

Business Process Management
Fundamentals and Applications

Kaushik Sengupta, Ph.D.
Associate Dean, Business Graduate Education
Online MBA Program Director
Zarb School of Business
Hofstra University

DRAFT BOOK CHAPTERS
In Preparation for Publication:
Business Expert Press

OVERVIEW

This is a book about business processes. The concepts underlying business processes are presented in a way which would make them applicable to any type of business out there, delivering value to customers in the form of either products or services. The discussions in this book are important because, irrespective of the type of business, the investment for a company lies in its people and its infrastructure. Infrastructure involves facilities, equipment, workstations and transport mechanisms. It is therefore important for businesses to have an efficient process which delivers the product or service in the most cost effective manner in the shortest possible time. An employee's daily work involves various activities that the person undertakes as part of their job responsibilities and hence is tied closely to the processes in which the employee is involved. It is therefore imperative that everyone in the corporate world understand the basic concepts behind effective business process management, and how such an understanding would apply to a specific company's setting.

The discussions in the book are presented in an easy-to-read format which I hope provides a level of understanding of the underlying concepts with no significant background or work related experience. A basic level of understanding of statistics is expected, and for those of you who need a quick overview of such concepts, an Appendix is provided for easy reference.

I hope you can find this book useful from the standpoint of what you do in the workplace on a daily basis and make it better for you and your organization going forward. I welcome feedback and comments on the book.

Kaushik Sengupta, Ph.D.,

Kaushik.Sengupta@hofstra.edu

Chapter 1: Introduction to Business Processes

To discuss the concepts behind Business Process Management (BPM), we have to first define BPM. Multiple sources define BPM as a set of structured processes or activities in organizations such that organizations can achieve the highest level of excellence in the pursuit of the organization's goals and objectives. There are a few keywords in the above definition – *structured, processes (or activities), excellence*.

Organizations are a collection of resources performing a set of processes. In this case, *resources* could be of many types. The most common categorization of resources are grouped as people, equipment or materials. In most service organizations, the primary resource is people. If you consider a banking operation, the employees in the bank are the primary resources. However, many of the tasks performed by a bank are done by other entities or constituents. Two such constituents that perform a significant amount of the work for a bank are: 1) a partner organization that performs some of the outsourced activities, and 2) customers who perform some of customer-facing activities. An example of the former would be when the bank outsources the underwriting and processing of loans to a different organization instead completing these tasks internally. An example of the latter would be the set of activities performed by the customers related to bill pay or activities undertaken by the customers at an ATM.

There are many benefits to this kind of outsourcing. Two primary reasons are cost and efficiency. From a cost perspective, the bank's operating costs would generally be lower as the number of employees required would be lower. By outsourcing to the partner organization, it is also generally assumed that the partner organization is the 'domain expert' in those outsourced activities and hence would be able to perform the tasks at a more efficient and lower cost level. In addition, many of the tasks performed by the customers are automated which implies there is an even lesser need of employees. The overall costs can therefore be significantly reduced. This is even more manifested in "online" banks which do not necessarily have typical branch locations as other

traditional banks do. A few or even a single centralized location for such “online” banks therefore results in operating costs to be much lower; in addition, many of the processes are simpler and more centralized because many of the activities are performed by the customers.

The second area of benefits from outsourcing is on efficiency. Instead of making customers avail of a bank’s services when the bank is open, customers can perform these tasks on their own at their convenience. This obviously makes the whole process more efficient since customers can perform the tasks as and when they need to, and many times outside of the bank’s normal operating hours. One can argue that the development and growth of online banking aided by the advancement of technology has enabled all of these changes to become a reality. Since for any organization, two of the main organizational goals are to reduce costs of operation and to make operations more efficient, the development of technology has simply been tools to assist organizations achieve these two objectives.

From the overall perspective of business processes and from the standpoint of what this book is all about, the tasks or activities have to be performed by someone. Whether this is performed by the internal organization (like the bank), or the partner organization (like the outsourced partner) or the customers, the activities or tasks have to be completed. In this sense, it is immaterial as to who actually performs the tasks. In the end, in order to complete a customer’s specific request or fulfill the demand, the set of activities or tasks have to be completed – therefore, the overall objective of making the processes as efficient as they could be does not go away. We take this more holistic approach of describing business processes in an organization ignoring the organizational boundaries. We consider the customers and the possible existence of the outsourced partners as part of the overall organizational structure. This approach is analogous to a traditional supply chain structure with suppliers and distributors – the holistic approach simply considers the entire supply chain as one entity, and the discussions have the sole objective of improving the business processes for the entire ‘supply chain’.

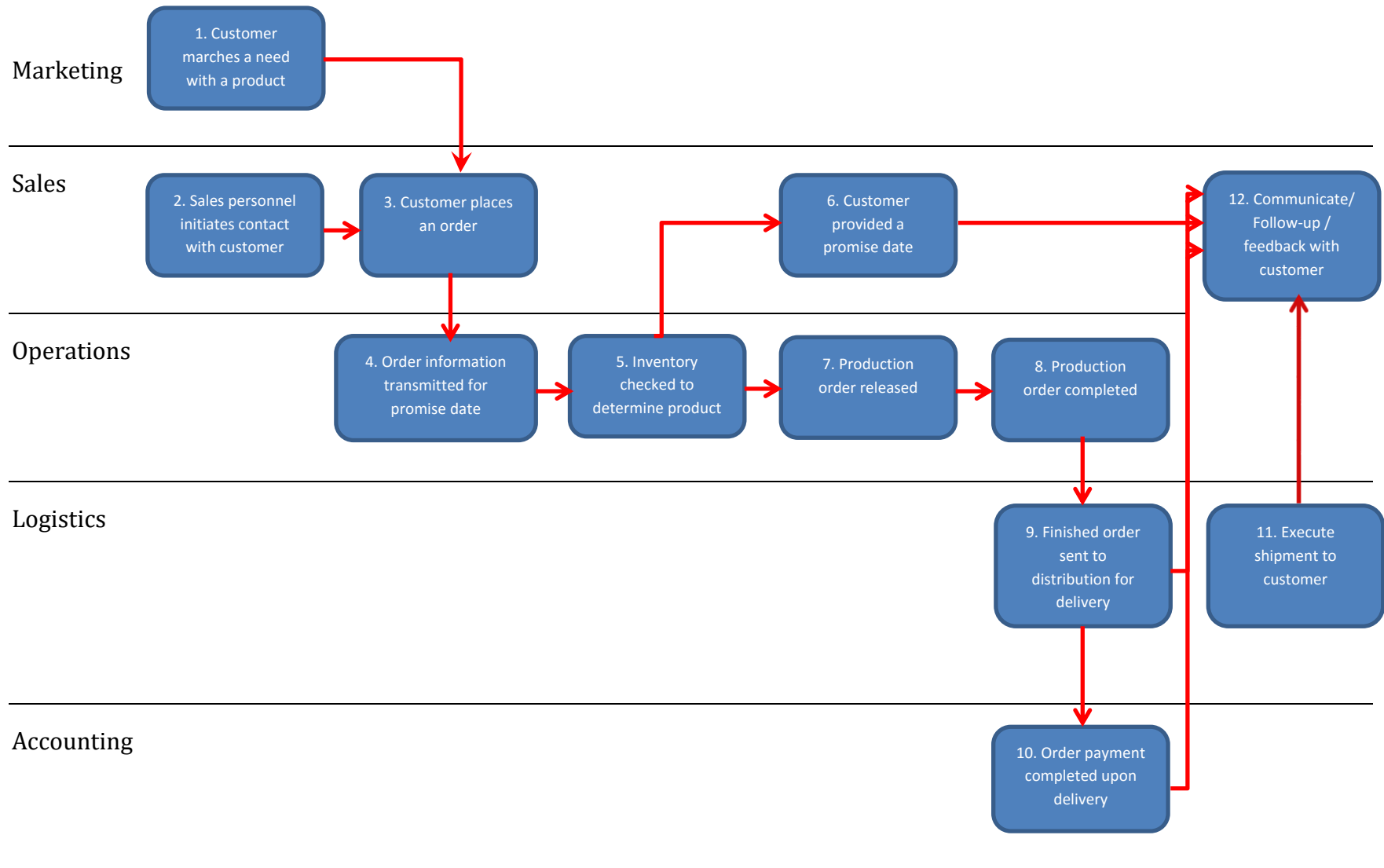
With the above, overall holistic organization in mind, the focus of this book is to explore the details on business processes. If you think of the daily operation of an organization, irrespective of the type of business, most of the investments and efforts spent in an organization have to do with specific tasks performed by various resources within the organizations. Therefore, reducing costs and increasing efficiencies can be possible with a better understanding of business processes.

The Strategic Importance of Business Processes

Many different types of processes exist within an organization. Consider two examples, one from manufacturing and the other from services. An industrial customer for a manufacturing organization places an order for a group of products from the manufacturer. A key metric that the customer would want to know when they place the order is the time it would take for the manufacturer to complete the order i.e. the order taking process in a manufacturing company. This includes the complete processing of the order including the time when the products in the order will be delivered to the customer and the payment for the order. A similar example from a service setting is the example for a restaurant. Various processes are executed from the time a customer walks into the restaurant for a meal. As we all have experienced, key metrics in this case would be the time it takes for the customer to sit at a table, the time it takes for the attendant to take the order, the time to get the food and finally, the time when the customer gets the check for payment. Numerous detailed activities are performed in both examples by different employees (and sometimes with mechanized equipment), some of which are visible to the customer and many which are not i.e. they are performed behind the scenes. Irrespective of whether these activities are visible or not to the customers, the satisfaction of the customer is the primary objective. The processes in both examples should not collectively take more time than the customer's expectation. In both examples, the customer is looking for

- An accurate estimate of how long it would take to complete the entire set of processes
- An adherence by the organization of how they actual fulfill the customer's requirements in relation to the initial estimate that was made to the customer.

Let us go back to the manufacturing example of the fulfillment of a customer's order. The fulfillment of a customer's order consists of numerous steps. Figure 1 shows the

Figure 1 – Customer Order Fulfillment Process

schematic of a typical order taking and fulfillment process. The steps shown in the overall process are performed by various functions within the organization and some of the linkages, as shown, transcend more than one function. Ultimately what is important is that the customer receives an acceptable promise date once the order is placed, and the variance or deviation of the actual fulfillment date from the order promise date. Associated aspects of the order would include delivering the order in the right quantity at the right quality that is acceptable to the customer. However, if one examines the various steps, the key question to ask is – where could the order process go wrong such that the order is delayed or otherwise not delivered as promised? This could happen at any of the steps. For instance, if there is a quality issue with the order, then the problem would lie within the operations function. If there is a delay in delivering the order after it has been manufactured, then the blame would be on logistics. But some of the steps within the specific functions are linked too – for example, who is responsible for understanding the details of the order as discussed between the customer and the sales personnel? You could argue that the initial understanding of these details would be on the sales personnel; however, a critical aspect of ensuring that the specifics are transmitted properly within the organization such that the right order is manufactured lies both with sales and operations, and particularly on the interaction between sales and operations. In other words, some of the process steps are contained within a function while others require an interface between the functions.

The above brings us to one aspect of process management that is important in ensuring that customers experience an efficient operating system from the company that allows customers to achieve the satisfaction from the fulfillment of the order. This has everything to do with the quality of the service offered including but not limited to the time it takes to complete the requirement, the quality of the output and whether all of these were completed to the best of the abilities of the provider organization. Responsibilities for a set of activities that lie within a specific function of an organization are easier to manage and track. For example, if the marketing department of a company is looking for a new advertising agency, the processes involved in evaluating potential advertising firms followed by an effective bidding, evaluation and decision process is usually carried out within the marketing function only. The set of processes involved in

this specific example is easy to track and manage since all of the necessary processes are within one function, and since a typical hierarchical organization structure is oriented by functions, specific activities can be assigned to specific resources and these could be subsequently executed as per the standards and expectations. Evaluation of these activities in terms of performance and efficiency can also be easily evaluated as these are naturally done as part of the overall employee performance evaluation. Consider a very different example – a cross-functional set of processes that are required for fulfillment of a customer order. The entire set of processes is shown in Figure 1. As you can observe, there are several process steps involved and all of these reside within a specific function. However, each process depends on other processes for proper completion – some of these dependencies are housed within the same function while others are dependent on processes from a different function. Wherever there is a dependency on processes in other functions, there are inter-functional interfaces that have to be executed properly in order to ensure efficient running of the entire set of processes. For instance, process step #2 and #3 are within the Sales function but process step #3 depends on step #1 which is contained within the Marketing function. Without an effective interaction and interface between #1 and #3, the customer needs will not be communicated properly for defining the order. Likewise, there are several instances where communications related to a pending order are executed by the sales personnel. But information related to the pending order have to come from other functions. An effective set of processes are required to ensure that the inter-functional linkages are properly set up and executed. This book provides guidelines and concepts that would help a practicing manager decide on the best set of processes and their interactions that would result in satisfying a customer's order.

An associated set of concepts is related to the issue of whether the operating system has sufficient capacity to meet customer demand. Demand forecasts for a company's products and/or services based on previous demand is an art – however, even the most expert forecaster would say it is impossible to come up with an accurate prediction of the future demand unless the product and/or service is so stable that there is no variability in the future sales. Most products/services have a reasonable to a high level of variability. In the light of this, forecasts become important but because even the best of

the forecasts are inaccurate to some extent, effective capacity planning is required to ensure that the organization has sufficient capacity to meet demand when required. Depending on the circumstances, this is a combination of the following –

- the right skill at
- the right time in
- the right quantity

The above is true for the entire operating system i.e. the entire ecosystem of the processes in the organization, not just one or two processes. Is it a magic that some restaurants have an atrocious reputation of poor service during a busy evening while at other more ‘effective’ restaurants, customers are generally happy although they know they may have to wait for some time before they are served on a busy evening? An analysis of the processes in the former case would naturally reveal many inefficiencies while in the latter case, it would reveal a carefully thought-out plan and strategy to ensure that the company has in place the most efficient and effective way of making sure customers are satisfied.

So, take the example of a really well-run restaurant on a busy Friday evening. What would you expect if you go for a dinner outing in such a place? If you are a reasonable customer, your expectations of service would be anchored in expecting good quality food and service based on certain factors like reputation, price and your previous experience at this restaurant. What you would usually not expect is to be able to find a table to sit the moment you walk into the place – it would be great if you do, but as a reasonable customer, you can expect to wait for some time before getting a table. The important question is: how much of a wait are you willing to incur? Is it 15 minutes? 45 minutes? An hour? Factors that determine the ‘reasonable’ extent of wait are many – you are probably willing to wait longer if the restaurant has really great food and/or has a great reputation for service and/or if the prices are great for the type of food you get there. It may simply be because you like eating there – which would be based on a combination of the factors indicated. However, if this expected wait is long, what do you do while you wait? From the business perspective, they could just have a few chairs and sitting areas,

and make you wait. Better businesses would likely have other options to keep the customers engaged while they wait – for example, it may have a nice bar area where you can get a couple of drinks before you get your table. Clearly, the latter is a better option for those who may choose to visit the bar area rather than simply wait. For those who do not want to take this option, it may still be a nice feeling to just know that they can go to the bar if they wish (although they may choose not to). From a design and strategy perspective, however, a conscious decision has to be made by the management about the type of restaurant they want to have. Coupled with these aspects would be factors related to overall capacity and how quickly they may want to serve customers. We discuss these aspects in great detail from a process perspective in the next few chapters. We will first discuss a brief history of the area of process management as this area has been in existence to close to three decades now; processes in organizations have existed ever since we commercialized the production and delivery of products and services. We then discuss some of the basic concepts related to process management. Subsequently we have extensive discussions and analysis on the various models and concepts that are in practice.

Chapter 2 – The History of Business Processes

Business processes have existed ever since we have had commercial businesses delivering a product or a service to customers. A certain set of tasks have always been required to accomplish the delivery of the final product or service. Historically, all of this started with the Industrial Revolution in the 1600s when the first large-scale industries were set up in Great Britain with the cotton mills, steel factories and other industries. It progressed through the next couple of centuries in various forms and the dissemination of these and other industries to the other parts of the world, particularly Western Europe and the United States. Modern day organizations and associated initial concepts of what it takes to make production and manufacturing more efficient (as most of the U.S. economy was in manufacturing at that point) started with the Scientific Management concepts of Fred Taylor with his Time and Motion studies, and the subsequent Assembly Line concept developed for the Model T at Ford. The latter was essentially a simpler version of the many more complicated processes that exist today.

Ford's Model T assembly line was based on a simple, yet cleverly designed series of processes where the car was manufactured and assembled. In the 1920s, Ford was a completely vertically integrated company as it owned the entire supply chain of the Model T – from the steel mills to the car dealers. However, the whole series of process steps were linked which made it possible for customers to buy the first commercially known and affordable car. As is well known, Ford's motto of "You can have any color as long as it is black" was built on the simple premise that standardization of the end-product leads to very efficient processes with little to no variation, thereby allowing accurate production planning to meet demand. In many ways, this was the original model for business process management. Of course, a lot has changed from those days and primarily the change has been that the color is not only just black, but has many different varieties not just in color but other components that goes into the making of a saleable automobile.

As the demand for different types of end products have grown, so has the complexity of the business processes. A bigger change was happening in mid-1900s to the latter half of

the century. American products were selling in hundreds and thousands because of the post-World War II expansion in the economy, fueling the need for U.S. companies to produce more volumes. Over a period of time, companies lost their attention on making quality products – for a while this was fine as the American consumer was more keen on buying a product rather than worrying whether that product had reliable quality or not. During the period from the 1950s to the mid-1970s, this was the status quo in the U.S. economy. Take the auto industry again as an example. Many different types of cars were manufactured and although these were primarily gas guzzlers with questionable quality standards, consumers largely bought them because of the steady demand. The oil shock of 1973 and the subsequent realization that better products can be bought and used changed this dynamic completely.

The economy in Japan was almost an opposite of that in the U.S. Post-war Japan was in shambles, there wasn't much natural resources within the country to start with and the business and government had to figure out a way by which the country could progress. With the lead of certain organizations like Nissan, Mitsubishi, Sony and others the art of making excellent products was slowly practiced and perfected over almost two decades in the 1950s-'70s. By the time the oil shock happened, these companies were ready to export their products to the outside world. It started with Toyota selling cars that were much more fuel efficient and with significantly better quality – aspects that were immediately noticed by American consumers. The whole world also quickly became aware of the prowess of the Japanese products.

Why do we refer to this development in the context of business processes? If you look at the development of the Just-in-Time system with its linked philosophy, it is a perfectly synchronized business process system. The Japanese companies took a lot of efforts to make this system as efficient as it could be over the period of 20-30 years, sometimes introducing aspects of the JIT system from lessons learned from elsewhere. It is well known that the system of warning lights or the *Andon* system to stop an entire production line if an employee faced a quality problem at his/her workstation was initially borrowed from a similar system at the U.S. supermarket check-out lines. Many of the other developments were related to making processes more efficient – starting

with the close network with supplier's processes to the SMED¹ principles of offline setups to minimize downtime. The entire premise of the JIT system was built on the seamless system concept where, based on the final demand of the product, all necessary process steps would be streamlined such that the product could be manufactured with a low inventory level and with few problems. It, therefore, is the essence of business process management.

In the mid-1980s, with the ever increasing demand for products made by Japanese companies (think Toyota, Nissan, Honda, Sony among others), U.S. companies started realizing the effects on their business bottom-line. Consumers started buying products that had better quality, which gave them better value for money. By the time the companies realized this, it was already too late – the quality of the U.S. products was not up to acceptable standards and once consumers realized, they balked. The U.S. based companies were making products that were not up to the mark on quality, processes were defective and the end-product suffered as a result. It took a while and some new ideas to rekindle the need for improvements in processes and quality².

The major concepts in BPM existed before this era. For instance, Little's Law connecting output with inventory and flow time in a process system was first coined in 1955³. What was missing was a context in which to apply these principles. Once the companies realized they need to bridge the gap with the competition in terms of quality and efficiency, these concepts came back to vogue in the framework of what we now know as Business Process Management.

¹ SMED stands for Single Minute Exchange Die → the context being, the die or setup of a machine is a non-productive downtime. So, if this can be offloaded to change the setup for the next product when the current product is running, and if the change can be made quickly (like, within a few minutes), then it saves productive capacity making the overall system that much more efficient.

² Business Process Reengineering (BPR) became a buzzword in the early '90s when companies started looking deeply into their existing set of processes and was looking to dismantle and 'reengineer' the system. Michael Hammer and John Champy's book *Reengineering the Corporation* became a best seller around that time.

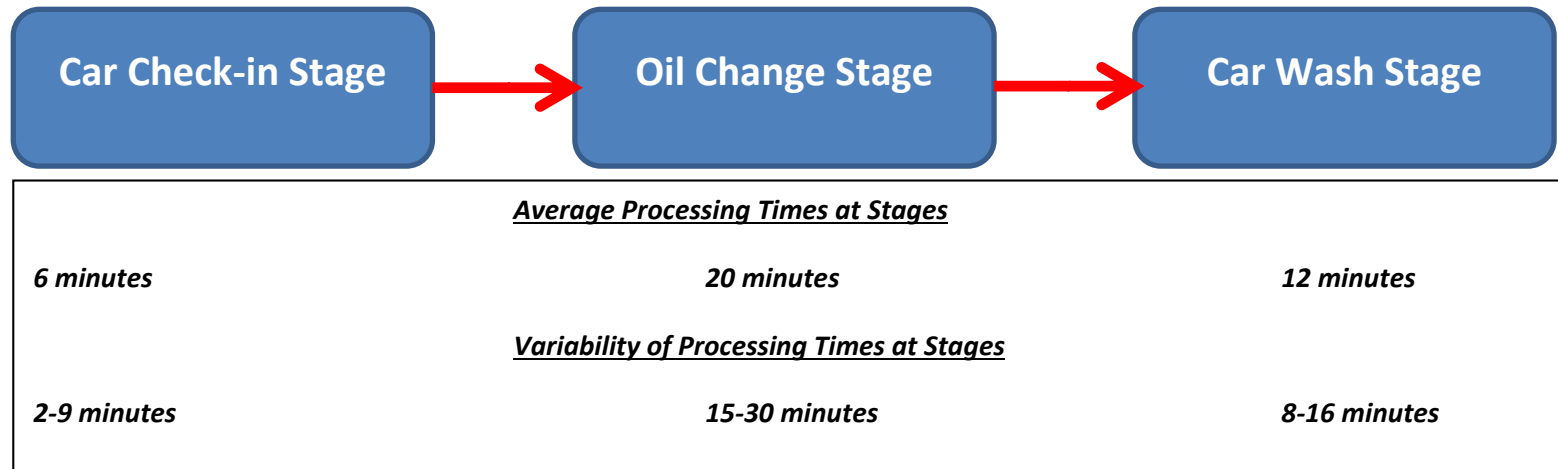
³ Arthur Little's law →

Chapter 3 – Basic Tools of Process Design

We start this chapter discussing some of the basic building blocks of process management and design. First, we provide a clarification on the overall definitions. Process Management entails the gamut of concepts and techniques associated with analyzing and managing business processes. Process design is more of a sub-area within process management where the objective is to either design a completely new set of processes, or redesign an existing set of processes to make it more efficient. In many situations, these two terms are used interchangeably and we will also treat the terms in a similar manner.

Processes are generally and organically defined as specific tasks or activities – an identifiable task or activity would be a specific work task that requires specific resources and takes a certain amount of time and effort by the resource performing the task. Such activities are connected to other activities within the same department or organization. Connections are also required across organizations. Refer back to Figure - 1, the different steps shown within the boxes are specific activities required to perform the overall process of fulfilling a customer's order across multiple departments. In this framework, the connections are shown by the arrows – some of these are within one department while others connect across departments. Figure – 1 is actually a slightly more complicated set of processes – some of these are executed simultaneously and there are instances where multiple activities 'feed' into a single activity. Let us look at a simpler set of processes to discuss the basic concepts of process design.

Take the simple example of a three stage process of an auto oil change and wash business. The first stage entails a customer check-in process; the second stage is the oil change process itself and the third stage is the car wash process. For simplicity, we can assume the oil change and wash are performed in single stages and the business does not offer any other service. We can also assume that each car visiting the facility does both the oil change and the wash. The framework is shown in Figure – 2.

Figure 2 – Basic Process Analysis Framework

Each car coming to the facility would be expected to take a variable amount of time at each of the three stages. Generally, the more complicated the service, the higher is the expected variability. So, if you look at the oil change process as a more complicated process than the wash process, the oil change would have a higher variability. However, to a certain extent there will be inherent variability in each stage. For example, the customer's credit card may be declined during the initial check-in stage. In this case, it will take longer than the average time to complete processing at the first stage. Likewise, a very dirty car will require more time for the wash stage. We will address this issue of variability shortly. Let's first look at the situation assuming each stage takes a constant amount of time.

System Analysis with Constant Processing Times

In all of the analysis below, we assume a **steady state system**. A steady state in this sense implies that we ignore start-up conditions i.e. when the system is ramping up. If you consider a steady arrival rate of the cars and assume the facility being a 24-hour operation, the state of the system after the initial start-up when all three stages are operating on cars would be what we know as steady state. In reality this is obviously not the case, but if you consider the long term aspect of the business, the start-up periods are much shorter than the steady state periods. So a general steady state assumption works well to understand how the system behaves over the long run. Since we view the steady state mode as a long-term view of the system after the initial start-up phase, the number of cars that can be processed by the system will determine its long term profitability. More about the profitability aspect later. Let us first determine some metrics in the system. These would be the basic process design tools that you would need to analyze the current state of the system and seek ways of improvement.

First, define a **bottleneck** as the stage that is the slowest in the system and the stage that eventually defines the overall system capacity. In this case, the slowest stage is the oil change stage as it takes 20 minutes on an average car to get through that stage. And since this is the slowest stage, this also determines the number of cars you can process in a given time period. So, on an average, we can process 3 cars per hour at the steady state. Note that although the other two stages can process a higher number of cars per

hour, the system output is determined by the bottleneck stage. In terms of standard terminology, **steady state output** or **steady state throughput** is 3 cars per hour. This part may be obvious in terms of the system output. However, what happens to the rest of the system? Let us look at the following additional aspects.

1. If a steady stream of cars arrives to the system, what happens to these cars after they check in? Since the first stage can process a car in 6 minutes, it can process 10 cars per hour. But since the next stage can only do 3 per hour, 7 cars are left queued between the two stages. You need to find a parking spot for them while they wait for the oil change. This brings two additional aspects –
 - a. The above happens during every hour of the steady state operation. So 7 cars queue up in the first hour, an additional 7 in the 2nd hour and so on. You can see how the system backs up and you quickly run out of the parking space. This is the notion of **blocking** as the first stage gets blocked and the cars cannot move forward.
 - b. Because you can see the above is unrealistic, it probably makes sense that you should not be checking in up to 10 cars in the first stage. In fact, if you don't have parking space between the first and second stage, you should not check more than 3 cars per hour as that is the number the bottleneck can handle. Blocking is not a good sign as that means your bottleneck cannot process items fast enough to meet demand – in this case, the solution would be to increase capacity if the demand is higher than capacity. More about that later – however, this also means your check-in stage should run at less than 100% capacity. In this example, it should actually run at 30% of its available capacity (3 car-capacity at the bottleneck/10 car-capacity at the first stage). Therefore, while blocking is created due to the imbalance in the stage capacities, it leads to **underutilization** of resources at the first (generically, the **non-bottleneck stage**). In this and other scenarios, the underutilization is actually desired because otherwise you would be accepting customers to check in but you will be making them wait because you cannot process them fast enough.

2. What is the situation at the third stage? The capacity for this stage is to process a car on an average of 12 minutes i.e. its capacity is 5 cars per hour. However, it cannot do this many as it is dependent on the previous stage which can only process up to 3/hour. In other words, this stage does not get blocked, but it gets **starved** of jobs. The bottom line is this stage is also idle for a portion of the time. In this scenario, it will effectively run at 60% capacity (3 car-capacity at the bottleneck/5 car-capacity at the 3rd stage).
3. How much time does it take for an average car to be processed through the system? If you just consider the process time, then this is the sum of the processing times at the three stages (38 minutes from start to finish). This time ignores any waiting time if there are a certain number of cars in queue before a particular car. This is defined as the **throughput time** and it equals *sum of (processing times + waiting/queue time)*. The concept of throughput time is important as you need to give an idea to your customers the expected time it would take for them get their car serviced through the facility. And, of course there is a level of uncertainty in this promise as the waiting time would be a variable number based on the number of cars in the system at a given time.
4. What can you say about the long-term state and performance of this system? In a simple sense, we can see that the system is not meeting external demand if the external demand is actually more than the bottleneck capacity. And if this is the case over the long term, the bottleneck capacity must be increased to meet the additional demand. In this particular example, your slowest stage, the bottleneck can only process 3 cars per hour. This essentially is also the output you get from the system at steady state. And, so if your demand is actually a situation where you have more than 3 customers per hour, you are leaving money on the table by not being to meet this demand. We will discuss this with regards to capacity addition later.

System Analysis with Variable Processing Times

Let us now consider the variability in the process times at the stages. Refer back to Figure-2 and realize that the variability is actually more of a reality than not. It is expected that different units will have different service requirements and so an oil change process on a particular car may take more or less time than that on a different car. This makes the system more complicated. Take the example with the variability numbers again. Note first that the numbers shown are the ranges of the processing times at each stage. So, the check-in stage could take anywhere from 2 minutes to 9 minutes. What happens if a car takes the maximum possible times at each of the three stages? Well, you can see the bottleneck is still the oil change stage, but the throughput time has increased to $9+30+16$ or 55 minutes. Quite long for a service like this!!

What happens if a car takes the minimum time at the first two stages and takes the maximum time at the 3rd stage? It could happen if the car came in very dirty increasing the time to wash it at the 3rd stage. The throughput time now becomes 33 minutes, lower than the average by 5 minutes. However, the 3rd stage is now the bottleneck. The bottleneck has shifted to this stage from the 2nd stage – this is case for the ***shifting bottleneck***. While you can see that the shifting bottleneck can happen, what is more annoying is that this can happen at random with any car depending on the specific time requirements. In fact, the ***shifting*** phenomenon only happens because of the variability and it creates a situation where the bottleneck is not uniquely defined. In this particular scenario, if the shifting occurs quite frequently causing the 3rd stage to be the bottleneck instead of the 2nd stage, we would view the 3rd stage as a ***Capacity Constrained Resource (CCR)***. So, a CCR would be defined as a resource that is not the bottleneck but its capacity is close to that of a bottleneck – and thereby can make the CCR a bottleneck at times. You can also look at the 1st stage in this context and note that given the specified ranges, the 1st stage can never be the bottleneck or the CCR. In other words, it would be defined as the ***non-bottleneck***.

Most of these concepts are also ingrained in the *Theory of Constraints (TOC)* methodology popularized by the late Eli Goldratt. Some of aspects that we just discussed tie to the specific points within the TOC as follows.

- *The system should pace to the bottleneck.* We discussed that the system capacity is determined by the bottleneck capacity.
 - *The bottleneck should have maximum (100%) capacity utilization.* This comes from the fact that if you do not run the bottleneck at 100% capacity, you lose output from the system.
 - *An hour lost at the bottleneck is an hour lost forever.* This is tied to the previous point that the bottleneck needs to be utilized to the fullest extent as it limits overall capacity.
 - *Bottlenecks should be closely monitored at all times.* This obviously comes from the above points which highlight the criticality of the bottleneck. Likewise, *CCRs should also be monitored closely* since these are close to the bottleneck in terms of capacity and sometimes become the bottleneck themselves.
 - *Idle time at non-bottlenecks should be planned idle time.* As discussed, the system is paced to the bottleneck where the non-bottlenecks run at less than 100% capacity.
 - *Idle time at the bottleneck should be avoided completely, if possible.* This notion obviously implies that the bottleneck must run 100% of the time. However, this may not happen in reality due to maintenance or other contingencies. This notion basically indicates the need for carefully managing the bottleneck resources because a productive time of less than 100% essentially implies lost productivity from the system
-

Chapter 4. Analysis of Process Flows

The discussion in the previous chapter should have provided you with an understanding of the basic concepts of Process Analysis and this forms the building block for the remainder of the chapters. In the next three chapters, we discuss important aspects of a systemized analysis of process systems by taking a few different examples. This illustrates the somewhat divergent complexities of the systems and their processes. The idea is that with a few varied examples, you should get a good understanding of how to approach process systems in the workplace. Although all the examples are specific to a particular setting, the concepts and methodology adopted are generic enough to apply to any similar or dissimilar process system.

The first step in process analysis is to understand the “as-is” system in terms of the different steps a unit goes through as it is worked on. In our previous example of the car oil change facility, the unit was the car. We will henceforth call this as a ***process unit (PU)***. Based on particular scenario, the *PU* could a car (like in our previous example), a person (like a patient in a hospital), a group of people (like a group visiting a restaurant), a package (like a mailed letter or envelope), a loan application (like a mortgage approval scenario in a lending institution) or a product (like an item being manufactured in a factory setting). There are obviously many other scenarios where the *PU* concept can be applied to specific settings.

Irrespective of the scenario, the processing of a *PU* requires you to understand and know a few details of the system it is going to go through. Let us take the example of an operating system for a pizza delivery business. Of course, all of us have experience in this, starting with ordering a pizza to receiving it by a delivery person. If you think of the pizza you ordered as a *PU*, the *PU* has the following characteristics (this is not specific to the pizza as such but is generic for any *PU*).

1. The *PU* goes through a number of stages in the system
2. The *PU* may be transformed as an entity as it goes through the stages (for example, the pizza as *PU* exists first as a ball of flour dough which eventually

becomes an unbaked pizza sheet. That transforms into an unbaked pizza with the toppings and eventually this becomes the final pizza after baking).

3. The PU will also have additional components which add value to it (for example, this would include aspects such as placing the pizza in the box, the warming bag, and also associated processing including billing and delivery).
4. As the PU goes through the various stages of processing, there are several handover points where one person handles the PU and the following stage is handled by a different person.
5. There are also potential fail points where potential problems may arise in the production and delivery of the pizza. Monitoring of these potential fail points is critical from the standpoint of quality assurance and control.

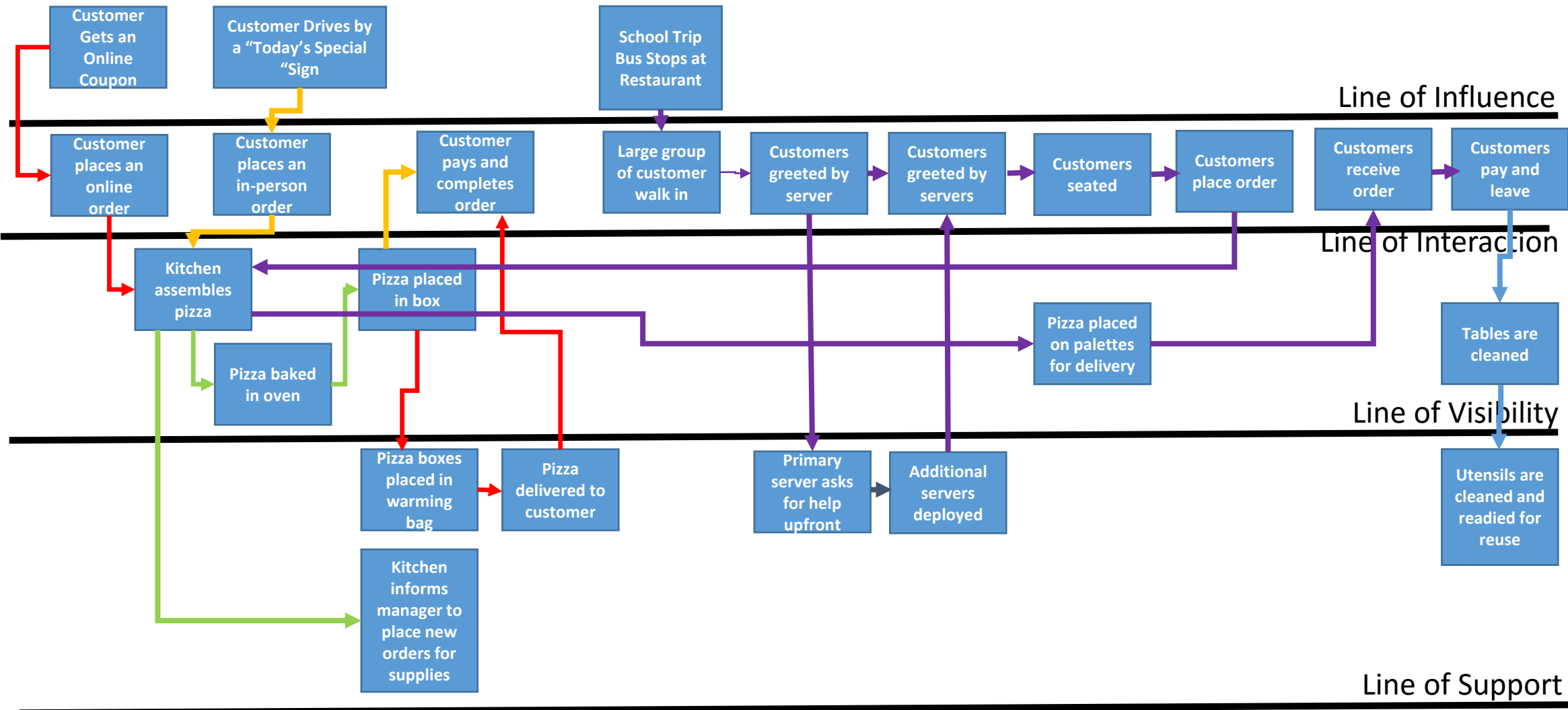
If we consider the above characteristics, and how a pizza gets ordered, manufactured and delivered to the customer, we must be able to map the various stages which the PU goes through to lay out the details. Figure -3 shows such a process. As you may observe, Figure – 3 provides a much greater level of detail regarding the steps in the process compared to the example of Figure – 2. Let us first discuss the elements of the process shown in Figure – 3.

Figure – 3 shows two main ways by which customers place orders – online ordering or in-person ordering. For the latter, we can have two scenarios – an individual customer order in the store and opting for a carryout order, OR eating in the restaurant. We have shown the case of a large group of customers eating in the restaurant as that brings up associated challenges in making sure the order is fulfilled effectively and efficiently.

Note the following aspects in the figure.

- 1) Four separate **process bands** are shown. These are the **Line of Influence**, **Line of Interaction**, **Line of Visibility** and **Line of Support**.
 - a. Line of Influence is the band where customers act on certain promotion/visual/location based factors to decide on placing an order for pizza. This is predominantly marketing based but location has a large

Figure 3: Service Blueprinting for a Pizza Delivery Business



influence in inviting customers to place an order with a specific pizza restaurant.

- b. Line of Interaction is essentially the band where customers directly interact with a part of the operating and delivery system. This could be in the form of interaction with an employee or (for the online mode) an interaction with the company's website.
 - c. Line of Visibility is the process band that includes all activities which are visible to the customer. In this case, we assume that customers in the restaurant can actually see what is going in the kitchen as their orders are prepared and delivered.
 - d. Line of Support is the final process band which includes all back-end support activities. In most cases, such activities are not visible or evident to the customers.
- 2) Each process band contains a number of distinct tasks or activities as shown.
 - 3) Activities within a process band have flows or interactions with activities within the same process band and across other process bands. For instance, the order placement by a customer (either online or in-person) is an activity within Line of Interaction. This leads to the actual order being worked on, starting with the activity '*Kitchen assembles pizza*' within the Line of Visibility. The arcs from one activity to the others show how these activities interact and many of these show the handoff from one activity to the next in sequence.
 - 4) Each activity has a specific processing time with a mean and variability as discussed before. There are also potential **fail-points** where potential quality and other problems may come up during the operation.

It is evident that apart from the Line of Support activities, all other activities are directly involving the customers. In some customers interact directly while in others, these activities are visible to customers. Therefore, any potential problems in these activities are directly experienced by customers, making it all the more critical that the operating system pays close attention to ensuring that customer orders are satisfied to their expected completion value and time.

As-Is vs. To-Be Process

Any operating system like the one depicted for the Pizza Restaurant in Figure – 3 is a set of activities performed on a daily basis. The way the activities are structured including the hand-offs between them and the processing times determine how efficiently a customer order is handled. Therefore, when mapping out a system like the one depicted in Figure -3, it is important to first document the operating system as an “**As-Is System**”, i.e. depict the system as it exists currently. Included in this analysis would be the following in order to get a complete picture of the current system. The list below presents the items in a question format.

- What are the specific activities that are performed as part of the regular operations?
- Can we get some details on the activities? Specifically:
 - What is the process time (mean and variability) at the activities?
 - Which are the activities that immediately precede and succeed the activity?
 - What is handed over from a preceding activity?
 - What is handed to a succeeding activity?
 - What resources are required to execute the activity? (Resources can be people, equipment, materials)
 - How often are the activities performed? (Frequency – hourly, daily, weekly, monthly)?
- What is the most appropriate placement of the activity within one of the process bands?
- What can go wrong during the execution of the activity?


Once all of the above information is gathered about all possible activities in the system, a flowchart like the one shown in Figure – 3 can be constructed to depict the ***As-Is Process***.








The next step is to analyze this process to see if the process can be more efficient in terms of meeting the customer demand. Specifically, analysis of customers waiting, the capacity in the system, the bottlenecks, the layouts and the process flows are the main objectives of such an analysis. We will go through these various steps in the next couple of chapters.

Once the analysis is complete, it may lead to a redesigning of the processes in terms of the work that is done as part of the activities, how they are interconnected and the process times. If the eventual scenario leads to a different process flowchart compared to the existing one, then we have essentially the “***To-Be Process***” as the one that is required to be designed to realize the benefits of the more efficient process.

Standard Notations for Process Flow Charts

Flow charts like the one shown in Figure – 3 provide a comprehensive picture of a complete operating system. Several standard notations are used to represent different elements of an operating system. For example, it is common to have *PU*s be delayed in the system simply because the *PU* may be waiting for the next process. In such a case, we will use a Delay block to show this wait; the block represents the letter D. There are frequently decisions within a process, where based on the decision, the output can be more than one possible paths. In such a case, we will use a Decision block which is shaped like a diamond. Other common notations are used to represent the various elements to comprehensively define a process. The following table shows the common, standard symbols used to create a process flowchartⁱ. Several other blocks not shown in the table are less commonly used in practice.

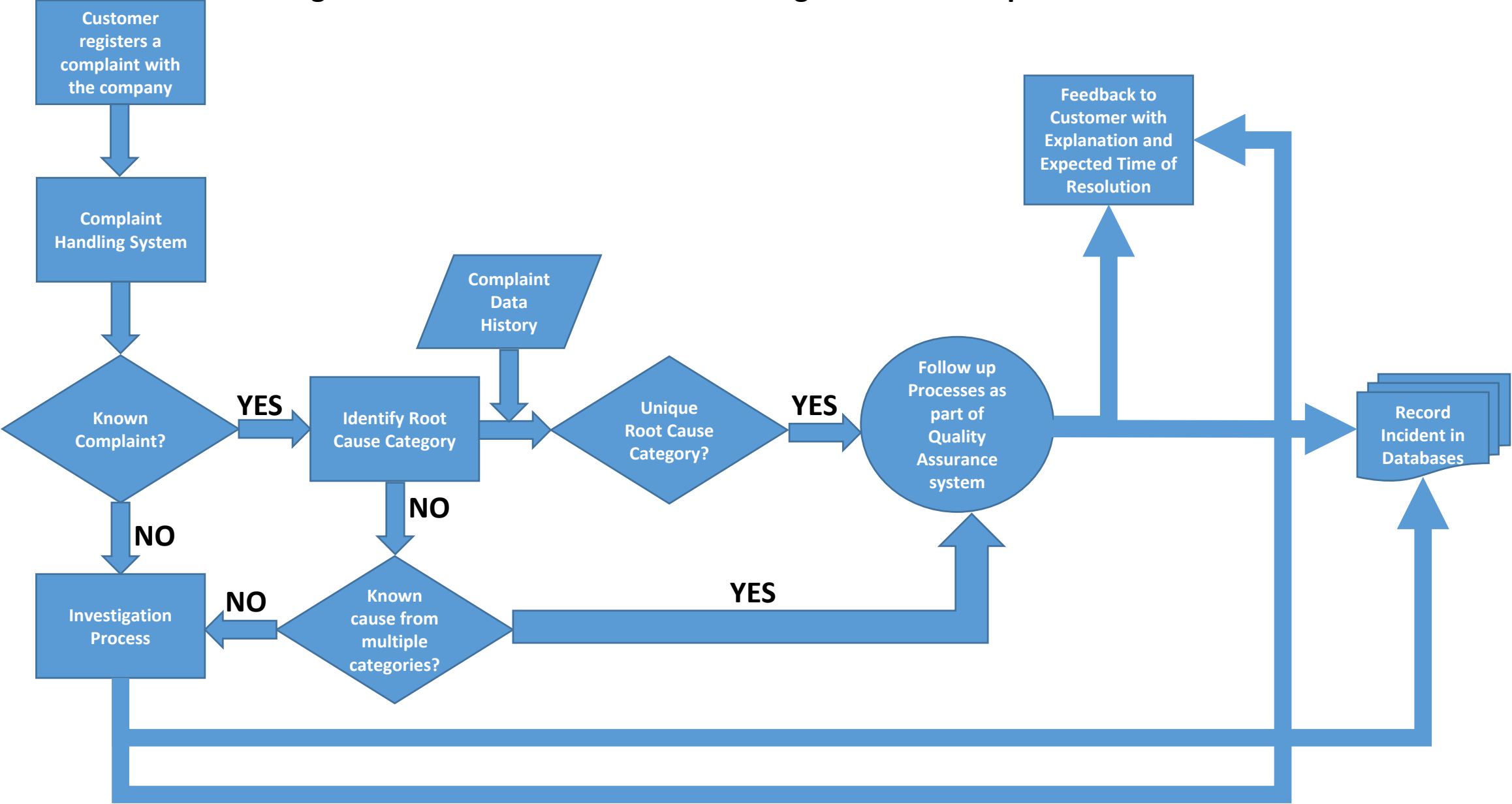
<u>Flowchart Notations</u>	<u>Descriptions</u>
	A Rectangle or Square shows a Task or Activity. Usually this is a process step. The block will usually have an input and/or output arrow.

<u>Flowchart Notations</u>	<u>Descriptions</u>
	A direction or arrow connecting one element of a process to another
	A diamond shows a decision point in the process with the question written within the block. Multiple arrows come out of the block, usually with a Yes or No qualifier to the arrows.
	A block shaped like a D denotes a delay or a wait.
	A circle shaped block that denotes a connector, or when the flowchart continues to a following page.
	A trapezoid block that shows an input or output.
	A block that indicates a document.
	A block similar to the Document block representing a cascaded set of documents or multiple documents.

The utility of being able to use the flowchart notations to depict different elements of a process and thereby the ability to provide a complete analysis of an operating system makes this approach of analysis versatile and generic. Such a methodology can be applied to any operating system irrespective of the organization or industry. Many such illustrations of flowcharts for various kinds of operating systems are available online. We present one such example in Figure 4 which shows a possible flowchart for a quality/customer complaint process.

Figure 4 shows the possible set of actions and decisions a company would follow when it receives a complaint from a customer regarding its product or service. Any small or large organizations must have a pre-defined set of actions when a customer complaint is

Figure 4: Process Flowchart for Handling Customer Complaints



registered in the company's systems. The process flowchart shows that when a company receives a complaint, the tasks that follow go through a systematic analysis of analyzing the complaint as to whether this is a known complaint. This particular step is a decision block. As shown, if the analysis at this step reveals that this is indeed a known complaint, then follow up activities are triggered to understand the root cause of the complaint. This subsequently leads to activities where the root cause may be due to a single cause or multiple causes. Follow-up activities are executed as part of the organization's efforts to improve quality in eliminating the known root causes of the problems in its products/services. As shown in the figure, the circle shaped block depicts the connector that leads to these subsequent activities (these are not shown in the figure). Final steps in the flowchart shown indicates two blocks – one is an activity where the customer receives an appropriate feedback from the company with an expected time of resolution that would address the complaint. The second block is a documentation step which records this incident and stores it in the database for future data analysis and retrieval. It may be noted that the original decision block of whether the customer complaint is a known complaint can only be determined if the company maintains a meticulous process of storing the data related to previous complaints and the associated root cause investigation and analysis. Therefore, the documentation step is a very important process in any organization's processes to be able to respond effectively to customer complaints with appropriate responses and a course of action if the identified root cause/s of the complaint is still in the process of being resolved.

As is evident from these couple of examples, process flows and its analysis is vital to an organization's success. Without a complete knowledge of the exact steps in a set of activities and how they relate to other activities within the organization it would be virtually impossible to understand how the day-day activities within a company lead to an effective product or service delivery system.

While the importance of a sound process flow analysis is well documented, there are many challenges to determining and developing an accurate picture of an organization's business processes. Due to human nature and partially due to lack of documentation, several organizational factors lead to scenarios where a sincere effort to understand a

company's business process from the perspective of making the operations more efficient may be derailed. Factors could include misguided organizational priorities and traditions, employee resistance to change because of perceived loss of control, lack of clear data, lack of technical skills and non-availability of critical resources. ⁱⁱ It may be obvious that a detailed business process analysis requires a significant amount of data and unless the data is very clearly and unambiguously maintained, the primary source of information would be the employees. Questions related to why a particular process is executed in a specific way can lead to answers such as – “this has always been the way we have done things”. Questions related to why a long standing process should be changed to make the overall system more efficient can lead to resistance as employees may not perceive the benefits from such a change. Worse – employees may actually see the benefits but they may still resist because they may feel a loss of control and even loss of jobs. Obviously such resistance can lead to barriers in effectively implementing a more efficient set of processes ⁱⁱⁱ. It is important to realize the potential derailments that may occur due to such factors. Many research studies have been conducted in this area. Suggested approaches of mitigating the effects of such factors include effective and strong project leadership from senior management, effective and clear communication to employees from senior management with regards to why the process analysis and improvement is necessary, proper training of employees to empower them with the knowledge of the benefits of process analysis, having a clear value equation which shows the bottom-line value of a process analysis exercise, documenting evidence of past successes, having a proper data management system so that the information is available and instituting appropriate reward and performance incentive systems such that employees are effectively recognized for being supportive and instrumental in the process analysis efforts ^{iv}.

You may look at the above factors and realize that many of the factors deal with senior management. A strong senior management perspective, vision and message to the company employees about the importance of an effective operating system is the key to dispelling the efforts of resistance that may otherwise come up. In addition, investments in IT systems, rewards and incentives can only be done by senior management. Therefore, it becomes critical that there is sufficient vision and support from the

company's senior management in ensuring that the benefits of process flow analysis are fully realized in an organization, especially when the process requires such a detailed level of investment of time and effort from those involved.

Chapter 5: Queuing Concepts and Waiting Lines

Now that we have discussed the overall, basic concepts of process design and business process flows in the previous two chapters, we will focus on a specific component of process analysis related to service systems. Service systems are essentially those operating systems where the primary delivery is a service, not a product.

So, why is this important? That's because any organization providing services is part of this discussion. And, because a majority of the industries in advanced world economies is in services. Think about any service business such as retail outlets, restaurants, banks, doctor's offices, hospitals, airlines, hotels etc. The examples are numerous and across a diverse set of industries. In fact, the North American Industry Classification System (NAICS)^v has 16 of the 20 industry categories in the service sector. The list, shown below, indicates the diversity of the industries in this sector:

1. Utilities
2. Construction
3. Wholesale Trade
4. Retail Trade
5. Transportation and Warehousing
6. Information
7. Finance and Insurance
8. Real Estate, Rental and Leasing
9. Professional, Scientific and Technical Services
10. Management of Companies and Enterprises
11. Administrative and Support and Waste Management and Remediation Services
12. Educational Services
13. Health Care and Social Assistance
14. Arts, Entertainment and Recreation
15. Accommodation and Food Services
16. Public Administration

Therefore, analysis of business processes in such a wide variety of organizations is an important topic. While the basic tools of process design and process flow framework discussed earlier are still valid in service industries, a key component of the analysis in such industries involve the customer interaction and the customer wait period. Let us elaborate this further.

All of us have experienced the services provided by service organizations. For example, when you go to a Division of Motor Vehicles (DMV) facility to say, renew your driver's license or get a new registration for your vehicle, there is a certain expectation in your mind that you will likely have a waiting time at the facility. You also have an expectation of how long it will take you to get your work done at the facility and, from the time when you arrive at the DMV, when do you expect to get out of there? You probably do not mind the time when your case is being processed, like the time between when you submit your application at the check-in counter to actually proceeding through the next few steps to get your work done. But you probably would mind if you have to wait or stand in line for a reasonably long period before you get to that first step in the process.

Is there a way to 'game' the system knowing that the above may happen when you get to the DMV? Well, you know that the DMV opens at 9AM. So, you say to yourself, what if I get there early? You act on that notion and think that if you arrive there at 830AM, i.e. 30 minutes before the place opens, you should be able to 'beat' the others to the line. You may get lucky on this and when you actually arrive, you may see there are about 5 people already standing in the line outside of the building. Or, you may get there at 830AM and see there are already 20 people standing there. You count your good fortune if the former happens; and you probably curse yourself that you should have arrived another 15 minutes earlier!! Either way, there is a lot of uncertainty in terms of how much time you think you need to spend at the DMV to get your work done.

The Uncertainties in a Service System

As evident from the DMV example, there are some uncertainties in the system which makes it harder for a customer to predict how long a service is going to take. Actually, if we isolate the elements of such a system, there are only two primary sources of

uncertainty – one is related to the customer arrivals and the other is related to the service times. Let us take a simpler example to illustrate this.

Consider a one-person walk-in teller counter at a community bank. Customers arrive to the bank and stand in the queue. Once the teller completes service with a customer, the next customer in the queue can walk up to the teller to start his/her service. If a customer arrives when the queue is empty, and there are no other customers at the teller counter, he/she can go directly for service. This is also a single phase system, i.e. the only service customers undertake is at the teller. Once they are done with their work at the counter, customers depart and exit out of the system. This single server, single stage system is the simplest service system that one can have. Two main components dictate the system behavior and how fast customers are processed – the arrival patterns of customers and the time taken to complete a typical service at the counter.

Arrival Times

The arrival rate of customers defines how frequently customers visit the service system. In fact, from statistics, we can assume that there is a certain inter-arrival time between two consecutive customers. Without loss of generality, and with data collection on the arrival patterns over time, we can also determine the average inter-arrival time between two such customers. There is bound to be some variability in this inter-arrival time. However, it is also expected that we can, with a reasonable degree of accuracy, determine the average inter-arrival time. Of course, if the arrival patterns change significantly from one portion of the day to another, then this average inter-arrival time will also change. For instance, in the example of the community bank, it is expected that more customers will likely visit the bank during the lunch hour or after their office hours compared to say between 10am and 12noon, or between 2pm and 4pm. In this case, the average inter-arrival time between 10am and noon, and 2pm-4pm would be different than that during the lunch hour of noon-2pm or after 4pm. If we partition the whole day into these four blocks of time we could easily come up with a situation as follows.

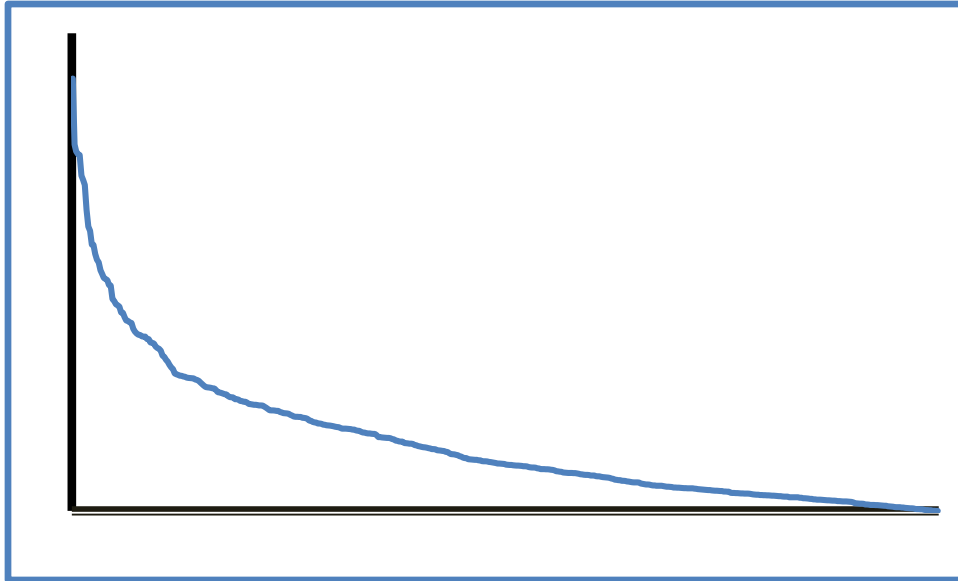
Time Interval	Possible Inter-Arrival Time between two Successive Customers	Average Number of Customers per Hour ⁴
10am-12noon	10 minutes	6
12noon-2pm	4 minutes	15
2pm-4pm	15 minutes	4
4pm-6pm	3 minutes	20

The above pattern of customer arrivals could be a stable pattern on a given weekday or multiple weekdays. Patterns during the weekend may follow a completely different arrival scenario. The bottom-line is that there are certain times during the day when the bank will have significantly more customers than at other times.

The arrival pattern as described above follows a *Poisson* process where the interarrival times are expected to be a steady value as shown within the time blocks in the example. It also implies that a customer arrival is independent of the previous arrival and the next arrival. This is known as the *memoryless property* – i.e. although there is an average interarrival time, the time interval between two customer arrivals is completely independent of the time interval between the previous customer arrivals. The *Poisson* process of the arrival patterns leads to this memoryless property where the interarrival times typically follows an *Exponential Distribution*. The shape of such a distribution will look as shown in Figure – 5.

Standard notation follows that if the arrival rate (number of customers per hour) is denoted by λ , then the inter-arrival time is $1/\lambda$ hours, or $60/\lambda$ minutes. Given that the inter-arrival time follows an exponential distribution, the mean is $1/\lambda$, the standard deviation is also $1/\lambda$ and the variance is $1/\lambda^2$. An exponential distribution with a mean of λ will look like as follows.

⁴ Note the average number of customers per hour is determined by $60/\text{Interarrival time in minutes}$. So when the average interarrival time is 10 minutes, the corresponding average number of customer arrivals is $60/10$ or 6 per hour.



In this case, the X-axis represents the inter-arrival times and the Y-axis represents the frequency. In other words, the distribution is really a histogram of the various inter-arrival times if the system data is collected and graphed over a long time. The shape of the distribution shows that the average inter-arrival time is skewed closer to the Y-axis and it has a long tail which implies there are some instances, albeit few, where the times are excessively long. This is, in fact, a typical behavior seen in many service systems.

One measure of the variability in a measure is the ***Coefficient of Variation (CV)***, where, $CV = \text{Standard Deviation} / \text{Mean}$

Since both the mean and the standard deviation for the exponential distribution is $1/\lambda$, $CV = 1.0$. A CV of 1.0 represents a distribution with a very high variation. So, although we say that a service system has an average inter-arrival time between consecutive customers, the variation in this time is very high leading to many instances where the actual inter-arrival time deviates significantly from the average value. While this may or may not be intuitive if you think about any typical service system, it is the reality is such systems where the systems get really busy at times (i.e. the inter-arrival times are low) to times when there aren't too many customers (i.e. the inter-arrival times are high). High inter-arrival times is the manifestation of the long tail property of an exponential distribution. However, a business will not be a very profitable business if you have high inter-arrival distributions (very low customers). Hence the long tail with the exponential

distribution makes practical sense as the reality of many service systems where there are few instances when the system has few customers but would otherwise be busy when there are a reasonable or high number of customers in the system.

Service Times

Let us discuss the service times next. As indicated before, this is the 2nd component which determines how fast a service can be rendered to a customer.

Once again, in any system, there is a notion of an average service time and a variability of service times. Think about the teller example again. Most customers likely get their work done at the bank in a few minutes (say an average of 3 minutes). However, there will always be cases where a customer will require a much higher time. You can't go too much lower than 3 minutes! But there will be a variability – in this case, we may look at the data again to see how much were these service times? It may show that the service times at the bank range from 2 minutes to 10 minutes.

Think about another common example – the checkout counter at a supermarket. The average time for checkout may be around 6 minutes (I am making up this number – it has to be measured with real observations). Some customers who will have only a single item to check out will take less than 6 minutes for sure. However there will also be customers with a trolley full of items to check out – these folks will take much longer to check out. Then, if there is a problem in checking out (the wrong price scanned on the item causing the checkout person to call the manager to check the price; or an issue with the customer's credit card), that will increase the checkout time too.

It so happens that with the notion of the average service times and the variability around it, the service times also follow an exponential distribution that is similar in shape to the inter-arrival times. Why is this true? Well, service times also follow the same memoryless property as the inter-arrival times. It is a valid and realistic assumption that each customer is unique in terms of what their service time would be compared to any other customer. All of us have been frustrated with standing in queue behind a customer at the counter who is taking an inordinately long time to complete the service – this is basically a manifestation of the uniqueness of the service times.

Thankfully, for all us who have had bad experience of waiting in lines, the frequency of the long service time instances is low – it gets back to our notion of the long tail distribution.

So, in summary, what we end up assuming is that the exponential distribution works as a really good distribution for both the inter-arrival time and the service time. Similar to the average inter-arrival time which is denoted by λ , the average service time is denoted by μ .

What has this to do with the process analysis? Well, unless you have an idea of how fast customers arrive to the system and how fast you can serve them, how would you know how many customers can you serve in a given time? And if your analysis of the service capacity shows that you don't have sufficient capacity to meet customer demand, you have to determine possible strategies which would improve capacity. In most cases, this is a fairly easy decision. Before we get there, let us look at some metrics that can be measured.

We come back to the reference system as a single stage, single phase system.

(END OF DRAFT CHAPTERS)

ⁱ American Society for Quality (ASQ) Knowledge Center, What is a Process Flow Chart? <http://asq.org/learn-about-quality/process-analysis-tools/overview/flowchart.html>

ⁱⁱ Taher, Nader Bin, and Vlad Krotov. 2016. "BUSINESS PROCESS REENGINEERING: ADDRESSING SOURCES OF RESISTANCE AND SABOTAGE TACTICS." *Journal Of Competitiveness Studies* 24, no. 3: 145-163.

ⁱⁱⁱ Ghatari, Ali Rajabzadeh, Zahra Shamsi, and Ali Vedadi. 2014. "Business process reengineering in public sector: ranking the implementation barriers." *International Journal Of Process Management & Benchmarking* 4, no. 3: 324-341.

^{iv} Paper, David, and Ruey-Dang Chang. 2005. "The state of business process reengineering: a search for success factors." *Total Quality Management & Business Excellence* 16, no. 1: 121-133.

^v 2012 North American Industry Classification System, <http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2012>