The second phase of the KPI life cycle consists of bringing together candidate KPIs for consideration. Of particular importance is excluding any obviously irrelevant KPIs, as it is common to end up with far too many.
3. *Set Composition:* The third step of the KPI life cycle is to choose from the KPI collection the specific set to implement. ISO 22400-1:2014 Part 1 [4] can help with this process, but the main focus is to ensure the chosen KPIs form a comprehensive set of measuring the business' objectives without being too onerous to implement and monitor.
4. *Implementation:* In this phase, the stakeholders define the process for assessment, examining KPI values and trends periodically and describing action plans for improving process control from KPI values.
5. *Assessment:* Stakeholders evaluate the relevance of the KPIs i.e. how well they align with the current performance objectives of the process, and how well they were implemented. If necessary, the implementation can be adjusted to improve the process.

The process is a cycle. KPIs should be periodically re-evaluated to ensure they're still meeting the needs and requirements of the business. As each KPI is chosen, the data to measure the KPI must be collected, and for that the data must be visible and transparent.

Visibility and transparency are key prerequisites for the optimisation of manufacturing processes. The more information available about a production process, the better performance can be measured through KPIs and better decisions can be made about how to react to events and issues.

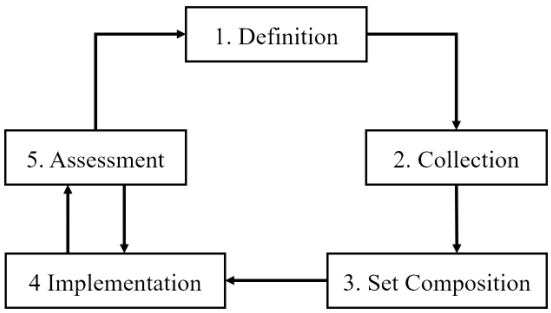

Figure 4.1-7 The KPI Life Cycle.

- *Visible Data:* Data which is easy to access with well-defined processes for doing so.
- *Transparent Data:* Data which is easy to understand and make decisions based upon it.

Despite the importance of process visibility and transparency, real-time reporting schemes with standardized KPIs are still missing in many enterprises. Even when KPI collecting is implemented these often require a manual input process through paper-based forms. Reported data is often based on quantities, e.g. produced units per shift, rather than more insightful metrics. Important data for process optimization such as setup times, change over times, processing times or downtimes are often missing and therefore cannot be reported. This is where modern technologies such as smart sensors and the Internet of Things can be beneficial and will be discussed more in Chapter 6.

4.2 Conventional Manufacturing Systems Analysis

Digital manufacturing techniques are changing the implementation of manufacturing systems and their analysis, with simulation and modelling now common practise across a wide range of sectors and company sizes. One key area in which these techniques can improve the performance of a system is through analysis of the system and identifying areas for improvement. However, the methods used to analyse systems with digital tools are built on a robust framework of more conventional production systems analysis, and understanding these will aid in understanding where and why more complex techniques should be used.

4.2.1 Production Analysis

Productivity (noun): The effectiveness of productive effort, especially in industry, as measured in terms of the rate of output per unit of input.

- Oxford University Press

Of all KPIs to monitor, the productivity of a manufacturing system is one of the most commonly desired, but also one of the most commonly misunderstood due its often-abstract nature. We all want to improve productivity, but what does that really mean?

Productivity is the ratio of output to input from a process. More formally:

$$Productivity = \frac{Units\ of\ Output}{Units\ of\ Input} \tag{4.2.1}$$

The units of output are the products the manufacturing process is creating, but the units of input can be one of several types, including:

- *Capital:* The produced output per unit of capital i.e. products manufactured per Euro/Pound/Dollar spent. This can include the non-recurring value of assets such as tools, plants, and equipment, as well as recurring costs such as maintenance or utility costs.
- *Labour:* The output per person, or more specifically per person-hour. Where manual processes are key, improving labour productivity can have significant effects. This could involve improved ergonomics are processing station, implementation of assistive technologies, or simple morale and motivation.
- *Material:* The output per unit of material, such as raw materials or parts used in the manufacturing process. Improving quality and reducing waste will improve material productivity.

Productivity is a top level goal, and identifying how to increase it requires identifying areas for improvement in a manufacturing system. Productivity is a key performance indicator – a measurement of how well the manufacturing process is performing, a desirable outcome. However, a lower level concept is that of a *metric*. A metric is a measured value which isn't necessarily desirable in isolation, but contributes to a KPI. Traditional manufacturing analysis typically measures metrics, which in turn identify areas for attention which – if improved – could affect the KPI which is the actual goal of the improvement process.

Measuring metrics will enable tracking the performance of production systems over time, allowing the tracking of improvements as they are implemented or to spot issues before they become critical. The techniques here can be applied from the smallest manufacturing lines (which could just be a machining centre with a worker loading and unloading parts) to larger, more complex manufacturing lines.

Calculating metrics does require some application of mathematical formulae, and also having data about the production operations. If you do not have access to this data, it is strongly recommended you start to collect it. Though it is possible to analyse your manufacturing systems via the methods in this chapter or with simulation and modelling tools using only estimated values, the results will be more accurate with real, measured data. Without access to real performance data, identifying ways to improve will be more guesswork than strategic, adding risk that investment in improvements (in terms of time or money) may be wasted. However, the results of calculations performed with informed estimates of performance figures may still offer broad insights and identification of large problems.

Note that these methods are generally concerned with discrete manufacturing systems, rather than continuous ones. There are alternative methods for analysing continuous processes such as chemical production or material processing.

4.2.2 Production Rate

The *production rate* is the number of work pieces a specific production process can produce per hour. Calculating this will enable understanding how long it will take to produce an order, the levels of utilisation of the system, and where bottlenecks might exist. The production capacity of the entire manufacturing line will be discussed in section 4.2.3. The production rate of a process can be calculated with a few short steps, starting with determining the cycle time and then calculating the production rate based on the batching strategy. The calculations should be based on real measured data to get the best results, but approximate measures could be used as a first estimation.

The first step to calculating production rate is to understand the *cycle times.* However, to calculate the cycle time, you must first identify what the work pieces that the process produces are.

- *Work Piece:* The discrete part or product being manufactured by the production system.

  These can be either complete products or parts depending on context. For example, a production line may manufacture entire watches, making each watch a work piece. Alternatively, the company's production system could manufacture only the watch casing as a supplier to a watch manufacturer, making those casings the work pieces. Work piece(s) is typically abbreviated as *pc* (pieces).

- *Cycle Time:* The time a work piece takes to have a single operation performed on it.

For example, the time it takes to mill the watch casing in a machining centre is a cycle time. The time it takes to polish the casing after milling will also have a cycle time.

This is expressed as a time, typically minutes (*m*). To calculate the cycle time, three pieces of information are required:

1. *Operation Time (T_o):* The time a work piece actually spends getting processed.
2. *Handling Time (T_h):* The time a work piece spends being loaded and unloaded from the production process.
3. *Tooling Time (T_t):* The *average* time necessary to set up the tools for the operation, including replacing worn out tools. A tool does not need to be replaced for every part, so the time that takes should be averaged out.

The cycle time is the sum of these three times.

$$Cycle\ Time = Operation\ Time + Handling\ Time + Tooling\ Time$$

$$T_c = T_o + T_h + T_t \tag{4.2.2}$$

For example, a watch casing that takes 5 minutes to mill, 1 minute to load into the machine, 30 seconds to unload from the machine, and requires 10 minutes to replace the mill head every 50 casings would have a cycle time of 402 seconds, or 6.7 minutes.

When you have calculated the cycle time for a process involved in creating the work piece, the production rate can be calculated.

- *Production Rate:* The number of work pieces produced by a production process per hour (pc/hour).

Given that the cycle time of a production process has been calculated, it might be assumed that the production rate is simply the number of cycles that fit into one hour. But this is an approximation that ignores the aspects of how a company batches its jobs together, which can have a significant impact on the production rate. There are four strategies for production that are considered here. These are:

- *Batch-Size-of-One:* Every product being manufactured is unique, and will require the manufacturing process to be setup specifically for each product every time. This is at the extreme end of customisation, and the cost and time of the setup will significantly increase the production cost of the product.
- *Sequential Batch Processing:* Grouping similar products together into batches, but each product still needs to be processed individually. This is the most common strategy for manufacturing, and helps reduce set-up costs.

- *Simultaneous Batch Processing:* A specialisation of sequential batch processing, simultaneous batch processing uses processes that can enable multiple work pieces to be processed simultaneously, such as heat treating.
- *Mass Production:* When a company manufactures parts in very high volumes, the costs of the setup times become so low they can effectively be ignored. This requires a product that almost never changes, and is in high demand.

Each method of production has a slightly different method of calculating the production rate, which are detailed in the following sections.

4.2.2.1 Batch-Size-of-One

At the extreme end of customisation, *batch-size-of-one* production rate is dominated by the change-over and set up times between the different products being manufactured. A company which specialises in job shop production may find itself dealing with low quantities of a product to manufacture, and one-offs are not outside the realm of possibility. The production time for a single item is the sum of two things:

1. *Set up Time (T_{su}):* The time it takes to set up the production process for the unique work piece. For example, loading the required CNC program, and adjusting the clamps to hold the work pieces in a milling machine.
2. *Cycle Time (T_c):* Equation 4.2.2, the time to process a single work piece.

$$Production\ Time\ per\ Work\ Piece = Setup\ Time + Cycle\ Time$$

$$T_p = T_{su} + T_c \tag{4.2.3}$$

The production rate (R_p) is then how many products can be made in one hour. These equations assume all times are expressed in minutes.

$$Production\ Rate = \frac{60}{Production\ Time\ per\ Work\ Piece}$$

$$R_p = \frac{60}{T_p} \tag{4.2.4}$$

For example, a small company makes bespoke watch casings. Each casing takes an average of 10 minutes to mill, but also takes 20 minutes to set up the milling machine for each unique casing. Using equation 4.2.3, *the production time is therefore 30 minutes per casing, and the production rate of the milling machine for bespoke casings is* 2 pc/hr *on average.*

4.2.2.2 Sequential Batch Processing

Batch processing is an extremely common approach to manufacturing, where a company makes a fixed quantity of identical products before switching over the production processes to a new product type. This minimises the impact of changeover time, while still allowing for changing products. Most batch processing is *sequential batch processing* – the products are batched but are still processed individually. To calculate the time required to process an entire batch on a production machine (T_b), three pieces of information are required:

1. *Set up Time (T_{su}):* The time it takes to set up the production process to process the products in the batch. For example, loading the required CNC program, and adjusting the clamps to hold the work pieces in a milling machine.
2. *Cycle Time (T_c):* Equation 4.2.2, the time to process a single work piece.
3. *Batch Quantity (Q):* The number of items in the batch, after which the process will be changed over for the next batch.

$$Batch\ Processing\ Time = Setup\ Time + (Cycle\ Time\ \times Batch\ Quantity)$$

$$T_b = T_{su} + (T_c Q) \tag{4.2.5}$$

The average time per work piece (T_p) is calculated by dividing the batch processing time (T_b) by the batch quantity (Q):

$$Production\ Time\ per\ Work\ Piece = \frac{Batch\ Processing\ Time}{Batch\ Quantity}$$

$$T_p = \frac{T_b}{Q} \tag{4.2.6}$$

The application of equation 4.2.4 will allow calculation of the production rate per hour.

It can be seen that the larger the batch, the smaller the impact of the setup time, as it is spread out over a larger number of products. Optimisation of the batching strategy is important to minimising the cost of each work piece.

For example, consider a company that is making a number of watch casings as a supplier for a company. They want to manufacture 100 casings. The cycle time is 10 minutes, and the setup time is 120 minutes. The company wants to split the order into two batches to allow them to produce other items between these batches to meet their other order deadlines. Each batch of 50 would therefore have a batch processing time of 620 minutes, and a production rate of 4.84 pc/hr.

Alternatively, if the company organised their scheduling to fit all 100 casings into one batch, the production rate would be 5.35 pc/hr, an improvement of 10.5%

over the two-batch strategy. It's up to the company to determine how best to batch its products while still delivering orders on time.

4.2.2.3 Simultaneous Batch Processing

Not all batch processing needs to occur sequentially like milling. Some batching can occur *simultaneously*, such as heat treating or electroplating components. Calculation of the production rate for a process that enables simultaneous processing is hence performed differently to sequential processing. The same method is used as for sequential batch processing, but replaces equation 4.2.5 with the following equation:

$$Batch\ Processing\ Time = Setup\ Time + Cycle\ Time$$

$$T_b = T_{su} + T_c \tag{4.2.7}$$

As can be seen, the processing time is no longer dependant on the size of the batch, assuming the entire batch can be processed in a single cycle. The batch quantity is still required for equation 4.2.6 however.

Consider the company making batches of watch casings. They electroplate the casings in silver, in batches of up to 50. The electroplating has a cycle time of 60 minutes. The setup time is 20 minutes. The batch processing time would be 80 minutes. And if the process is run at the maximum batch size of 50, the production time per work unit would be just 1.6 minutes, giving a production rate of 37.5 pc/hr. The company may not be able to mill casings fast enough to have a high degree of utilisation for their electroplating process unless they purchase multiple milling machines.

4.2.2.4 Mass Production

Mass production is a situation where a company effectively never stops producing a single product, as demand for the product is sufficiently large that a dedicated production process is financially viable. In this circumstance, the impact of the setup time is negligible. The mass-production rate (R_{mp}) is then simply the number of work pieces that a process can produce per hour:

$$Mass - Production\ Rate = \frac{60}{Cycle\ Time}$$

$$R_{mp} = \frac{60}{T_c} \tag{4.2.8}$$

A major watch manufacturer produces cases for their most popular product continuously, including a milling process. Though the milling process did take time

to set up originally, the time this took divided by the tens of thousands of parts produced since then is a tiny fraction of a second and can be ignored.

Note that though the setup time can be ignored, the handling times and tooling times in equation 4.2.2 cannot be ignored and will affect the cycle time of the process.

4.2.3 Production Capacity

Whereas section 4.2.2 was concerned with the expected production rate of individual processes and pieces of equipment per hour, this section looks at the overall production capacity that the equipment enables the company to achieve. This represents the maximum number of work pieces that can be manufactured in a time period, such as pieces per day, week, or year.

Understanding the maximum possible production rate you can achieve with a production line or other sequential set of processes is important for several reasons. It ensures jobs are not over allocated to a facility, as this would result in missed deadlines and delays. It also helps understand the utilisation of production processes, and identify underutilised areas where more value could be made.

4.2.3.1 Production Operating Hours

In section 4.2.2 the number of work pieces which can be produced by individual pieces of equipment per hour was calculated. The next step to understanding the production capacity is to understand how many hours a day the production process is operating. Some companies work a single shift on week days. Others can approach 24 hours a day, 7 days a week. Understanding this is the first step to calculating the production capacity.

The production operating hours per year (assuming all shifts are the same length) is calculated by:

$$Hours\ of\ Production = Number\ of\ Shifts \times Shift\ Length \times Days\ per\ Week \times Weeks\ per\ Year \tag{4.2.9}$$

For example, a company which operates a single 8-hour shift on weekdays, and operates 50 weeks a year has 2000 hours of production per year. A different company which operates two 8-hour shifts 7 days a week, 50 weeks a year would have 5600 hours of production per year.

To approximate the hours of production per day or per week, divide the production per year by 365 or 52 respectively. The figures may need adjusting if you're calculating for a period that includes a Christmas break, for example.

4.2.3.2 Simple Production Capacity

In many cases a company has a quantity of machines, and they produce parts at a roughly similar rate. For example, a company has five milling machines which produce watch casings, with each machine producing at a similar production rate. In this case, calculation of the production capacity of the facility can be calculated with the following information:

1. *Number of Machines* (*n*): The number of similar machines in the company that produce parts at approximately the same rate.
2. *Hours of Production* (H_{pc}): The number of hours over which to calculate the production capacity, calculated with equation 4.2.9. Hours per week or per month can be used here to calculate the production capacity for periods shorter than a year.
3. *Production Rate* (R_p): The number of work pieces each machine produces per hour. This is calculated using the methods in section 4.2.2, specifically equation 4.2.4.

$$Production\ Capacity = Number\ of\ Machines\ \times\ Hours\ of\ Production\ \times\ Production\ Rate$$

$$PC = nH_{pc}R_p \tag{4.2.10}$$

For example, the company with five similar milling machines operates one, eight-hour shift five days a week, 50 weeks a year. The milling machines have a production rate of 4.84 pc/hr. This company therefore operates 2000 hours a year, and has a maximum yearly production capacity of 48,400 pc/year.

4.2.3.3 Advanced Production Capacity

For situations where different pieces of equipment operate at different production rates (PR), a modification to equation 4.2.10 is used. Instead, the production rate of each individual machine needs to be considered separately:

- *Production Rate of Machine i* (R_{pi})*:* for a set of *n* machines, the production rate of a specific one. This is calculated using the methods in section 4.2.2, specifically equation 4.2.4.

$$Production\ Capacity = Hours\ of\ Production \times (PR\ of\ Machine\ 1 + PR\ of\ Machine\ 2\ etc)$$

$$PC = H_{pc}\sum_{i=1}^{n} R_{pi} \tag{4.2.11}$$

Consider a company with three milling machines. They all produce watch casings, but operate at different speeds due to being different models from different manufacturers. The machines are numbered, their production rates are calculated individually as per section 4.2.1, and the results are put in a table below:

Machine Number	Machine Name	Production Rate
1	Faithful Workhorse	4 pc/hr
2	Cheap and Cheerful	3 pc/hr
3	State of the Art	6 pc/hr

Table 4.2-1 The milling machines available to the watch casing manufacturer.

The company works 2000 hours per year. Their weekly production capacity is hence:

$$Weekly\ Production\ Capacity = \frac{Yearly\ Production\ Hours}{Weeks\ in\ a\ Year} \times \left(R_{p1} + R_{p2} + R_{p3}\right)$$

$$Weekly\ PC = \frac{2000}{52} \times (4 + 3 + 6)$$

$$Weekly\ PC = 38.46 \times 13$$

$$Weekly\ PC = 500\ pc/week$$

4.2.4 Capacity Insights

Gathering data and processing the calculations for individual production stations is simply the first step in analysing manufacturing processes. It's important to look at the results and understand what it's telling you. This will enable a manufacturing engineer to make better decisions about their business, improving productivity and profitability. Of particular concern is matching production capacity to the capacity required. Failure to meet the required production capacity will cause order backlogs. Having unused production capacity represents unutilised equipment which could otherwise be producing value for the company.

A company cannot make more pieces per time period than its calculated production capacity. If the order book requires rates higher than your production capacity, the company needs to increase the production capacity or risk delays to delivery times. Similarly, making fewer pieces than the production capacity implies the company could be generating more value. If equipment is sitting idle, it is not generating as much revenue as it could be.

There are many ways to adjust production capacity up or down as required, some which can be short term considerations, and some which are longer term. It's important to consider the time scales involved. In this section, we describe some

options to increase capacity or to mitigate overcapacity in the short, medium, and long term.

4.2.4.1 Increasing Capacity

When an enterprise needs to produce more work pieces than is has the capacity to deliver, there are a number of options available. Which option is chosen will depend on how long the enterprise expects to be over capacity for.

- *Increase hours worked per shift* [Short Term]: For short-term capacity issues, one of the simplest ways is to increase production capacity is to ask existing workers to work overtime on their existing working days. This will increase labour costs, especially if the company needs to offer improved hourly pay to incentivise working overtime, but this is quick and simple to implement.
- *Repurpose existing equipment* [Short Term]: Where an enterprise has multiple production lines and makes multiple products, changing some equipment from one process to another is a way to boost production capacity. The availability of tooling to do this, or the time required to reconfigure and/or reprogram the equipment will determine the speed of this approach, but it is often short compared to other approaches. This can also include moving workers from one line to another.
- *Backlog orders* [Short Term]: Depending on the nature of the orders and the enterprise's relationship with the customer, deliberately delaying orders during short periods of overcapacity may be less financially damaging than adding more workers or equipment. The effect on the company's reputation must be considered.
- *Subcontract out work* [Short Term]: If the period of overcapacity is expected to be short, and delaying the delivery of products is not an option, subcontracting some of the work to other companies may be a solution. This can help fix bottlenecks in the manufacturing process and improve overall production capacity, but beware of increased costs and overheads related to organising the subcontract.
- *Increase the number of shifts per day/week* [Medium Term]: If overcapacity issues are likely to continue beyond the short term, establishing an additional shift to maximise machine processing times may be an option.
- *Increase production rate of bottleneck processes* [Medium Term]: The limiting process of a manufacturing line is the bottleneck. Improving the production rate of that process will improve the production rate of the whole production line. This could involve retraining operators, optimising CNC programs, improved tooling, or other technical improvements.
- *Purchase additional equipment* [Long Term]: If there is a real opportunity to increase revenues by increasing the production capacity, acquisition of more equipment (or more manual workers) may be the ideal choice. Be aware of the potentially long lead times on equipment, and that it will be difficult to get optimised output from new equipment until employees are experienced in its use.

Focus equipment acquisition on bottleneck processes, as these will enable higher production capacities.

- *Redesign manufacturing process* [Long Term]: If a product line has been produced for a long period of time without changing the way it is manufactured, possible efficiencies could be made by changing the manufacturing process and taking advantage of new experience and equipment. Also consider whether some degree of redesigning the product could improve production rates.

4.2.5 Mitigating Unused Capacity

Where far more production capacity exists than is being utilised, cost savings could be made by reducing the production capacity of a production facility or line. Alternatively, strategies to make use of idle capacity could be adopted, creating value from your assets.

- *Repurpose existing equipment* [Short Term]: Where an enterprise has multiple production lines and makes multiple products, if demand for a product is low, consider re-using equipment for other products where demand is higher. This can also include moving workers from one line to another.
- *Stockpile inventory* [Short Term]: If the under-capacity is temporary, and the enterprise makes products that they know will continue to sell in the future, is it possible to use unused capacity to stockpile inventory which will help smooth out overcapacity periods later. This is effectively gambling against the demand for the products, as the stockpiled inventory has no value until sold.
- *Reduce the number of shifts per week* [Short Term]: One of the simplest (and least popular) ways to address unused labour capacity is to reduce the workforce. This could mean implementing redundancy for workers, but an alternative is to reduce the number of shifts. Workers may accept a move from a five day week to a four day week if it means surviving a short-term period with a low number of orders without laying anyone off permanently.
- *Take on extra work* [Medium Term]: If production capacity is unused, consider offering it as subcontracted capacity to other companies. The ability to do this relies on the relationship with other manufacturers and the nature of the spare capacity, but taking on additional work can create value from otherwise idle equipment.
- *Sell equipment* [Long Term]: If equipment is unlikely to be made use of in the medium to long term, it may be worthwhile to claim back some of the value of the equipment by selling it, which can often be a significant return. The ability to sell equipment, and the price it sells for depends on the demand for that equipment. Carefully consider expected future requirements and the value of the equipment before committing to sell.

4.2.5.1 Bottlenecks

Many manufacturing plants do not manufacture single-stage products such as the watch casing discussed in the previous sections, where the casings only have one process applied to them. Instead, multiple operations are executed in sequence to produce the part, forming a production line. A production line may not be a physical entity on the shop floor, but a process followed using multiple distributed pieces of hardware instead, but the approach is the same.

However, calculating the production capacity of a manufacturing line is simplified by a single problem: in almost all manufacturing lines there is a bottleneck. The bottleneck is the process that limits the production capacity, by producing slower than any other process. By calculating the production rates of every process, you can identify the bottleneck as the machine (or set of machines) with the lowest summed production rate. The calculation of the production capacity of the entire manufacturing line is then simply the production capacity of the bottleneck process.

For example, the watch manufacturer mills watch casings with 2 milling machines at a rate of 5 pc/hr per machine, batch electroplates them at a rate of 20 pc/hr, and polishes the casings at a rate of 6 pc/hr. The two milling machines can together produce 10 cases per hour, the electroplating process can electroplate 20 per hour, but the bottleneck is the polishing process. You can increase the production rates of the milling and the electroplating, but you'll just end up with bigger piles of unpolished cases.

Informally, bottlenecks are often easily identified by looking at what equipment is constantly in use, and which equipment typically has large queues of work waiting in front of them. However, formally understanding queues and buffers in a manufacturing system can give insight into how best to arrange your manufacturing processes and at what speed to run them. The next section discusses queueing analysis, and it is highly applicable to manufacturing system optimisation. Manufacturing systems are connected series of systems, either as connected manufacturing processes to form a production line, or as connected suppliers in a supply chain to produce a large or complex product. Understanding how products move and wait in these systems will aid in analysis where optimisations can be made to improve the flow of materials in the system.

4.3 Queueing Analysis

4.3.1 Introduction to Queueing Theory

Queue (noun, British): a line or sequence of people or vehicles awaiting their turn to be attended to or to proceed.

- Oxford University Press

Buffer (noun): a means or device used as a cushion against the shock of fluctuations in business or financial activity.

- Merriam-Webster.com. 2011

Manufacturing typically involves multiple steps as parts flow between machining stations and assembly processes on the shop floor. Unless you're using a pulse flow production line where everything moves between stations in sync it is inevitable that parts will arrive at stations before the station is ready, or stations will be ready before parts are available. Using and understanding queues and the flow of parts will aid in understanding and optimising the production line.

The manufacturing process involves multiple operational steps, converting raw materials into finished products. In order to make the process efficient (e.g. maximising the production rate of a line) and cost effective, analytical tools such as *queueing theory* have been used extensively. They play an important role in the performance analysis, design, planning and control of manufacturing processes, as parts and products in queues are not generating value. Queuing theory is a branch of mathematics that studies and models the act of waiting in lines, and originated in the analysis of telecommunication exchanges handling phone calls. Queuing theory has been successfully applied in modelling production lines to study performance. Queueing Theory requires simplifying the processes in your manufacturing system and modelling them so you can apply formulas to calculate performance measures. By understanding what affects the performance of queues, you can start to record that information, and begin to gather insights into how changes to your production system could affect queueing performance, and therefore overall system performance.

In order to understand how this theory works, first some basic components and characteristics of a manufacturing production line and its queues are first defined. Manufacturing systems have different product-flow layouts that are classified depending on the method of part transfer and the number of part types produced by the system. Part transfer can be carried out in three different ways:

1. *Synchronous* (also known as transfer lines): Where parts are transported simultaneously between workstations.

2. *Asynchronous* (also known as production lines): Where each part moves independently from other parts.
3. *Continuous:* Where parts move continuously at a constant speed.

This product-flow is a waiting line or *queue*, where a sequence of objects (in this case manufactured parts) are waiting to be processed. Owing to the origins of queueing theory lying in human queueing analysis, these objects are generally referred to as *customers*, and the process being carried out on them is referred as a *service*. The term *buffer* is often frequently used in the manufacturing context. A queue includes both a buffer (the 'waiting area') and the service that feeds from the buffer, which in a manufacturing context will mean the production process or station. In practise, the terms are often used interchangeably. Figure 4.3-1 shows the general product-flow layout of mass manufacturing systems.

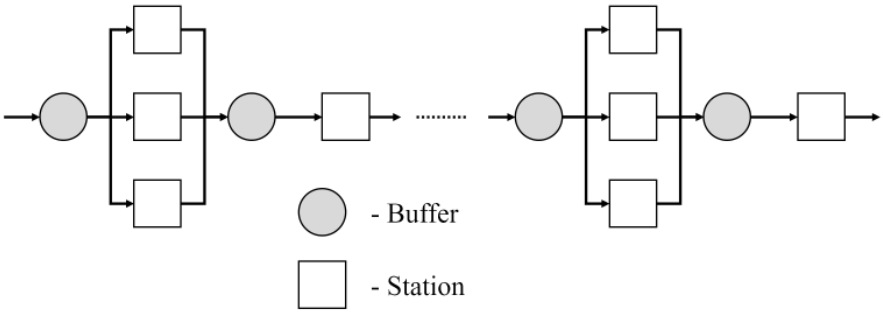

Figure 4.3-1 Product flow layout of mass production systems [5]. It can be seen that when there are parallel stations, those stations can take parts from a single shared buffer.

Production lines are used to produce parts which have a high-volume turnover, and they are characterised by a product-flow layout, low product flexibility (the line is restricted to producing a small variety of part types) and asynchronous part transfer. *Blocking* and *starving* of parts are the main causes of inefficiency in production lines.

- *Blocking:* A part is waiting to be processed, but cannot as the required machine is being used by another part.
- *Starving:* A machine is idle, as it has no input parts to start processing.

These phenomena are mainly caused by variable processing times, and by disruptions to the line caused by the unreliability of stations. To increase the efficiency of the lines, queues are placed between stations.

4.3.2 Queueing Analysis

The impact of blocking and starving on the productivity of a production line with a high production volume is significant. Understanding the queueing in your system will also help identify bottlenecks and inefficiencies as targets for specific processes to be improved. In many cases, simply being aware these problems exist can go a long way to intuitively solving issues. However, more detailed analysis can enable the identification of subtler problems and implementing broader optimisation approaches. To do this, you must model your manufacturing system.

4.3.2.1 Modelling the Problem

Model (noun): a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions.

- Oxford University Press

In order to understand the behaviour of a production system, it can be analysed as a stochastic process (a process with random elements to it). The main interest concerns the distribution of the number of jobs in the system at an arbitrary point in time. From this distribution, it is possible to define how the number of jobs in the system fluctuates, which will allow the computation of important performance characteristics such as the mean number of jobs in the system. The first step for modelling a production line is to characterise the system. The system has different features:

- First, the *arrival* process of the parts is characterised, defining how the parts arrive in the system.
- Then the *service* duration is characterised, which is the time the part is at the station where some operations are performed.
- It is also necessary to specify the *number of stations*, in which several parts could be processed in parallel.
- Parts wait in the *buffer* if all stations are busy, so the total number of parts in the system is estimated by including both the ones being served as well as the ones that are waiting.
- Finally, the *scheduling policy* needs to be specified. This determines in what order parts in the buffer are released for processing.

Common scheduling policies (also called service disciplines) include:

- *First in first out (FIFO):* First come, first served; the type of queue you are intuitively familiar with. Like a queue of customers in a shop, in manufacturing the part which has been waiting the longest is served next.

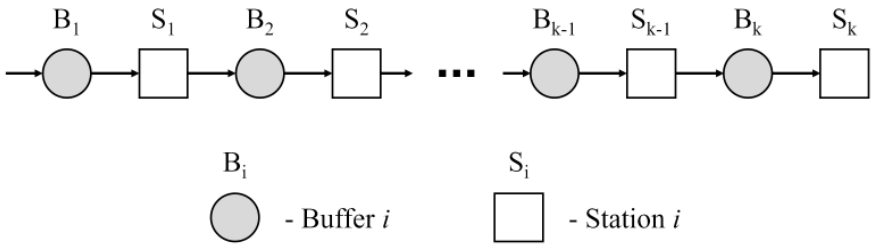

Figure 4.3-2 The basic queueing network model of a production line [5].

- *Last in first out (LIFO):* Imagine a stack of trays in a canteen. The mostly recently replaced tray is placed on the top of the pile, and is also the next tray that will be taken. Sometimes used in manufacturing where the parts are buffered in a stack.
- *Priority:* Where not all parts are the same, parts can be allocated a priority. Products that need to be shipped earlier could be assigned a higher priority, and the next part to be processed is the part with the highest priority assigned.
- *Shortest processing time:* The part that will take the least time to process is used next. This is sometimes used when buffers are reaching their limit and space needs to be cleared.

Figure 4.3-2 shows a common basic model of a production line. The production line consists of *k* stations arranged in series. Each station (S_i) has a buffer (B_i) preceding it. The buffer before the first station may be finite or infinite, all inter-station buffers are finite. Parts enter the system at station 1 and pass through all stations in order. At each station, an operation is performed on each of the parts by a single machine. The parts leave the final station (S_k) in finished form.

The common underlying assumptions when modelling a production line as a queueing network are:

- The line is operated at steady-state conditions (conditions always remain constant through all the production line).
- All random variables are independent.
- All the transport times between stations are zero.
- All failures are single-machine failures and they are operation-dependant (they can only fail while they are operating).
- No parts are scrapped.
- Only a single part type is modelled.
- All buffers utilise the FIFO policy.
- There are enough repair personnel.

4.3.2.2 Part Arrivals

To model the production line, the continuous arrival of the parts at a given workstation needs to be defined. In simple models, it is assumed that the successive inter-arrival times, U_1, U_2, U_3 ... U_i between parts are mutually independent and that they follow the same probability distribution.

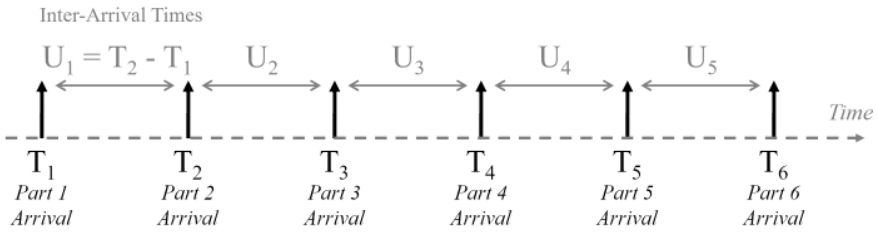

Figure 4.3-3 Part arrivals. T_i is the arrival time for Part *i.* U_i is the time between the arrival of two parts $U_i = T_{i+1}-T_i$

The arrival process of a queuing system is often modelled as a Poisson process. In this process, the inter-arrival times $U_1, U_2, U_3\ldots$ of each part are independent, and arrive according to an exponential distribution with mean λ. If you are familiar with normal distributions or "bell curves", a Poisson process is similar but is discrete (i.e. you can't have fractions of objects arriving). It is described with a single variable – the average number of events per time unit, in this case the number of part arrivals.

If variance is important, the probability that the inter-arrival time U_i is greater than a given value u is equal to $exp(-\lambda u)$. In this case, when the inter-arrival times are independent and identically distributed according to an exponential distribution with parameter λ, the arrival process is said to be a Poisson process of rate λ, where the arrival rate λ is the average number of arrivals per time unit.

For example, λ is typically expressed as the average number of arrivals per time unit. Where 12 parts arrive in a queue per hour, λ = 12/h. Ensure when calculating queue behaviour that all time quantities are expressed for the same time unit e.g. per hour.

4.3.2.3 Service Duration

In order to characterise the system, it is also necessary to properly define the service duration, which refers to the time it takes for the station to perform an operation on the part. The *time of service* is the elapsed time between beginning of service and departure, independently of any waiting time in the queue. For a production process, the time of service is equivalent to the cycle time. In general, it is assumed that:

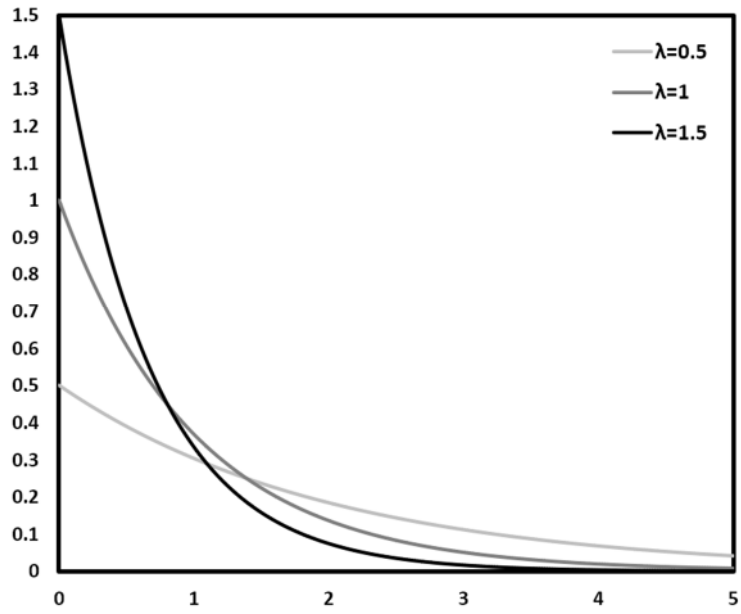

Figure 4.3-4 Three examples of exponential distributions with varying means. In all cases, it can be seen that shorter arrival times are much more likely than longer ones, but occasionally long arrival times will occur – a part has arrived late to a station. These rarer events are typically the ones that cause problems with queueing systems.

- Service times are independent and identically distributed.
- Service times are characterised by their probability distribution.

Much like part arrival times, the exponential distribution is a common model of service duration. It is expressed as the number of processed parts per time unit. In this case it is assumed that if S_i denotes the service duration for a part i, then the probability that S_i is greater than a given value s equals $exp(-\mu s)$, where μ is the mean of the exponential distribution and s is a duration.

A few metrics can be calculated with this information. The *service rate* is equal to μ and is the average number of parts processed (i.e. served) per time unit if the machine (the server) is always busy e.g. parts per minute.

For example, if a machine can process a part every 4 minutes, μ = 15/h (i.e. 15 processed parts per hour). Remember when calculating queue behaviour that all time quantities are expressed for the same time unit.

The *offered load* (sometimes called traffic intensity) is another important metric, representing the expected amount of traffic at a station. The offered load ρ is defined as:

$$\rho = \frac{\lambda}{\mu} \tag{4.3.1}$$

where λ is the average number of arrivals per time unit and μ is the average number of parts a station is able to handle if it is always busy. Offered load is the average proportion of time a server will be occupied, and is an important value for calculating many useful metrics. Offered load must be less than one. If it's greater than one, the station can't process parts fast enough, and the queue will continue to grow forever.

With our example, λ = 12/h and μ = 15/h so the offered load is 0.8.

4.3.2.4 The M/M/1 Queue

Different classes of queues are defined using Kendall's Notation, which defines the features of the queue necessary to perform queueing analysis. In Kendall's Notation, a queue is described with 5 different parameters *A/B/c/K/Z* where the value for each parameter describes the following:

- *A* is the inter-arrival time distribution. M is for Markovian (i.e. exponential as discussed previously), D is for Deterministic (constant), and G is for General Distribution (i.e. an unknown distribution). Other values exist for less common distributions.
- *B* is the service time distribution, and can generally have the same values as the inter-arrival distribution.
- *c* is the number of servers that take parts from the queue.
- *K* is the system's capacity i.e. the maximum length of the queue, plus the number of servers. For this reason it is sometimes written as *K+c*. If the value is omitted, the queue is infinite.
- *Z* is the service discipline e.g. FIFO, LIFO, Priority. Where this is left blank, the discipline is assumed to be FIFO.

The simplest type of queue and the one discussed in this section is the M/M/1 queue, which is a queue with Markovian inter-arrival and service durations, a single server to process parts, and a queue with no maximum length. An M/M/1 queue could be more fully written as M/M/1/∞/FIFO.

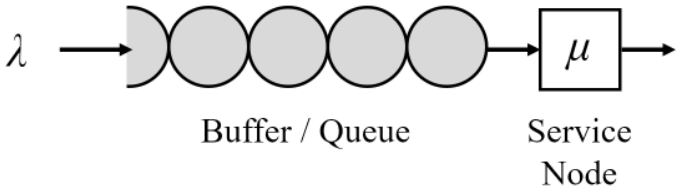

Figure 4.3-5 An example M/M/1 queue.

Although in reality a production system would never be as simple as this model, it contains most of the essential characteristics of a production system and shows the basic ideas and methods of Queuing Theory. By simplifying the model, it is possible to create "good enough" conclusions without the process to arrive at the conclusions being so complex as to no longer be useful.

The analysis of the system consists of studying the evolution of the value $N(t)$, which refers to the number of parts in the system N at time t. "Part in the system" is the number of parts in the queue, plus the parts actively being processed. The value of $N(t)$ can change in two different ways during a time period, representing a transition from $N(t)$ to $N(t+\Delta t)$. If there are n parts in the system, then the following changes can happen:

- If an arrival occurs, the state of the system increases from n to $n+1$. The rate of increase is represented by λ, the arrival rate.
- If a process is completed, the state of the system decreases from n to $n-1$. The rate of decrease is represented by μ, the service rate.

Predicting the value of $N(t)$ allows you to predict the average queue length of the modelled queue, as well as the proportion of the time the queue is over a specific length. Although an M/M/1 queue has an "infinite" queue to simplify the calculations, the real system being modelled will have a limit you don't want to exceed. The transition from one state to another state can take place as shown in Figure 4.3-6.

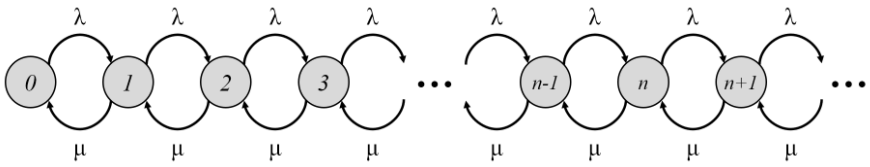

Figure 4.3-6 States of an M/M/1 queue.

4.3.3 Event Probabilities and Performance Measures

Having modelled the queue, you can start to calculate the probabilities of events occurring, such as the average number of parts in the queue, and the average time a part spends in the queue. This will enable identifying issues with your existing queueing system, where buffers may be too large or too small, or where a buffer is at risk of exceeding its capacity. The assumptions made about the system (i.e. Poisson arrivals, exponential processing times, FIFO) make it possible to describe the state of the system at an arbitrary point in time by simply specifying the number of parts in the system. Without these assumptions, the state description would be very complicated and would have to contain not only the number of parts in the system, but also, for example, the residual processing time of the part in service.

At any given point, if the system is in the state *n*, which is the number of parts in the queueing system (including in the service node(s) itself)*:*

- The state of the system moves from *n-1* to *n* at the rate of $(P_{n-1})(\lambda)$
- The state of the system moves from *n* to *n-1* at the rate of $(P_n)(\mu)$

Where P_{n-1} and P_n are the probabilities of being in states *n-1* and *n* respectively. Assuming the system is in a steady-state:

$$(P_n) = (\rho)(P_{n-1}) \tag{4.3.2}$$

Where ρ is the offered load discussed in section 4.3.2.3. Although queues are dynamic and changing systems, by computing the probabilities of being in each state a number of performance measure can be calculated to give insight into the queue and its expected behaviour.

4.3.3.1 Number of Parts in System Probabilities

To calculate the performance measures of the M/M/1 model, first the probability of having *n* parts in the system is defined. If we know the offered load as calculated in section 4.3.2.3, the odds of the system being empty and idle is:

$$P_0 = 1 - \rho \tag{4.3.3}$$

From equation 4.3.2, the odds of having one part in the system is:

$$P_1 = \rho P_0 \tag{4.3.4}$$

That is to say, the odds of having one part in the system is equal to the odds of having zero parts multiplied by the offered load. Then, similarly the odds of having two parts in the system is the odds of having one part multiplied by the offered load:

$$P_2 = \rho P_1 \tag{4.3.5}$$

And by extension the odds of having n parts in the system is calculated as:

$$P_n = \rho^n P_0 \tag{4.3.6}$$

As an M/M/1 queue can only process one part at a time, the queue length is *n-1* (to a minimum of zero). Though an M/M/1 queue is assumed to be infinite, in reality there is likely to be a practical limit above which significant inconvenience or extra cost will occur. The odds of the system having *n* or more parts in it is equal to:

$$P_{n\ or\ more} = \rho^n \tag{4.3.7}$$

The result will be the proportion of the time the system spends with a queue equal to or in excess of length *n-1* (as one part will be in the station being processed).

4.3.3.2 Performance Measures

With the aid of the formulas from the previous sub-section, the following performance measures can be calculated:

- Average number of parts in the queue (L_q).
- Average number of parts in the system i.e. the queue and the station (L_s).
- Average time a part spends in the queue (W_q).
- Average time a part spends in the system i.e. queueing time plus processing time (W_s).

The average number of parts in the entire system (i.e. both the queue and the processing station) L_s is calculated with the offered load ρ, as the offered load represents the amount of 'traffic' in the system:

$$L_s = \frac{\rho}{1-\rho} \tag{4.3.8}$$

The average number of parts in the queue L_q is calculated by multiplying the average number of parts in the system L_s by the offered load ρ:

$$L_q = \rho L_s \tag{4.3.9}$$

Which is equivalent to:

$$L_q = \frac{\rho^2}{1-\rho} \tag{4.3.10}$$

An important law in Queueing Theory is Little's Law, and it states that the number of parts in the system L_s is equal to the arrival rate λ multiplied by the time the part spends in the system W_s:

$$L_s = \lambda W_s \tag{4.3.11}$$

L_s and λ, this can be rearranged as:

$$W_s = \frac{L_s}{\lambda} \tag{4.3.12}$$

As L_s is calculated with the offered load ρ, and ρ depends on the arrival rate λ and service rate μ, this can be simplified even further to be:

$$W_s = \frac{1}{\mu - \lambda} \tag{4.3.13}$$

Note that the time unit of W_s is the same as the time unit of λ and μ. *If* λ = 12/h and μ = 15/h then W_s = 0.333 hours, or 20 minutes.

Lastly, the time a part spends waiting in the queue W_q is the total time in the system, minus the processing time. The average processing time is equal to $1/\mu$ and therefore the average waiting time is simply:

$$W_q = W_s - \frac{1}{\mu} \tag{4.3.14}$$

$$W_q = \frac{1}{\mu - \lambda} - \frac{1}{\mu} \tag{4.3.15}$$

4.3.3.3 Queue Performance Example

Queueing Theory is an extremely effective tool for understanding networks of connected processes, but does require some practise and thought to understand. This section shows a worked example of a queueing theory applied to an example process, which should help you understand how to apply the formulas.

For a machining station on the watch casing manufacturing line, the average rate of arrival of parts is 10 per hour. On average, the station can process parts with a rate of one part every five minutes. Assume the arrival of parts follows a Poisson distribution and the processing of parts at the station follows an exponential distribution. Find the average number of parts waiting in the queue and the average number of parts in the system. Find the average waiting time in the queue and the overall time of a part in the system. Find the odds of the queue exceeding ten parts in length, as this is the longest the queue can be without blocking the previous process.

Poisson and exponential distributions may sound complex, but if your inter-process and service times follow a bell curve (so the average time is common, and much higher or lower times are uncommon) and you can calculate an average time, then they probably fall into this category. It is the variability in these two values that

makes queueing theory important. If part arrival times and service times were constant, queues would be entirely predictable. However, processing is rarely entirely predictable, and even a short run of short part arrival times combined with long service times could overwhelm a buffering system.

Poisson arrivals, exponential service and a single station means this example follows an M/M/1 model and we can use the formulas we have learned. Before we can calculate the performance measures, we need to calculate the arrival and service rates, and the offered load.

- The part arrival rate is λ = 10/h

The part arrival rate is how often parts arrive at the processing station, on average. It is unlikely the parts will arrive exactly every 6 minutes, but instead some parts will arrive slightly faster, some will arrive slightly slower, and more rarely there may be more significant variation. But the average is ten parts per hour. This is a number you will have to measure from your actual production line.

- The part service rate is μ = 1 in 5 minutes = 12/h

The part service rate, is how fast the processing station can process parts, and is also called the cycle time. Like the part arrival rates, here we give an average but each specific processing time can fluctuate. Again, this is something you will have to measure from the processing station itself, remembering to include factors such as the loading and unloading of parts, and any tooling time (see section 4.2.2 for some considerations here). If your processing station is able to process parts in a constant time i.e. exactly five minutes for each part, the formulas will be slightly different. See the next section for information on M/D/1 queues.

- The offered load ρ is therefore:

$$\rho = {}^{\lambda}/_{\mu} = {}^{10}/_{12} = \mathbf{0.833}$$

The offered load is the part arrival rate divided per the service rate, and is a measure of how 'busy' the processing station is. It is an important value for the remaining calculations. Note that the offered load must be lower than 1, otherwise the processing station will be unable to keep up with demand.

- The average number of parts in the system L_s is:

$$L_s = {}^{\rho}/_{1-\rho} = {}^{0.833}/_{1-0.8333} = \mathbf{5}$$

The "queueing system" referred to here includes the processing station and the buffer before it. A single part in the system implies the part is being processed and

the buffer is empty. 5 parts here implies that on average there is one part being processed and 4 parts in the buffer.

- The average number of parts waiting in the queue L_q is:

$$L_q = {\rho^2}/{1-\rho} = {0.8333^2}/{1-0.8333} = {0.694}/{0.167} = \mathbf{4.156}$$

It may seem counter intuitive that the average number of parts waiting is not just the average number of parts in the system minus one, but this is due to parts arriving at variable times during the part processing operation (called "residual time"), combined with the queue sometimes being empty.

- The average time a part spends in the system W_s is:

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{12-10} = \mathbf{0.5}\ \boldsymbol{hours}\ (30\ minutes)$$

The total time a part spends in the queue is the combination of the waiting time, and the processing time. This will of course vary considerably as the length of the queue fluctuates.

- The average waiting time of a part in the queue W_q is:

$$W_q = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{1}{12-10} - \frac{1}{12} = \mathbf{0.416}\ \boldsymbol{hours}\ (25\ minutes)$$

As stated previously, the time a part spends in the system is the waiting time plus the processing time. So the time the part spends waiting is just the total system time minus the average processing time.

- The probability of the queue equalling or exceeding 10 parts is:

$$P_{11\ or\ more} = \rho^{11} = 0.8333^{11} = \mathbf{0.135}$$

The M/M/1 queue is assumed to have an infinite length buffer as this significantly simplifies the mathematics behind it. This is obviously not a realistic assumption for real systems however, so a common application of queueing theory is to calculate how often the maximum queue length is exceeded. Remember that P_n is the number of parts in the entire system (including being processed) not just the queue. Hence the probability of the queue exceeding 10 is the probability of the number of parts in the system exceeding 11.

The consequences of a queue exceeding its maximum limit can vary significantly, but a common issue would be that the previous production station in the production

line is "blocked" i.e. cannot output its parts as there is nowhere for them to go. This can negatively affect productivity, so understanding if this is going to be a regular occurrence or highly infrequent is important. Conversely, an oversized buffer can be expensive or take up a lot of space, particularly for larger parts and products. If the calculations show the system will never require such a large buffer, the buffer can be reduced in size or never purchased in the first place.

4.3.4 Queueing Theory Conclusions

As Kendall's notation implies, M/M/1 queues are a single type in a huge range of possibilities, but are often discussed as they are both simple to calculate metrics for, and encompass a large number of real-world queueing systems with only a few simplifying assumptions. However, they do not cover all queueing types. For example, where the service duration is a known constant (for example, an automated CNC milling operation with a fixed program) you would have an M/D/1 queue. A buffer that feeds into three similar processing stations could be an M/M/3 queue.

For each of these types of queues formulas exist to calculate the performance measures. For example, the average number of parts in the system for an M/M/1 queue is given by equation 4.3.8:

$$L_s = \frac{\rho}{1-\rho} \tag{4.3.8}$$

By comparison, the average number of waiting parts in an M/M/c queue (i.e. an M/M queue with *c* servers) is given by:

$$L_s = \frac{\rho}{1-\rho} C\left(c, {}^{\lambda}/_{\mu}\right) + c\rho \tag{4.3.16}$$

where $C\left(c, {}^{\lambda}/_{\mu}\right)$ is equal to:

$$C\left(c, {}^{\lambda}/_{\mu}\right) = \frac{1}{1=(1-\rho)\left(\frac{c!}{(c\rho)^c}\right)\sum_{k=0}^{c-1}\frac{(c\rho)^k}{k!}} \tag{4.3.17}$$

which is quite clearly a much more complex situation!

This is where computer-based simulation and modelling packages become important – by hiding and solving the complexity for the users, and enabling the calculation of far more complex situations with a lower risk of error and less time investment on behalf of the user.

4.4 Conclusions

Manufacturing analysis is a combination of several aspects – understanding how to formally approach the decision making process, what the question being asked really is in terms of KPIs, gathering data to inform the analysis process, and finally performing calculation and modelling to find an answer to the question. Understanding the process is as important as the formulas and models themselves, as once the process is understood the relevant formulas can be looked up and implemented.

However, as shown at the end of the Queueing Theory section, seemingly simple calculations for manufacturing analysis can rapidly become more complex as the situation being analysed expands beyond small examples. Similarly for manufacturing capacity and production rate analysis, the simpler the situation, the simpler the mathematics required. However, real manufacturing systems are rarely so simple, and the calculations required can rapidly become complex, time consuming, and prone to error.

Though it is important to understand the basics of mathematical basis of manufacturing systems analysis, there exists a substantial range of computer tools to aid in analysis and decision making, removing or hiding much of the complexity and allowing the user to be more accurate, productive, and more able to respond to change. The decision-making process, understand and correctly applying KPIs, and understanding what a tool is suited for and its limitations are all as important for computer-based tools as they are for manual calculations.

The next chapter (Chapter 5) will focus on modelling and simulation computer-based digital tools which are offline i.e. not connected in real-time to a manufacturing system. Chapter 6 will discuss the developing area of digital twins, which are modelling tools connected directly to physical systems and update in real time. It will also discuss decision support systems, which are tools to directly aid in the formal decision-making process.

4.5 References

[1] R. H. Hayes and S. C. Wheelwright, Restoring our competitive edge: competing through manufacturing (Vol. 8), New York: Wiley, 1984.

[2] H. A. Simon, The new science of management decision, Harper & Brothers, 1960.

[3] M. Davidson, "28 Manufacturing Metrics that Actually Matter (The Ones We Rely On)," LNS Research, 9th October 2013. [Online]. Available: https://blog.lnsresearch.com/blog/bid/188295/28-manufacturing-metrics-that-actually-matter-the-ones-we-rely-on.

[4] International Organization for Standardization (ISO), *ISO 22400-1:2014 Automation systems and integration — Key performance indicators (KPIs) for manufacturing operations management — Part 1: Overview, concepts and terminology*, 2014.

[5] H. T. Papadopolous, C. Heavey and J. Browne, Queueing theory in manufacturing systems analysis and design., Springer Science & Business Media, 1992.

Co-funded by the
Erasmus+ Programme
of the European Union