Chapter 3

Mathematical Modeling of Production Systems

Motivation: All methods of analysis, continuous improvement, and design described in this textbook are model-based, i.e., their application requires a mathematical model of the production system under consideration. Therefore, the issue of mathematical modeling is of central importance. The main difficulty here is that *no two production systems are identical*. Even if they were designed identically, numerous changes and adjustments, introduced in the course of time by engineering and equipment maintenance personnel, force them to evolve so that they become fundamentally different. Thus, there are, practically speaking, infinitely many different production systems. Nevertheless, it is possible to introduce a small set of standard models to which every production system may be reduced, perhaps at the expense of sacrificing some fidelity of the description. The purpose of this chapter is to discuss these standard models and indicate how a given production system can be reduced to one of them. The issue of parameter identification is also addressed.

Overview: The mathematical model of a production system is defined by the following five components:

- *Type of a production system*: It shows how the machines and material handling devices (or buffers) are connected and defines the flow of parts within the system.
- *Models of the machines*: They quantify the operation of the machines from the point of view of their productivity, reliability, and quality.
- *Models of the material handling devices*: They quantify their parameters, which affect the overall system performance.
- *Rules of interactions between the machines and material handling devices*: They define how the states of the machines and material handling devices affect each other and, thus, facilitate uniqueness of the resulting mathematical description.

- *Performance measures*: These are metrics, which quantify the efficiency of system operation and, thus, are central to analysis, continuous improvement, and design methods developed in this book.

This chapter describes each of these components and comments on parameter identification and model validation.

3.1 Types of Production Systems

3.1.1 Serial production lines

Serial production line – a group of producing units, arranged in consecutive order, and material handling devices that transport parts (or jobs) from one producing unit to the next.

Figure 3.1 shows the block diagram of a serial production line where, as in Chapter 1, circles represent producing units and rectangles are material handling devices.

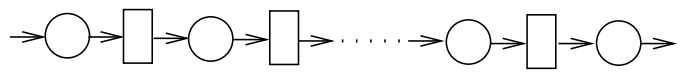


Figure 3.1: Serial production line

The producing units may be either individual machines or work cells, carrying out machining, washing, heat treatment, and other operations. If assembly operations are performed, the parts to be attached to the one being processed are viewed as produced by another production system and, therefore, the line is still serial (rather than an assembly system – to be considered in Subsection 3.1.2). The producing units may also be departments or shops of a manufacturing plant. For instance, they may represent the body shop, paint shop, and the general assembly of an automotive assembly plant. Finally, the producing units may even be complete plants, representing various tiers of a supply chain. However, since the emphasis of this book is on parts flow rather than on the technology of manufacturing, we refer to all producing units as *machines*.

The material handing devices may be boxes, or conveyors, or automated guided vehicles, when the producing units are machines or work cells or shops in a plant. They may be trucks, trains, etc., when the producing units are plants. Whatever their physical implementation may be, we refer to them as *buffers*, since the most important feature of material handling devices, from the point of view of the issues addressed in this textbook, is their storing capacity.

The buffers, discussed above, are called *in-process buffers*. In addition, serial production lines may have *finished goods buffers* (FGB). The purpose of the latter is to filter out production randomness and, thereby, ensure *reliable satisfaction of customers demand by unreliable production systems*. An example of a serial line with a FGB is shown in Figure 3.2.

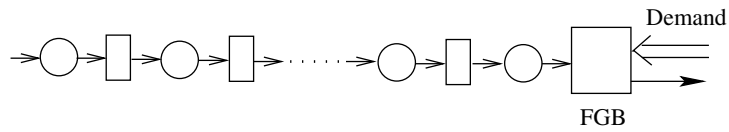


Figure 3.2: Serial production line with a finished goods buffer

In some cases, parts within a serial line are transported on *carriers*, sometimes referred to as pallets, skids, etc. Such lines are called *closed with respect to carriers* (see Figure 3.3). Here, raw materials must be placed on a carrier, and the finished parts must be removed from the carrier, returning the latter to the *empty carrier buffer*. Thus, the performance of such lines may be impeded, in comparison to the corresponding *open* lines, since the first machine may be *starved for carriers* and the last machine may be *blocked by the empty carrier buffer*. Too many carriers lead to frequent blockages of the last machine; too few carriers lead to frequent starvations of the first machine. Thus, an additional problem for closed lines is selecting a "just right" number of carriers.

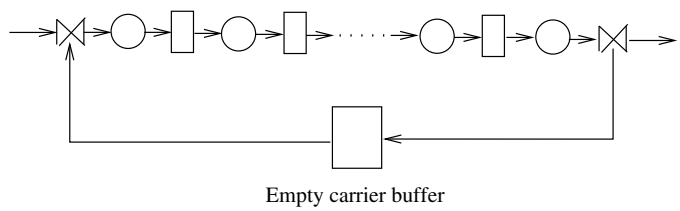


Figure 3.3: Closed serial line

Along with producing units, serial lines may include *inspection operations* intended to identify and remove defective parts produced in the system. Such a line is shown in Figure 3.4 where the shaded circles are the machines, which may produce defective parts, and the black circles are the inspection machines; the arrows under the inspection machines indicate scrap removal.

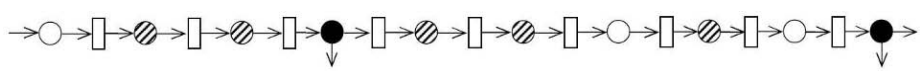


Figure 3.4: Serial line with product quality inspection

Another variation of serial lines is production lines with *rework*. Here, if a defective product is produced, it is repaired and returned to an appropriate operation for subsequent re-processing. An example of a serial line with rework is shown in Figure 3.5. Such lines are typical, for instance, in paint shops of automotive assembly plants.

A generalization of lines with rework are the so-called *re-entrant* lines, illustrated in Figure 3.6, where some of the machines are represented by ovals to

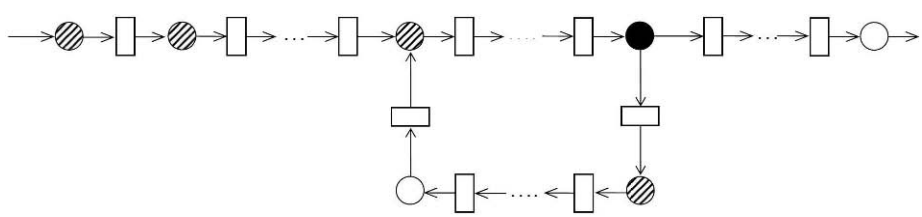


Figure 3.5: Serial line with rework

better indicate the flow of parts. Here, each part may visit the same machine multiple times. Typically, this structure arises in semiconductor manufacturing where, on the one hand, equipment costs are extremely high, and, on the other hand, the products have a layered structure, which necessitates/permits the utilization of the same equipment at various stages of the production process. Clearly, these lines may have even more severe problems with blockages and starvations and, therefore, their performance is typically inferior to corresponding "untangled" serial lines. In addition, since each machine serves several buffers, priorities of service become an important issue.

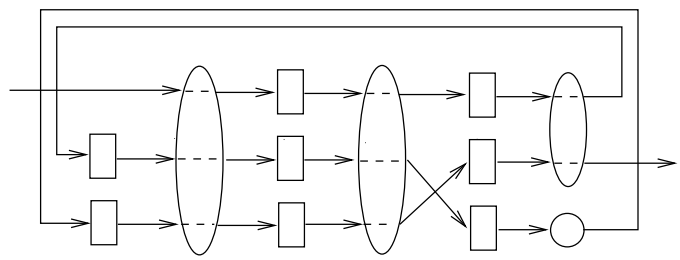


Figure 3.6: Re-entrant line

The serial production line is a "work horse" of manufacturing. It is hardly possible to find a production system, which would not include one or more serial lines. Moreover, all other production systems may be broken down into serial lines connected according to a certain topology. Thus, the study of serial lines is of fundamental importance in Production Systems Engineering, and it is a major component of this textbook (Parts II and III).

3.1.2 Assembly systems

Assembly system – two or more serial lines, referred to as *component lines*, one or more *merge operations*, where the components are assembled, and, perhaps, several subsequent processing operations performed on an assembled part.

Figures 3.7 and 3.8 show the block diagrams of typical assembly systems where, as before, the circles represent the machines and rectangles are the buffers. Systems similar to that of Figure 3.8 are typical in automotive en-

gine plants where the horizontal line represents the general engine assembly (with engine blocks as "raw materials"), while the vertical lines are various departments producing engine parts, such as crank shaft, camshaft, etc.

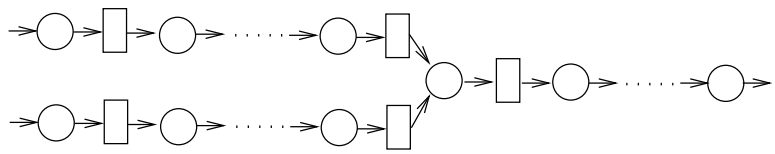


Figure 3.7: Assembly system with a single merge operation

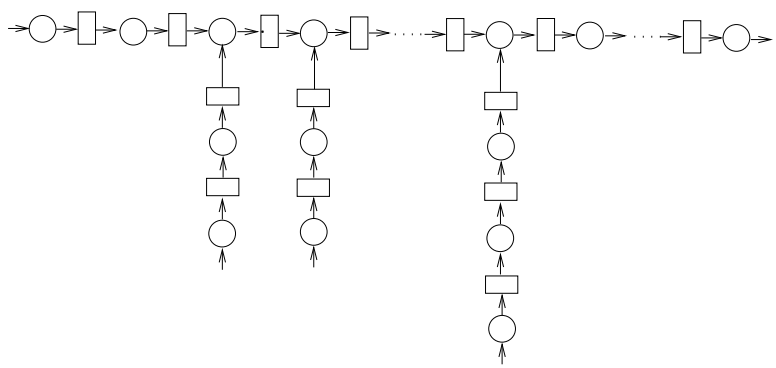


Figure 3.8: Assembly system with multiple merge operations

Clearly, assembly systems may be viewed as several serial production lines connected through their finished goods buffers. Each of these component lines may have all other variations described above, e.g., being closed with respect to carriers or re-entrant. In this book, assembly systems are studied in Part IV.

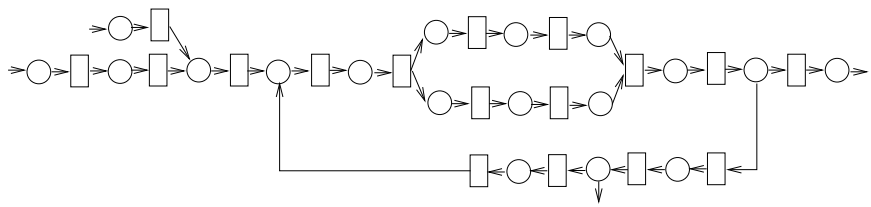


Figure 3.9: Complex production system

While it is highly desirable that a production system under consideration be reduced to either a serial line or an assembly system, it is possible to carry out some analyses (for instance, performance evaluation) for more complex modelsTypes of production systems!complex lines. Figure 3.9 shows an example of such a model.

3.2 Structural Modeling

It is quite seldom that production systems on the factory floor have *exactly* the same structure as one of those shown in Figures 3.1 - 3.9. For instance, a serial line may have multiple machines in some operations, as shown in Figures 3.10 and 3.11. The situation in Figure 3.10 typically happens because no machines of the desired capacity are available for some technological operations. Figure 3.11 exemplifies the situations where a machine performs several synchronous dependent operations in the sense that all operations are down if at least one of them is down. In all these cases, the production systems must be reduced to one of the standard types discussed above (see Figure 3.12) in order to carry out their analysis and design using the tools described in this book. We refer to this process as *structural modeling*.

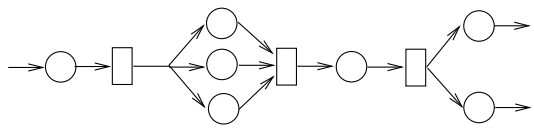


Figure 3.10: Serial production line with parallel machines

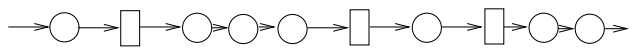


Figure 3.11: Serial production line with synchronous dependent machines

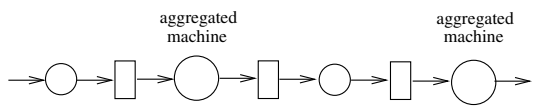


Figure 3.12: Structural model of serial production lines of Figures 3.10 and 3.11

The general approach to structural modeling is based on the maxim attributed to Einstein: "*The model should be as simple as possible, but not simpler.*" The last clause makes the process of modeling more an art than engineering and, like the arts, must be taught through examples and experience. A few examples described below illustrate how this process is carried out, while Subsection 3.3.5 shows how the characteristics of the aggregated machines of Figure 3.12 can be calculated. Case studies in Section 3.10 offer additional examples.

Consider an automotive ignition module production system shown in Figure 3.13, which operates as follows: The raw materials for parts A_1 and A_2 are loaded on conveyors at operations 1 and 9, respectively, and then transported to other operations. At operation 8, parts A_1 are unloaded into the buffer, which is another conveyor and which transports them to the mating (or merge) operation 13, where the assembly of A_1 and A_2 takes place. Operations 14 - 18 perform additional processing.

As it follows from this description, this system can be modeled as shown in Figure 3.14, which is a standard assembly system.

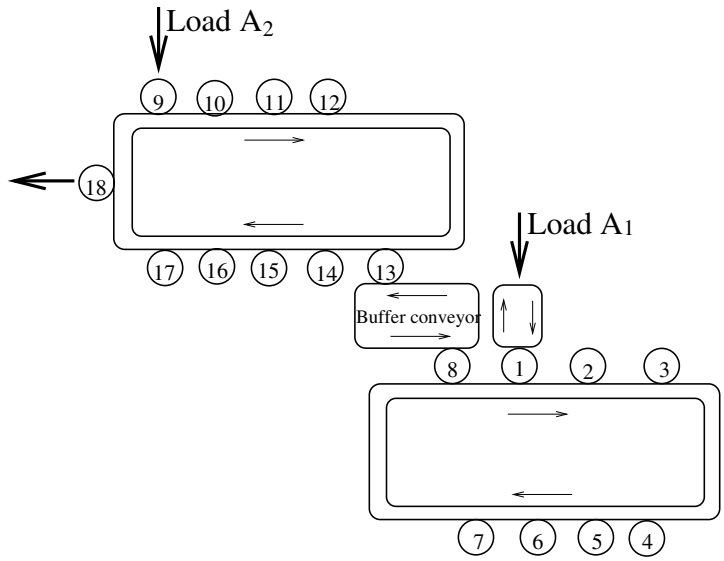


Figure 3.13: Layout of automotive ignition module assembly system

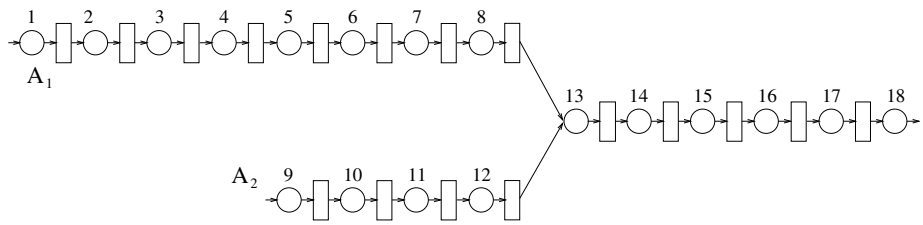


Figure 3.14: Structural model of the automotive ignition module assembly system of Figure 3.13

The situation with the system of Figure 3.15 is more complex. Here, 13 injection molding machines produce seven different part types necessary for the assembly. Which part is produced by a specific injection molding machine depends on scheduling. A physical model of this system is shown in Figure 3.16. To simplify it, we note that from the point of view of the in-process buffers, it is not important which particular machine is producing a specific part type at each time moment. What is important is the rate of parts flow into each buffer. Therefore, it is possible to substitute the 13 real machines by 7 *virtual* machines (see Figure 3.17), each producing a specific part type. Also, the additional processing operations can be aggregated into one assembly machine. If it is possible to calculate the parameters of the virtual machines,

based on the parameters of the real machines and scheduling procedures (which, in fact, can be done with a certain level of fidelity), then the production system of Figure 3.16 is reduced to a standard assembly system, shown in Figure 3.17.

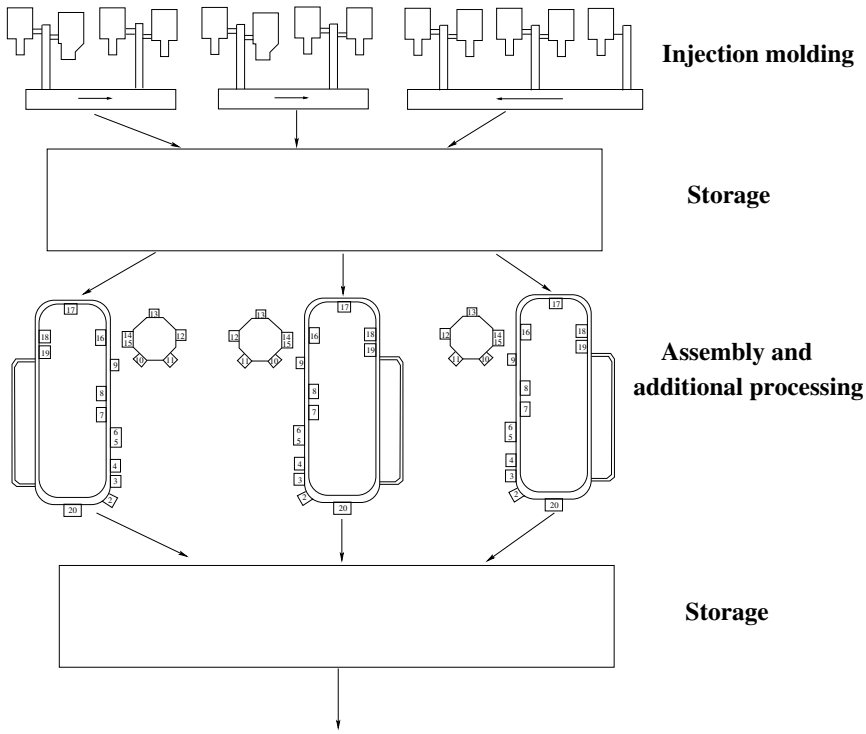


Figure 3.15: Layout of injection molding - assembly system

The development of a simple, but still relatively precise structural model of a production system, is one of the most important stages of production systems analysis, continuous improvement, and design. Since no formal methods for such modeling exist (or, perhaps, are even possible), production system engineers and managers must develop these skills through practical experience.

3.3 Mathematical Models of Machines

3.3.1 Timing issues

Cycle time (τ) – the time necessary to process a part by a machine. The cycle time may be constant, variable, or random. In large volume production systems, τ is practically always constant or close to being constant. This is the case in most production systems of the automotive, electronics, appliance, and other

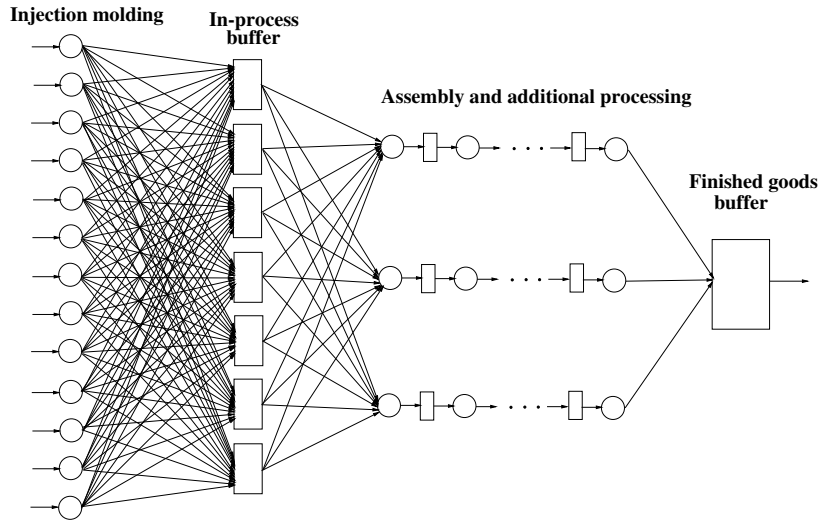


Figure 3.16: Physical model of injection molding - assembly system

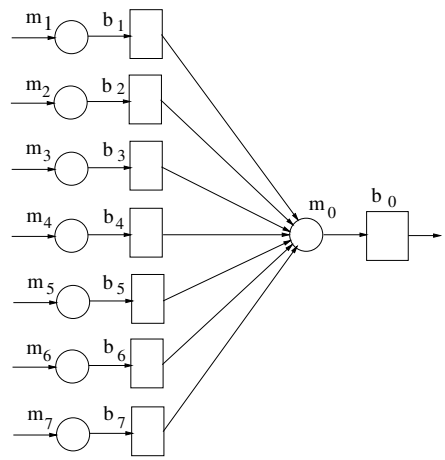


Figure 3.17: Structural model of injection molding - assembly system with virtual machines

industries. Variable or random τ takes place in job-shop environments where each part may have different processing specifications. In this book, we consider only machines with a *constant* cycle time; similar developments, however, can be carried out for the case of random (e.g., exponentially distributed) processing time.

Machine capacity (c) – the number of parts produced by a machine per unit of time when the machine is up. Clearly, in the case of constant τ,

$$c = \frac{1}{\tau}.$$

Machines in a production system may have identical or different cycle times. In the case of identical cycle time, the time axis may be considered as slotted or unslotted.

Slotted time – the time axis is slotted with the slot duration equal to the cycle time. In this case, all transitions – changes of machines' status (up or down) and changes of buffers' occupancy – are considered as taking place only at the beginning or the end of the time slot. Production systems satisfying this convention are called *synchronous*.

Unslotted or *continuous time* – the above mentioned changes may occur at any time moment. If the cycle times of all machines are identical, such a system, with a slight abuse of the definition, is still referred to as *synchronous*. If the cycle times are not identical, the system is called *asynchronous*.

Production systems with machines having different cycle times are typically considered as operating in unslotted time.

In the unslotted case, production systems can be conceptualized as discrete event systems or as flow systems.

Discrete event system – a job (i.e., part) is transferred from the producing machine to the subsequent buffer (if it is not full) only after the processing of the whole job is complete. In this case, the buffer occupancy is a non-negative integer.

Flow system – infinitesimal parts of the job are (conceptually) transferred from the producing machine to the subsequent buffer if it is not full. Similarly, an infinitesimal part of a job is taken by a downstream machine from the buffer, if the machine is not down and the buffer is not empty. In this case, there is a continuous flow of parts into and from the buffers. Clearly, the buffer occupancy in this situation is a non-negative real number.

Obviously, the discrete event conventions are closer to reality. However, flow systems are sometimes easier to analyze and often lead to reasonable conclusions. In this textbook both conventions are addressed.

3.3.2 Machine reliability models

Machine reliability model – the probability mass functions (pmf's) or the probability density functions (pdf's) of the up- and downtime of the machine in the slotted or unslotted time, respectively.

The reliability models considered in this book are listed below.

Reliability models for the slotted time case: Two models are addressed.

- *Bernoulli reliability model (B)* – at the beginning of each time slot, the status of the machine – up or down – is determined by a chance experiment, according to which it is up with probability p and down with probability $1-p$, independently of the status of this machine in all previous time slots.

This is the simplest reliability model. Indeed, first, it is static, i.e., no past status of the machine affects its status in the upcoming slot and, second, its pmf is very simple. Nevertheless, it is still practical, especially for describing assembly operations where the downtime is typically very short and comparable with the cycle time of the machine. Most of the analysis, continuous improvement, and design problems, considered in this book, are first solved using this simplest case and then extended to more complex scenarios.

- *Geometric reliability model (Geo)* – uptime and downtime pmf's are given by the geometric pmf's (2.14), (2.15), i.e.,

$$\begin{aligned} P_{t_{up}}(t) &= P[t_{up} = t] = P(1-P)^{t-1}, \quad t = 1, 2, \ldots, \\ P_{t_{down}}(t) &= P[t_{down} = t] = R(1-R)^{t-1}, \quad t = 1, 2, \ldots. \end{aligned} \tag{3.1}$$

As it is clear from the discussion in Chapter 2, these pmf's are generated by the transition diagram of Figure 2.7, which implies that the state of the machine (up or down) in any time slot is defined by that of the previous time slot with the probabilities of breakdown and repair P and R, respectively. Clearly, in this case the machine is a dynamic system with one step memory and, as it follows from Chapter 2, it can be described by a Markov chain. Methods of analysis of production systems with this reliability model are more complex than in the memoryless case. In comparison with the Bernoulli model, this is a more realistic description of a machine.

Production lines with Bernoulli and geometric reliability models are usually considered as discrete event systems, while all reliability models for the continuous time case (listed below) are viewed in the framework of flow systems.

Reliability models for the continuous time case: The continuous time case is, perhaps, more realistic than the slotted time and, therefore, a larger set of reliability models is addressed. They are as follows:

- *Exponential reliability model (exp)* – the uptime and downtime pdf's of the machine are given by the exponential distributions (2.27), (2.28), i.e.,

$$\begin{aligned} f_{t_{up}}(t) &= \lambda e^{-\lambda t}, \quad t \geq 0, \\ f_{t_{down}}(t) &= \mu e^{-\mu t}, \quad t \geq 0. \end{aligned} \tag{3.2}$$

The transition diagram, generating this reliability model, is shown in Figure 2.12, which implies that, if up, the machine may go down in each infinitesimal interval δt with rate λ and, if down, it may go up during δt with rate μ.

Clearly, this is also a dynamical system, and it can be described by a continuous time, discrete space Markov process. Methods of analysis of production systems with this reliability model are roughly of the same complexity as those for the geometric case and can be applied to the same technological operations. The main drawback of this model is that the breakdown and repair rates are constant, which is hardly true in reality. Therefore, other reliability models are introduced.

- *Rayleigh reliability model (Ra)* – the uptime and downtime pdf's of the machine are given by Rayleigh distributions (2.29), (2.30), i.e.,

$$\begin{aligned} f_{t_{up}}(t) &= \lambda t e^{-\frac{\lambda t^2}{2}}, \quad t \geq 0, \\ f_{t_{down}}(t) &= \mu t e^{-\frac{\mu t^2}{2}}, \quad t \geq 0. \end{aligned} \tag{3.3}$$

The breakdown and repair rates of a Rayleigh machine, as it has been shown in Chapter 2, are λt and μt respectively, implying that both are linearly increasing in time. The resulting system is, of course, still dynamic but it is not described by a Markov process anymore, and no rigorous analytical methods for analysis of such systems are available. In this book, we present *empirical methods* for analysis of this and other non-Markovian situations (see Chapter 12).

A deficiency of both exponential and Rayleigh reliability models is that their coefficients of variation cannot be placed at will – they are fixed, respectively, at 1 and 0.52, for all values of λ and μ. In reality, however, machines on the factory floor may have widely different coefficients of variation. The experimental evidence indicates, however, that in most cases they take values between 0 and 1. This necessitates considering other reliability models.

- *Weibull reliability model (W)* – the uptime and downtime pdf's of the machine are given by Weibull distributions (2.37), (2.38), i.e.,

$$\begin{aligned} f_{t_{up}}(t) &= \lambda^{\Lambda} e^{-(\lambda t)^{\Lambda}} \Lambda t^{\Lambda-1}, \quad t \geq 0, \\ f_{t_{down}}(t) &= \mu^{\mathrm{M}} e^{-(\mu t)^{\mathrm{M}}} \mathrm{M} t^{\mathrm{M}-1}, \quad t \geq 0. \end{aligned} \tag{3.4}$$

The Weibull distribution is commonly used in reliability theory. Being defined by two parameters, λ and Λ or μ and M, it allows us to place both its expected value and variance and, thus, the coefficient of variation, at will.

The next two reliability models, also with the ability of placing their CVs at will, are used for the sake of generality:

- *Gamma reliability model (ga)* – the uptime and downtime pdf's of the machine are given by the gamma distributions (2.35), (2.36), i.e.,

$$\begin{aligned} f_{t_{up}}(t) &= \lambda e^{-\lambda t} \frac{(\lambda t)^{\Lambda-1}}{\Gamma(\Lambda)}, \quad t \geq 0, \\ f_{t_{down}}(t) &= \mu e^{-\mu t} \frac{(\mu t)^{\mathrm{M}-1}}{\Gamma(\mathrm{M})}, \quad t \geq 0, \end{aligned} \tag{3.5}$$

where

$$\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds.$$

When Λ and M are positive integers, these distributions coincide with the Erlang distribution. Although there are some analytical methods for analysis of production systems with Erlang reliability models, they are so computationally intensive that the analysis of even two machine lines with buffer of capacity more than 5 is practically impossible. Therefore, these methods are not included in this textbook.

- *Log-normal reliability model (LN)* – the uptime and downtime pdf's of the machine are given by the log-normal distributions (2.39), (2.40), i.e.,

$$\begin{aligned} f_{t_{up}}(t) &= \frac{1}{\sqrt{2\pi}\Lambda t} e^{-\frac{(\ln t - \lambda)^2}{2\Lambda^2}}, \quad t \geq 0, \\ f_{t_{down}}(t) &= \frac{1}{\sqrt{2\pi}\mathrm{M}t} e^{-\frac{(\ln t - \mu)^2}{2\mathrm{M}^2}}, \quad t \geq 0. \end{aligned} \tag{3.6}$$

This model is considered since, unlike all others, it does not coincide with the exponential case when its $CV = 1$.

All reliability models described above imply that both up- and downtime are distributed according to the same type of pdf. In reality, of course, this is not necessarily the case. Therefore, we introduce

- *Mixed reliability model (M)* – the uptime and downtime are defined by different distributions selected from the set $\{exp, Ra, W, ga, LN\}$.

One more reliability model is considered in this book – for situations when no data on the type of the distributions involved is available:

- *General reliability model (G)* – the up- and downtime may be distributed according to arbitrary pdf's.

Clearly, rigorous analysis in this situation is impossible. However, in the framework of a number of problems addressed in this book, we show that the solutions practically do not depend on the distributions involved and are defined mostly by their first two moments. In this way, the results obtained for specific pdf's are extended to the general model of machine reliability.

Time-dependent vs. operation-dependent failures: As it was pointed out on a number of occasions above, machines in a production system can be blocked or starved. If a machine is, indeed, blocked or starved, it is forced down and does not perform its technological operation. In this situation, is its uptime "ticking" or not? The answer depends on the nature of the machine and its technological operation. For instance, tool wear does not occur when the machine is idle, while power failures are practically independent of the machine

status. To distinguish between these two situations, the following notions are introduced:

Operation-dependent failures – machine breakdowns cannot occur while it is blocked or starved.

Time-dependent failures – machine breakdowns may occur even while it is blocked or starved.

It turns out that all performance measures, considered in this book, take practically identical values under either time- or operation-dependent failures: in most cases the difference is within 1% - 3% (in particular, when buffers are not too small). Therefore, selection of a failure mode – time- or operation-dependent – can be made on the basis of convenience for subsequent analyses. Since, as it turns out, time-dependent failures are simpler for analysis, this is the convention considered throughout this book. Extensions to operation-dependent failures are possible.

3.3.3 Notations

In the slotted time case, each machine is denoted in this book by a pair

$$[P_{t_{up}}(t), P_{t_{down}}(t)],$$

where $P_{t_{up}}(t)$ and $P_{t_{down}}(t)$ are the pmf's of up- and downtime, respectively. For instance, a machine can be denoted as $[B_{up}, B_{down}]$ or $[Geo_{up}, Geo_{down}]$ or $[B_{up}, Geo_{down}]$.

Similarly, in the continuous time case, each machine is denoted as

$$[f_{t_{up}}(t), f_{t_{down}}(t)],$$

where the first symbol defines the uptime pdf and the second the downtime pdf, each belonging to the set $\{exp, Ra, W, ga, LN, G\}$. Examples of this notation can be given as $[exp_{up}, exp_{down}]$, $[W_{up}, ga_{down}]$, $[G_{up}, LN_{down}]$, etc.

Analogous notations are used for production systems consisting of several machines. For example, a serial line with M machines is denoted as

$$\{[f_{t_{up}}(t), f_{t_{down}}(t)]_1, [f_{t_{up}}(t), f_{t_{down}}(t)]_2, \ldots, [f_{t_{up}}(t), f_{t_{down}}(t)]_M\},$$

where $[f_{t_{up}}(t), f_{t_{down}}(t)]_i$ denotes the reliability model of the i-th machine in the system. If all machines in the system have identical reliability models, the serial line is denoted as

$$\{[f_{t_{up}}(t), f_{t_{down}}(t)]_i, \quad i = 1, \ldots, M\}.$$

Similar notations are used for production systems operating in slotted time.

The above notations are extended to machines with non-identical cycle times as follows: Each machine is denoted by a triple

$$[\tau, f_{t_{up}}(t), f_{t_{down}}(t)],$$

and a serial production line by the expression

$$\{[\tau, f_{t_{up}}(t), f_{t_{down}}(t)]_1, [\tau, f_{t_{up}}(t), f_{t_{down}}(t)]_2, \ldots, [\tau, f_{t_{up}}(t), f_{t_{down}}(t)]_M\}.$$

3.3.4 Machine model identification

Machine model identification – the process of determining the machine characteristics described above.

These characteristics include the machine cycle time, τ, and its reliability model, i.e., $P_{t_{up}}(t)$, $P_{t_{down}}(t)$ or $f_{t_{up}}(t)$, $f_{t_{down}}(t)$. The cycle time in most cases can be easily identified using a stop watch and measuring the time used by the machine to process a part. If manual loading and unloading operations take place, the loading and unloading time must be included in the machine cycle time. In some production systems, manual loading and unloading take more time than part processing. In these cases, the machine has a large cycle time even if its "own" (i.e., technological) cycle time is small.

Also, it should be pointed out that the "official" cycle times of the machines, recorded in an appropriate log, may be far from the real one. This happens because equipment maintenance personnel often makes adjustments, which modify the cycle time, without properly recording it or even realizing that a change has occurred. In the course of time, these changes are accumulated and lead to the situation mentioned above. Nevertheless, machine cycle time can be viewed as a rather "stable" or slowly changing machine characteristic. However, in every application it must be carefully measured.

The machine reliability model is much more difficult to identify. Strictly speaking, it requires the identification of the histograms of up- and downtime, which, in turn, require a very large number of measurements during a long period of time. The result is that the pmf's or pdf's of up- and downtime of the machines on the factory floor are, practically, **never** known. What is typically known is the average up- and downtime of the machines, T_{up} and T_{down}, often referred to as the mean time to failure (MTTF) and mean time to repair (MTTR), respectively. These characteristics are determined by measuring the durations of randomly occurring up- and downtime, $t_{up,i}$ and $t_{down,i}$ and then calculating their averages according to

$$T_{up} = \frac{\sum_{i=1}^{n} t_{up,i}}{n}, \tag{3.7}$$

$$T_{down} = \frac{\sum_{i=1}^{n} t_{down,i}}{n}, \tag{3.8}$$

where number n is sufficiently large to guarantee statistically reliable estimates (see Problem 2.8 of Chapter 2).

It is advisable to continually monitor T_{up} and T_{down} since they may (and often do) change in time. In this case, $T_{up}(s)$ and $T_{down}(s)$ are calculated, where $s = 1, 2, \ldots$, is the index of the period of observation, consisting of n occurrences of up- and downtime, i.e.,

$$T_{up}(s) = \frac{\sum_{i=1}^{n} t_{up,i}(s)}{n}, \tag{3.9}$$

$$T_{down}(s) = \frac{\sum_{i=1}^{n} t_{down,i}(s)}{n}, \tag{3.10}$$

$$s = 1, 2, \ldots.$$

If $T_{up}(s)$ and/or $T_{down}(s)$ exhibit undesirable trends (i.e., $T_{up}(s)$ decreasing and/or $T_{down}(s)$ increasing), appropriate actions must be taken, for instance, re-evaluation of preventative maintenance procedures.

As it will be shown in Parts II - IV, the knowledge of T_{up} and T_{down} is important and, moreover, necessary for the production line management technique referred to as *measurement-based management* (MBM).

It should be pointed out that values of T_{up} and T_{down} are often available from equipment manufacturers. However, they may be quite different from their real values. The same can be said for some operator log data. Therefore, for any analysis, design, and continuous improvement project, it would be prudent to accurately identify T_{up} and T_{down}.

As it is shown in Parts II - IV, MBM requires the knowledge of not only T_{up} and T_{down} but also the coefficients of variation, CV_{up} and CV_{down}. Fortunately, they can be calculated based on the measurements, which are used to calculate T_{up} and T_{down}. Indeed, since

$$\begin{aligned} Var(t_{up}) &= \frac{\sum_{i=1}^{n}(t_{up,i} - T_{up})^2}{n-1}, \\ Var(t_{down}) &= \frac{\sum_{i=1}^{n}(t_{down,i} - T_{down})^2}{n-1}, \end{aligned}$$

the CVs can be calculated as

$$CV_{up} = \frac{\sqrt{Var(t_{up})}}{T_{up}}, \tag{3.11}$$

$$CV_{down} = \frac{\sqrt{Var(t_{down})}}{T_{down}}. \tag{3.12}$$

If continuous monitoring of up- and downtime takes place, $CV_{up}(s)$ and $CV_{down}(s)$ may be calculated as follows:

$$CV_{up}(s) = \frac{\sqrt{Var(t_{up}(s))}}{T_{up}(s)}, \quad s = 1, 2, \ldots, \tag{3.13}$$

$$CV_{down}(s) = \frac{\sqrt{Var(t_{down}(s))}}{T_{down}(s)}, \quad s = 1, 2, \ldots. \tag{3.14}$$

Realistically speaking, T_{up}, T_{down} and CV_{up}, CV_{down} (or $T_{up}(s)$, $T_{down}(s)$ and $CV_{up}(s)$, $CV_{down}(s)$) may be the only characteristics of reliability models available from the factory floor. Fortunately, they are also *sufficient*, as it is shown in Parts II - IV, to solve most of the analysis, continuous improvement, and design problems of practical importance, while the knowledge of T_{up} and T_{down} alone is *not sufficient* for these purposes.

Finally, it should be pointed out that even when T_{up}, T_{down} and CV_{up}, CV_{down} (or $T_{up}(s)$, $T_{down}(s)$ and $CV_{up}(s)$, $CV_{down}(s)$) are available, in most cases it cannot be assumed that these values are, in fact, precise. Experience shows that in most realistic cases, these data may have up to 5% - 10% errors

as compared with their real values. This discrepancy is largely due to the fact that machine characteristics are changing in time and the measurement process of the realizations $t_{up,i}$ and $t_{down,i}$ is not fault free.

3.3.5 Calculating parameters of aggregated machines

As it was mentioned in Section 3.2, structural modeling of a production system may require "combining" parallel machines (as in Figure 3.10) or consecutive dependent machines (as in Figure 3.11) into one aggregated machine (as in Figure 3.12). Below, methods for calculating parameters of aggregated machines are given.

Aggregating parallel machines: This aggregation process is illustrated in Figure 3.18. Below, we provide expressions for the parameters of the aggregated machine m_{agg}^{par}, i.e., $\{\tau_{agg}^{par}, T_{up,agg}^{par}, T_{down,agg}^{par}\}$.

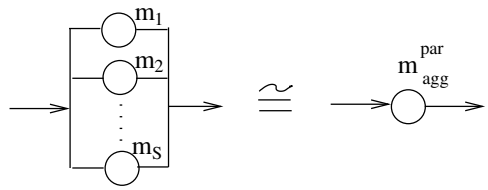


Figure 3.18: Aggregating parallel machines

(a) *Identical machines.* Assume that each machine m_i, $i = 1, \ldots, S$, is characterized by $\tau_i = \tau = 1/c$, $T_{up,i} = T_{up}$, and $T_{down,i} = T_{down}$. Then c_{agg}^{par} is selected as

$$c_{agg}^{par} = Sc,$$

which implies that

$$\tau_{agg}^{par} = \frac{1}{c_{agg}^{par}} = \frac{\tau}{S}. \tag{3.15}$$

The average uptime and downtime of the aggregated machine are selected as

$$T_{up,agg}^{par} = T_{up}, \tag{3.16}$$
$$T_{down,agg}^{par} = T_{down}. \tag{3.17}$$

When machines m_i, $i = 1, \ldots, S$, are exponential with parameters λ and μ, these expressions become

$$\lambda_{agg}^{par} = \lambda, \tag{3.18}$$
$$\mu_{agg}^{par} = \mu. \tag{3.19}$$

(b) *Nonidentical machines.* Assume that each machine m_i, $i = 1, \ldots, S$, is characterized by $\tau_i = 1/c_i$, $T_{up,i}$, and $T_{down,i}$. Then c_{agg}^{par} is selected as

$$c_{agg}^{par} = \sum_{i=1}^{S} c_i,$$

i.e.,

$$\tau_{agg}^{par} = \frac{1}{\sum_{i=1}^{S} \frac{1}{\tau_i}}. \tag{3.20}$$

The average up- and downtime of the aggregated machine are selected based on the following considerations. The parallel system has the capacity c_{agg}^{par} only when all S machines are up; when less than S machines are up, the capacity is lower. To model this situation, we assume that the capacity of the aggregated machine remains the same while its uptime is reduced appropriately. For instance, in the case $S = 2$, the following four events can occur: both machines are up, both are down, the first is up and the second down, and the first is down and the second is up. Then, $T_{up,agg}^{par}$ can be defined as:

$$T_{up,agg}^{par} = \frac{T_{up,1}T_{up,2} + \frac{\tau_2}{\tau_1+\tau_2}T_{up,1}T_{down,2} + \frac{\tau_1}{\tau_1+\tau_2}T_{up,2}T_{down,1}}{\frac{1}{2}(T_{up,1} + T_{down,1} + T_{up,2} + T_{down,2})}.$$

Here, the first term in the numerator indicates the case when both machines are up, and the latter two terms refer to the situation where one machine is up and the other down, while the denominator is the average (up+down)-time of the machines. Given this, $T_{down,agg}^{par}$ is selected so that the throughput of the aggregated machine is the same as that of the parallel system, i.e.,

$$TP_{agg}^{par} := c_{agg}^{par} \frac{T_{up,agg}^{par}}{T_{up,agg}^{par} + T_{down,agg}^{par}} = \sum_{i=1}^{2} c_i \frac{T_{up,i}}{T_{up,i} + T_{down,i}}.$$

In other words,

$$T_{down,agg}^{par} = \frac{T_{down,1}T_{down,2} + \frac{\tau_2}{\tau_1+\tau_2}T_{down,1}T_{up,2} + \frac{\tau_1}{\tau_1+\tau_2}T_{down,2}T_{up,1}}{\frac{1}{2}(T_{up,1} + T_{down,1} + T_{up,2} + T_{down,2})}.$$

In the case $S > 2$, these arguments lead to the following expressions:

$$T_{up,agg}^{par} = \frac{\tau_{agg}^{par} \sum_{i=1}^{S} \left[\frac{1}{\tau_i T_{down,i}} \prod_{j=1,j\neq i}^{S} (\frac{1}{T_{up,j}} + \frac{1}{T_{down,j}}) \right]}{\frac{1}{S} \sum_{i=1}^{S} \left[\frac{1}{T_{up,i} T_{down,i}} \prod_{j=1,j\neq i}^{S} (\frac{1}{T_{up,j}} + \frac{1}{T_{down,j}}) \right]}, \tag{3.21}$$

$$T_{down,agg}^{par} = \frac{\tau_{agg}^{par} \sum_{i=1}^{S} \left[\frac{1}{\tau_i T_{up,i}} \prod_{j=1,j\neq i}^{S} (\frac{1}{T_{up,j}} + \frac{1}{T_{down,j}}) \right]}{\frac{1}{S} \sum_{i=1}^{S} \left[\frac{1}{T_{up,i} T_{down,i}} \prod_{j=1,j\neq i}^{S} (\frac{1}{T_{up,j}} + \frac{1}{T_{down,j}}) \right]}. \tag{3.22}$$

If all S machines are exponential with parameters λ_i and μ_i, these expressions become

$$\lambda_{agg}^{par} = \frac{\sum_{i=1}^{S}\left[\lambda_i\mu_i\prod_{j=1,j\neq i}^{S}(\lambda_j+\mu_j)\right]}{S\tau_{agg}^{par}\sum_{i=1}^{S}\left[\frac{1}{\tau_i}\mu_i\prod_{j=1,j\neq i}^{S}(\lambda_j+\mu_j)\right]}, \tag{3.23}$$

$$\mu_{agg}^{par} = \frac{\sum_{i=1}^{S}\left[\lambda_i\mu_i\prod_{j=1,j\neq i}^{S}(\lambda_j+\mu_j)\right]}{S\tau_{agg}^{par}\sum_{i=1}^{S}\left[\frac{1}{\tau_i}\lambda_i\prod_{j=1,j\neq i}^{S}(\lambda_j+\mu_j)\right]}. \tag{3.24}$$

Clearly, the above expressions are only approximations of the real system. However, in most cases, they lead to accurate estimates of the original system performance. For instance, consider a system shown in Figure 3.19 and assume that the machines are exponential with

$$\begin{aligned}\tau &= \{1.2, 2, 2.6, 1, 3.4, 2.5, 3, 1.4\},\\ T_{up} &= \{100, 33.3, 52.6316, 20, 50, 100, 33.3, 100\},\\ T_{down} &= \{10, 5, 6.6667, 8.3333, 10, 5.8824, 4, 12.5\},\end{aligned}$$

i.e.,

$$e = \{0.91, 0.87, 0.89, 0.71, 0.83, 0.94, 0.89, 0.89\}.$$

Then the exponential machines of the aggregated system, shown in Figure 3.20,

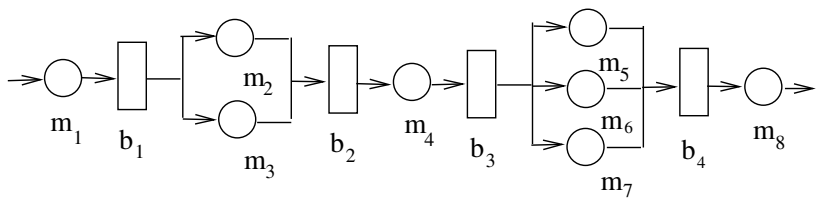


Figure 3.19: Serial production line with parallel machines

are defined by the following parameters:

$$\begin{aligned}\tau_{agg}^{par} &= \{1.2, 1.13, 1, 0.97, 1.4\},\\ T_{up,agg}^{par} &= \{100, 40.8558, 20, 50.8103, 100\},\\ T_{down,agg}^{par} &= \{10, 5.7091, 8.3333, 5.9039, 12.5\},\end{aligned}$$

i.e.,

$$e_{agg}^{par} = \{0.9091, 0.8774, 0.7059, 0.8959, 0.8889\}.$$

The throughput of these two lines, TP and TP_{agg}, has been evaluated by simulations for buffer capacities varied from $N = 5$ to $N = 100$. The accuracy of the aggregation has been quantified by

$$\epsilon_{TP} = \frac{TP - TP_{agg}}{TP} \cdot 100\%.$$

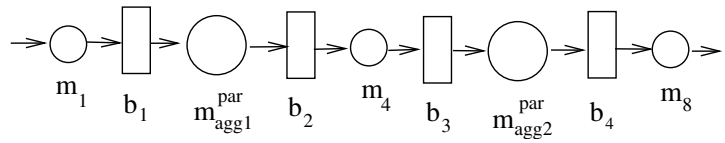


Figure 3.20: Aggregated version of the serial line of Figure 3.19

The results, along with 95% confidence intervals, are given in Table 3.1 and illustrated graphically in Figure 3.21. Clearly, the accuracy is sufficiently high, especially when N can accommodate at least one downtime of the machines.

Table 3.1: Accuracy of aggregation of parallel machines

N	TP	TP_{agg}	ϵ_{TP}
5	0.5388 ± 0.0002354	0.5105 ± 0.0002409	-5.25%
10	0.5874 ± 0.0002444	0.5699 ± 0.0002215	-2.98%
20	0.6202 ± 0.0001873	0.6138 ± 0.0002238	-1.03%
30	0.6292 ± 0.0002383	0.6272 ± 0.0002062	-0.32%
40	0.6326 ± 0.0001769	0.6318 ± 0.0002004	-0.13%
50	0.6339 ± 0.0002363	0.6332 ± 0.0002182	-0.11%
100	0.6347 ± 0.0001998	0.6346 ± 0.0002008	-0.02%

Aggregating consecutive dependent machines: This aggregation is illustrated in Figure 3.22.

(a) *Identical machines.* Again, it is first assumed that machines are identical and each machine, m_i, $i = 1, \ldots, S$, is characterized by $\{\tau, T_{up}, T_{down}\}$. The parameters of the aggregated machine m_{agg}^{con}, i.e., $\{\tau_{agg}^{con}, T_{up,agg}^{con}, T_{down,agg}^{con}\}$, are selected as follows:

$$\tau_{agg}^{con} = \tau, \tag{3.25}$$

$$T_{up,agg}^{con} = \Big(\frac{T_{up}}{T_{up}+T_{down}}\Big)^{S-1} T_{up}, \tag{3.26}$$

$$T_{down,agg}^{con} = \Big[1-\Big(\frac{T_{up}}{T_{up}+T_{down}}\Big)^{S}\Big](T_{up}+T_{down}). \tag{3.27}$$

In the case of exponential machines with parameters λ and μ, we obtain

$$\lambda_{agg}^{con} = \frac{\lambda(\lambda+\mu)^{S-1}}{\mu^{S-1}},$$

$$\mu_{agg}^{con} = \frac{\lambda\mu}{(\lambda+\mu)\Big[1-\Big(\frac{\mu}{\lambda+\mu}\Big)^{S}\Big]}.$$

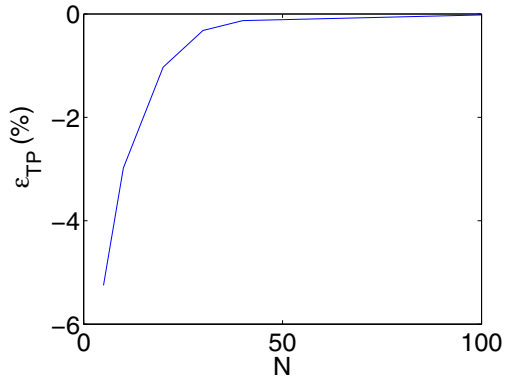


Figure 3.21: Illustration of accuracy of parallel machines aggregation

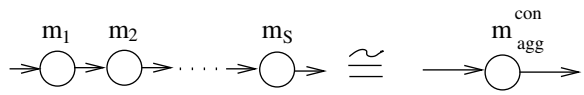


Figure 3.22: Aggregating consecutive machines

(b) *Nonidentical machines.* Assume each machine m_i, $i = 1, \ldots, S$, is characterized by $\{\tau_i, T_{up,i}, T_{down,i}\}$. The cycle time of the aggregated machine m_{agg}^{con}, i.e., τ_{agg}^{con}, is selected as follows:

$$\tau_{agg}^{con} = \max_i \tau_i. \tag{3.28}$$

The average up- and downtime of the aggregated machine are selected based on the following considerations. The consecutive dependent system is working only when all S machines are up. For example, when $S = 2$, $T_{up,agg}^{con}$ can be understood as the time that both machines are up within an average (up+down)-time of the machines, i.e.,

$$T_{up,agg}^{con} = \frac{1}{2}(T_{up,1} + T_{down,1} + T_{up,2} + T_{down,2}) \frac{T_{up,1}}{T_{up,1} + T_{down,1}} \cdot \frac{T_{up,2}}{T_{up,2} + T_{down,2}}.$$

Given this, $T_{down,agg}^{con}$ is again selected so that the throughput of the aggregated machine is the same as that of the consecutive dependent system, i.e.,

$$e_{agg}^{con} := \frac{T_{up,agg}^{con}}{T_{up,agg}^{con} + T_{down,agg}^{con}} = \frac{T_{up,1}}{T_{up,1} + T_{down,1}} \cdot \frac{T_{up,2}}{T_{up,2} + T_{down,2}}.$$

In other words,

$$T_{down,agg}^{con} = \frac{1}{2}(T_{up,1}+T_{down,1}+T_{up,2}+T_{down,2}) \Big(1 - \frac{T_{up,1}T_{up,2}}{(T_{up,1} + T_{down,1})(T_{up,2} + T_{down,2})}\Big).$$

In the case $S > 2$, these arguments lead to the following expressions:

$$T^{con}_{up,agg} = \frac{1}{S}\sum_{i=1}^{S}(T_{up,i}+T_{down,i})\prod_{i=1}^{S}\Big(\frac{T_{up,i}}{T_{up,i}+T_{down,i}}\Big), \tag{3.29}$$

$$T^{con}_{down,agg} = \frac{1}{S}\sum_{i=1}^{S}(T_{up,i}+T_{down,i})\Big[1-\prod_{i=1}^{S}\Big(\frac{T_{up,i}}{T_{up,i}+T_{down,i}}\Big)\Big].. \tag{3.30}$$

If all S machines are exponential with parameters λ_i and μ_i, these expressions become

$$\lambda^{con}_{agg} = \frac{S}{\sum_{i=1}^{S}\Big(\frac{1}{\lambda_i}+\frac{1}{\mu_i}\Big)\prod_{i=1}^{S}\frac{\mu_i}{\lambda_i+\mu_i}}, \tag{3.31}$$

$$\mu^{con}_{agg} = \frac{S}{\sum_{i=1}^{S}\Big(\frac{1}{\lambda_i}+\frac{1}{\mu_i}\Big)\Big[1-\prod_{i=1}^{S}\Big(\frac{\mu_i}{\lambda_i+\mu_i}\Big)\Big]}. \tag{3.32}$$

The accuracy of this aggregation is of the same order of magnitude as that for parallel machines.

PSE Toolbox: The process of aggregating parallel and consecutive dependent machines is implemented as the first two tools of the toolbox function **Modeling**. For a description and illustration of these tools, see Section 19.2.

3.3.6 Machine quality models

In some manufacturing operations, machines can produce defective parts, along with non-defective ones. To formalize this situation, we introduce *machine quality models* – the pmf or pdf of time intervals during which the machine produces good or defective parts. Examples of quality models are listed below:

- *Bernoulli quality model* - each part produced during a cycle time is good with probability g and defective with probability $1-g$, independent of the quality of parts produced during previous cycles.
- *Exponential quality model* - when up, the intervals of time during which a machine produces good or defective parts are distributed exponentially with parameters γ and β, respectively.
- *General quality model* - when up, the intervals of time during which a machine produces good or defective parts are distributed according to arbitrary pdf's.

The Bernoulli quality model is appropriate when the defects are due to independent random events (for instance, dust or scratches in automotive painting operations). The exponential and general quality models are appropriate when the defects are due to deterioration in the machine operation, such as wearing of the cutting tools or vibration of the workpiece.

3.4 Mathematical Models of Buffers

3.4.1 Modeling

For the purposes of this textbook, the mathematical model of a buffer is very simple – it is its capacity, N, i.e., the maximum number of parts that it can store. It is assumed throughout that

$$N < \infty,$$

implying that buffers are finite.

The number of parts contained in a buffer at a given time is referred to as its *occupancy*. Since in a production system, the occupancy of a buffer at a given time (slot or moment) depends on its occupancy at the previous time (slot or moment), buffers are dynamical systems with the occupancy being their states. If the machines are modeled as discrete event systems, the state of the buffer is an integer between 0 and N. In flow models, states are real numbers between 0 and N.

It is assumed that a part, produced by a machine, is immediately placed in the downstream buffer, if it is not full. Similarly, it is assumed that a part is immediately available for processing by a machine, if the upstream buffer is not empty. Although these assumptions are introduced to simplify the analysis, most production systems are designed and operated so that they hold. In some cases, where they do not hold but the times for placing a part in the buffer and transferring a part to the subsequent machine are relatively constant, this assumption might be "compensated" by an appropriate increase of the machine cycle time. In any case, experience shows that these assumptions do not lead to erroneous results as far as steady state performance measures are concerned.

In some cases, buffers are relatively complex material handling devices, e.g., robots, automated guided vehicles, etc., and, therefore, may experience breakdowns in the same manner as machines do. In these cases, a material handling device can be modeled as a machine followed by a buffer, as illustrated in Figure 3.23.

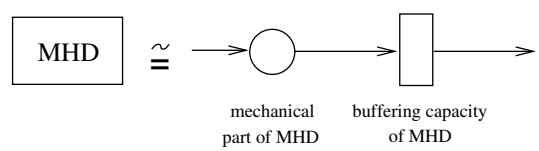


Figure 3.23: Modeling material handling devices with breakdowns

3.4.2 Buffer parameters identification

Buffer model identification – the process of determining the capacity of the buffer, N.

As it has been mentioned above, buffers on the factory floor may take the form of boxes, which store the work-in-process, or conveyors, or silos, or robotic storage devices, or automated guided vehicles. In most cases, their capacity can be relatively easily identified. For instance, in a kanban system, N is determined by the number of kanban cards between each pair of consecutive machines. In robotic storage devices, the capacity also can be evaluated through the analysis of the storage spaces available and part size. In the case of conveyors or similar material handling devices (e.g., silos), where the parts are transported from one operation to another on carriers, the identification of buffer capacity is carried out as follows:

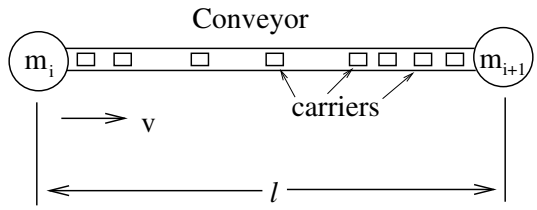


Figure 3.24: Conveyor material handling

Assume that the length of the conveyor between machines m_i and m_{i+1} is l, its speed is v, and the cycle time of the machines is τ (see Figure 3.24). Then the time to travel from m_i to m_{i+1} is

$$T_{travel} = \frac{l}{v},$$

and the number of carriers necessary to sustain continuous operation is

$$N_0 = \left\lceil \frac{T_{travel}}{\tau} \right\rceil,$$

where $\lceil x \rceil$ is the smallest integer larger than or equal to x.

If only N_0 carriers are available between m_i and m_{i+1}, this conveyor provides no buffering capabilities. However, if there is space for $K_i > N_0$ carriers on the conveyor between m_i and m_{i+1}, the buffer capacity is

$$N_i = K_i - N_0.$$

Similar arguments can be used for identification of buffering capacity of other material handling devices when the cycle times of the machines are not identical.

3.5 Modeling Interactions between Machines and Buffers

This section defines conventions, which specify how the states of the machines and buffers affect each other. The purpose of these conventions is to ensure uniqueness of the mathematical description of the production systems at hand.

3.5.1 Slotted time case

State changing convention: *Machine state* (or status in the Bernoulli reliability case) is determined at *the beginning of each time slot*. This implies that a chance experiment, carried out at the beginning of each time slot, determines whether the machine is up or down during this time slot.

Buffer state is determined at *the end of each time slot*. This implies that buffer occupancy may change only at the end of a time slot. For instance, if the state of the buffer was 0 at the end of the previous time slot, then the downstream machine does not produce a part during the subsequent time slot, even if it is up. If the buffer state was $h \neq 0$ and the downstream machine was up and not blocked, then the state of the buffer at the end of the next time slot is h if the upstream machine produces a part, or $h-1$ if the upstream machine fails to produce.

Blocking and starvation conventions: *Blocked before service* (BBS) – a machine cannot operate if it is up (as determined at the beginning of the time slot), the downstream buffer is full (as determined at the end of the previous time slot), and the downstream machine does not take a part from this buffer at the beginning of this time slot. In other words, the part, which is being processed by a machine, is viewed as being already in the subsequent buffer.

Blocked after service (BAS) – if the machine is up (as determined at the beginning of the time slot) and the upstream buffer is not empty (as determined at the end of the previous time slot), the machine operates on the part, even if the downstream buffer is full. The machine becomes blocked if the same conditions persist during the next time slot.

Thus, the capacity of buffers under BBS and BAS conventions are related as

$$N_i^{BBS} = N_i^{BAS} + 1, \qquad i = 1, \ldots, M-1.$$

In other words, while $N_i^{BAS} \geq 0$, $N_i^{BBS} \geq 1$. All performance measures under these conventions are quite similar. In this textbook, we use the blocked before service convention, since it leads to a simpler description.

As far as *starvations* are concerned, it is assumed that a machine is starved during a time slot if it is up (as determined at the beginning of this time slot) and the upstream buffer is empty (as determined at the end of the previous time slot).

3.5.2 Continuous time case

In analytical investigations of flow models, i.e., when state transitions of the machines and buffers may take place at any time moment, the state changing conventions (either before or after service) do not affect system analysis or performance. In numerical investigations of flow models, the slotted time conventions are used with slot duration $\delta t \ll \tau$.

3.6 Performance Measures

The mathematical models of machines and buffers, described above, are necessary, in particular, for calculating *performance measures* of production systems at hand. Although some of them have been mentioned in Chapters 1 and 2, below we provide formal definitions.

3.6.1 Production rate and throughput

Production rate (PR) – average number of parts produced by the last machine of a production system per cycle time in the steady state of system operation.

This metric is appropriate for production systems with all machines having identical cycle times. In the asynchronous case, this metric is referred to as

Throughput (TP) – average number of parts produced by the last machine of a productions system per unit of time in the steady state of system operation.

Clearly, TP can be used in the synchronous case as well; in this case,

$$TP = c \cdot PR,$$

where c is the machine capacity.

Since PR and TP are steady state performance measures, due to conservation of flow, the same average number of parts is produced by any machine in the system.

Both PR and TP are functions of machine and buffer parameters. For instance, in the case of a serial line with M Bernoulli machines, i.e.,

$$\{[B_{up}, B_{down}]_1, \ldots, [B_{up}, B_{down}]_M\},$$

this function can be denoted as

$$PR = PR(p_1, \ldots, p_M, N_1, \ldots, N_{M-1}).$$

In the case of

$$\{[exp_{up}, exp_{down}]_1, \ldots, [exp_{up}, exp_{down}]_M\},$$

this function becomes

$$PR = PR(\lambda_1, \mu_1, \ldots, \lambda_M, \mu_M, N_1, \ldots, N_{M-1}).$$

In the asynchronous exponential case, this function is

$$TP = TP(\tau_1, \lambda_1, \mu_1, \ldots, \tau_M, \lambda_M, \mu_M, N_1, \ldots, N_{M-1}).$$

Due to the complex nature of interference among the machines and buffers (through blockages and starvations), these functions are extremely complicated, and their closed-form expressions are all but impossible to derive, except for the case of two-machine systems. For $M > 2$, several approximation techniques, based on decomposition or aggregation of longer lines into two-machine systems,

have been developed. Specific aggregation techniques, used in this book, are described in Parts II - IV.

The situation becomes even more complex for systems with non-Markovian machines. For instance, if the machines are characterized by the Weibull reliability model (3.4), the production rate has the form

$$PR = PR(\lambda_1, \Lambda_1, \mu_1, \mathrm{M}_1, \ldots, \lambda_M, \Lambda_M, \mu_M, \mathrm{M}_M, N_1, \ldots, N_{M-1}),$$

and this function cannot be expressed in closed form even for $M = 2$. Therefore, the approach to non-Markovian systems, used in this book, is based on empirical expressions, derived through extensive experimentation with systems at hand (see Part III).

All of the above remarks remain valid in the case of more complex serial lines and assembly systems. For example, in the closed serial line of Figure 3.3 with Bernoulli machines, this function is

$$PR = PR(p_1, \ldots, p_M, N_1, \ldots, N_{M-1}, S, N_0),$$

where S is the number of carriers and N_0 is the capacity of the empty carrier buffer. For the assembly system of Figure 1.2 with Bernoulli machines,

$$\begin{aligned} PR \quad = \quad & PR(p_{11}, \ldots, p_{1M_1}, N_{11}, \ldots, N_{1M_1}; p_{21}, \ldots, p_{2M_2}, N_{21}, \ldots, N_{M_2}; \\ & p_{01}, \ldots, p_{0M_0}, N_{01}, \ldots, N_{0(M_0-1)}). \end{aligned} \tag{3.33}$$

In production lines that include non-perfect quality machines, PR denotes the production rate of non-defective parts. In this case, for the Bernoulli reliability and quality models,

$$PR = PR(p_1, \ldots, p_M, g_1, \ldots, g_M, N_1, \ldots, N_{M-1}, S, N_0).$$

The average number of parts used by the first machine per cycle time is another performance measure, referred to as the *consumption rate* (CR). The difference between CR and PR is the *scrap rate* (SR).

3.6.2 Work-in-process and finished goods inventory

Work-in-process of the i-th buffer (WIP_i) – average number of parts contained in the i-th in-process buffer of a production system in the steady state of its operation.

Total work-in-process (WIP) – average number of parts contained in all in-process buffers of a production system in the steady state of its operation. For instance, for serial lines with M machines,

$$WIP = \sum_{i=1}^{M-1} WIP_i.$$

Finished goods inventory (FGI) – average number of parts contained in the finished goods buffer of a production system in the steady state of its operation.

Clearly, each of these performance measures is a function of system parameters. For instance WIP_i for the line

$$\{[B_{up}, B_{down}]_1, \ldots, [B_{up}, B_{down}]_M\},$$

can be denoted as

$$WIP_i = WIP_i(p_1, \ldots, p_M, N_1, \ldots, N_{M-1}).$$

For the line

$$\{[exp_{up}, exp_{down}]_1, \ldots, [exp_{up}, exp_{down}]_M\},$$

this function is

$$WIP_i = WIP_i(\lambda_1, \mu_1, \ldots, \lambda_M, \mu_M, N_1, \ldots, N_{M-1}).$$

As in the case of the production rate, there are no closed-form expressions for these functions, except for $M = 2$ and machines with Bernoulli, geometric, or exponential reliability models. Again, for $M > 2$, aggregation techniques are available (see Parts II - IV).

A similar situation takes place in the case of the finished goods inventory. The function of interest here is

$$FGI = FGI(\lambda_1, \mu_1, \ldots, \lambda_M, \mu_M, N_1, \ldots, N_{M-1} N_{FGB}),$$

which corresponds to serial lines with exponential machines. Unfortunately, there are no closed-form expressions for this performance measure, even for $M = 2$, and only approximations are available.

3.6.3 Probabilities of blockages and starvations

While PR and WIP are performance measures widely used in manufacturing, both in practice and theory, the metrics discussed in this subsection are used quite rarely, at least in practice. Nevertheless, as it is shown in this textbook, they play a central role in measurement-based management of production systems.

While T_{up}'s and T_{down}'s characterize the reliability of the machines in isolation, and N's characterize the storing efficacy of the buffers, *the blockages and starvations characterize the production systems as a whole*, i.e., machines and buffers placed in specific positions within a production system. As a result, these metrics, as shown in Parts II - IV, define the system bottlenecks and other system-wide characteristics. Consequently, they may and should be used for managing production systems. While the discussion of measurement-based management is included in Parts II - IV, below we define these performance metrics (for the blocked before service convention).

Blockage of machine i (BL_i) – steady state probability that machine i is up, buffer i is full, and machine $i + 1$ does not take a part from the buffer.

Starvation of machine i (ST_i) – steady state probability that machine i is up and buffer $i - 1$ is empty.

For the case of serial lines in slotted time, these performance measures can be expressed as

$$\begin{aligned} BL_i &= P[\{m_i \text{ is up at the beginning of the time slot}\} \cap \{b_i \text{ is full at the end of the previous time slot}\} \cap \{m_{i+1} \text{ does not take a part from } b_i \text{ at the beginning of the time slot}\}], \quad i = 1, \ldots, M-1, \\ ST_i &= P[\{m_i \text{ is up at the beginning of the time slot}\} \cap \{b_{i-1} \text{ is empty at the end of the previous time slot}\}], \quad i = 2, \ldots, M. \end{aligned}$$

For the case of serial lines in continuous time, these performance measures can be expressed as

$$\begin{aligned} BL_i &= P[\{m_i \text{ is up at time } t\} \cap \{b_i \text{ is full at time } t\} \cap \{m_{i+1} \text{ does not take material from } b_i \text{ at time } t\}], \quad i = 1, \ldots, M-1, \\ ST_i &= P[\{m_i \text{ is up at time } t\} \cap \{b_{i-1} \text{ is empty at time } t\}], \quad i = 2, \ldots, M. \end{aligned}$$

It is typically assumed that m_1 is never starved and m_M is never blocked.

Each of the above probabilities are functions of the machine and buffer parameters and their locations within the system. For instance, for $\{[exp_{up}, exp_{down}]_1, \ldots, [exp_{up}, exp_{down}]_M\}$ these functions become

$$\begin{aligned} BL_i &= BL_i(\lambda_1, \mu_1, \ldots, \lambda_M, \mu_M, N_1, \ldots, N_{M-1}), \\ ST_i &= ST_i(\lambda_1, \mu_1, \ldots, \lambda_M, \mu_M, N_1, \ldots, N_{M-1}). \end{aligned}$$

Again, closed-form expressions for these performance metrics can be derived only for $M = 2$ and Bernoulli, geometric, or exponential machines. For $M > 2$, the aggregation procedures mentioned above lead to estimates of these quantities (see Parts II - IV).

3.6.4 Residence time

Residence time (RT) – average time a part spends in the system in the steady state of its operation.

As it follows from the Little's Law, RT can be easily evaluated, if PR (or TP) and WIP are known. Indeed,

$$RT = \frac{WIP}{PR}[\text{cycle time}] \quad \text{or} \quad RT = \frac{WIP}{TP}[\text{units of time}].$$

The residence time is important for quoting product delivery time to the customer. Note that RT can be quite large even if the total processing time, i.e., $\sum_{i-=1}^{M} \tau_i$, is small. Note also that in industry RT is sometimes referred to as the flow time or the system cycle time.

3.6.5 Due-time performance

Due-time performance (DTP) – steady state probability to ship to the customer the desired number of parts during a given time period.

Assume that the customer requires D parts to be shipped during each shipping period T. Then

$$DTP = P[\{\text{ship to the customer } D \text{ parts every shipping period } T\}].$$

This performance metric is typically used for production systems with finished goods buffers (FGB). For instance, for the line $\{[exp_{up}, exp_{down}]_1, \ldots, [exp_{up}, exp_{down}]_M\}$ with a FGB of capacity N_{FGB} this function can be denoted as

$$DTP = DTP(\lambda_1, \mu_1, \ldots, \lambda_M, \mu_M, N_1, \ldots, N_{M-1}, N_{FGB}, D, T).$$

Analytical expressions for this function are derived only for a single-machine production system, and lower bounds are available for $M > 1$ (see Parts II - IV).

3.6.6 Transient characteristics

Transient characteristics – a group of performance measures that describe how PR, WIP, and the probabilities of buffer occupancy reach their steady state values.

These properties are described in terms of several metrics. Perhaps, the most important of them is the *settling time*, t_s, which is the time necessary to reach and remain in a $\pm 5\%$ neighborhood of the steady state, starting from zero initial conditions (i.e., all empty buffers). Thus, in this textbook, we characterize the transient behavior by t_{sPR} and t_{sWIP}, (i.e., the settling times of PR and WIP, respectively). In addition, we investigate the properties of the second largest eigenvalue of the transition matrix of Markov chains that characterize production systems at hand, which, as it turns out, defines the behavior of t_{sPR} and t_{sWIP}.

3.6.7 Evaluating performance measures on the factory floor

Clearly, all the performance measures mentioned above can be evaluated by monitoring the appropriate metrics on the factory floor and calculating their average values. In current practice, PR or TP are monitored in most production systems (but, astonishingly, not in all). WIP is monitored, on a continuous basis, less often. The blockages and starvations are monitored very rarely. We show in this book that all of them must be monitored continuously in order to exercise measurement-based management of production systems.

3.7 Model Validation

Model validation – process of assessing accuracy of the mathematical model of the production system.

Typically, this process is carried out by comparing predictions of the model with factory floor measurements. For instance, let PR and $\widehat{PR}$ be the production rates identified on the factory floor and calculated using the model, respectively. Then the value of the error, defined as

$$\epsilon_{PR} = \frac{|PR - \widehat{PR}|}{PR} \cdot 100\%, \tag{3.34}$$

gives a measure of model fidelity.

The values of ϵ_{PR}, which can be viewed as acceptable, are of the same order of magnitude as the accuracy with which parameters of the model are identified. Typically, as it was mentioned above, machine parameters are identified with a 5% - 10% accuracy (mostly due to the low accuracy of the data available on the factory floor). Thus, an accuracy of model prediction at the level of 5% - 10% is often viewed as acceptable.

In cases where the errors are outside of this "fuzzy" region, the process of modeling must be repeated anew. In other words, the process of production systems modeling is iterative in nature, as it is the case in all engineering disciplines.

In applications where not only PR but also WIP_i, ST_i, and BL_i are available from factory floor measurements, the latter can be used for model validation as well. The measure of fidelity (3.34), however, must be modified to avoid small denominators when either WIP_i or ST_i or BL_i are very small. Therefore, the following measures for model validation are recommended:

$$\epsilon_{WIP_i} = \frac{|WIP_i - \widehat{WIP}_i|}{N_i} \cdot 100\%, \quad i = 1, \ldots, M-1, \tag{3.35}$$

$$\epsilon_{ST_i} = |ST_i - \widehat{ST}_i|, \quad i = 2, \ldots, M, \tag{3.36}$$

$$\epsilon_{BL_i} = |BL_i - \widehat{BL}_i|, \quad i = 1, \ldots, M-1, \tag{3.37}$$

where N_i is the i-th buffer capacity and $\widehat{WIP}_i$, $\widehat{ST}_i$ and $\widehat{BL}_i$ are calculated using the model.

When a production system is at the design stage and no measurements of its performance are available, the above model validation process cannot be carried out. In this case, opinion of manufacturing system specialists is the only guide for assessing model adequacy.

3.8 Steps of Modeling, Analysis, Design, and Improvement

3.8.1 Modeling

To summarize, the process of production system modeling consists of the following five steps:

- *Layout investigation.* The goal here is to identify the physical layout of the production system at hand, as exemplified, for instance, by Figures 3.13 and 3.15.
- *Structural modeling.* The goal here is to reduce the identified physical layout to one of the standard types of production systems, as exemplified by Figures 3.14 and 3.17.
- *Machine parameters identification.* The goal here is to determine cycle time τ_i, average up- and downtime, $T_{up,i}$ and $T_{down,i}$, and coefficients of variation, $CV_{up,i}$ and $CV_{down,i}$ of each machine in the system. If the process of structural modeling included aggregation of some of the machines, parameters of the aggregated machines must also be calculated, as described in Subsection 3.3.5.
- *Buffer parameter identification.* The purpose here is to identify the storing capacity of each buffer, as described in Subsection 3.4.2.
- *Model validation.* The purpose is to compare predictions of the model with factory floor measurements, as described in Section 3.7. If it turns out that the errors are too large, all steps must be repeated anew.

It must be pointed out that the process of mathematical modeling is, perhaps, the most important stage of production systems analysis, design, and continuous improvement. Engaging in this process, one should keep in mind that, as it was mentioned before, *the model should be as simple as possible but not simpler.*

3.8.2 Analysis, continuous improvement, and design

Typically, mathematical models are used for analysis, continuous improvement, and design of production systems based on one or more of the following steps:

- Evaluation of the model performance measures (e.g., PR, TP, CR, SR, WIP, BL, ST, t_s).
- Identification of system bottlenecks, i.e., machines and buffers that impede the system performance in the strongest manner.
- Investigation of "what if" scenarios, i.e., prediction of system performance if some of its elements are changed (i.e., improvement/replacement of a machine, increase/re-allocation of buffer capacity, re-assignment of workforce, etc.).

- Design of lean buffering, i.e., the smallest buffer capacity, which is necessary and sufficient to ensure the desired throughput of the system.
- Based on the above data, determination of the most promising direction for system improvement, taking into account all practical constraints (i.e., financial and personnel resources, space and time availability, etc.).

Parts II - IV of this textbook provide tools for carrying out these steps for a wide range of production systems in large volume manufacturing environments.

3.9 Simplification: Transforming Exponential Models into Bernoulli Models

3.9.1 Motivation

As it turns out, production systems with Bernoulli machines are easier to analyze than similar systems with exponential machines. Therefore, in some cases it is beneficial to "simplify" an exponential line to a Bernoulli one, if such a simplification does not lead to a substantially lower accuracy. In this section, this simplification, referred to as *exp-B* (exponential-to-Bernoulli) *transformation*, is described and its accuracy with respect to PR calculation is evaluated.

After the exp-B transformation is performed and the subsequent analysis and improvement of the Bernoulli model is carried out, it is often necessary to return to the exponential description (for instance, to guide the implementation of the improvement measures developed). This is accomplished by the so-called *B-exp* (Bernoulli-to-exponential) *transformation*, which is also described below.

3.9.2 Exponential and Bernoulli lines considered

Consider a serial production line with M exponential machines denoted as

$$[\tau_i, exp_{up,i}, exp_{down,i}], \qquad i = 1, \ldots, M.$$

This implies that the i-th machine is capable of producing

$$c_i = \frac{1}{\tau_i} \text{ parts/unit of time,}$$

and up- and downtime are distributed exponentially with parameters λ_i and μ_i, respectively, i.e.,

$$\begin{aligned} f_{up,i}(t) &= \lambda_i e^{-\lambda_i t}, \qquad t \geq 0, \\ f_{down,i}(t) &= \mu_i e^{-\mu_i t}, \qquad t \geq 0. \end{aligned}$$

Thus, the i-th machine is characterized by the triple:

$$(c_i, \lambda_i, \mu_i), \qquad i = 1, \ldots, M.$$

The i-th buffer is characterized by its capacity, N_i, $i = 1, \ldots, M-1$.

Assume that this serial line operates according to the following conventions:

(a) Flow model. (b) Machine status is determined independently from each other. (c) Time-dependent failures. (d) Blocked before service. (e) The first machine is never starved and the last machine is never blocked.

Along with this system, consider a Bernoulli line with M machines denoted as

$$[B_{up,i}, B_{down,i}], \qquad i = 1, \ldots, M.$$

In other words, the i-th machine produces a part during a cycle time with probability p_i^{Ber}, and the cycle time, τ, of all machines is the same. Denote the buffer capacities of this line as N_i^{Ber}, $i = 1, \ldots, M-1$, and assume that the line operates according to the following assumptions:

(a) Synchronous model with slot duration τ. (b) The status of the machines is determined at the beginning of each time slot and the state of the buffer at the end of the time slot. (c) Machine status is determined independently from each other. (d) Time-dependent failures. (e) Blocked before service. (f) The first machine is never starved and the last machine is never blocked.

Given these two production lines, the purpose of the exp-B transformation (illustrated in Figure 3.25) is to calculate p_i^{Ber}, $i = 1, \ldots, M$, N_i^{Ber}, $i = 1, \ldots, M-1$, and τ, so that the throughputs of the two lines, TP^{exp} and $TP^{Ber} = (1/\tau)PR^{Ber}$, are sufficiently close to each other.

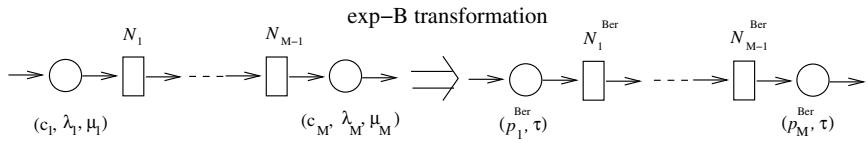


Figure 3.25: Illustration of exp-B transformation

The inverse operation, i.e., the B-exp transformation, is illustrated in Figure 3.26. The goal here is to calculate the parameters of the *transformed* exponential line, i.e.,

$$c_i^{tr}, \lambda_i^{tr}, \mu_i^{tr}, \qquad i = 1, \ldots, M$$

and

$$N_i^{tr}, \qquad i = 1, \ldots, M-1$$

so that the throughput of the transformed exponential line, TP^{tr}, is sufficiently close to $(1/\tau)PR^{Ber}$.

Both of these transformations are described below.

3.9.3 The exp-B transformation

Given the exponential line with machines (c_i, λ_i, μ_i) and buffers N_i, introduce the Bernoulli line with cycle time τ, machines p_i^{Ber}, and buffers N_i^{Ber} defined

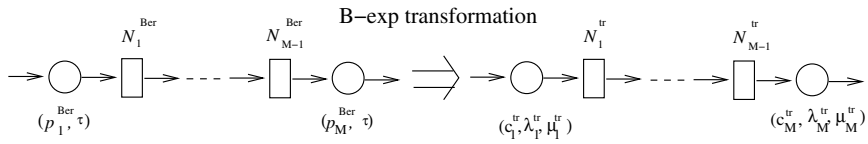


Figure 3.26: Illustration of B-exp transformation

as follows:

$$\tau = \frac{1}{c_{max}}, \tag{3.38}$$

$$p_i^{Ber} = \frac{\frac{c_i}{c_{max}} \cdot \frac{1}{\lambda_i}}{\frac{1}{\lambda_i} + \frac{1}{\mu_i}} = \frac{c_i e_i}{c_{max}}, \qquad i = 1, \ldots, M, \tag{3.39}$$

$$N_i^{Ber} = \min\Big(\frac{N_i}{c_{i+1}}\mu_i, \frac{N_i}{c_i}\mu_{i+1}\Big) + 1, \qquad i = 1, \ldots, M-1, \tag{3.40}$$

where

$$\begin{aligned} c_{max} &= \max(c_i), \qquad i = 1, \ldots, M, \\ e_i &= \frac{\mu_i}{\lambda_i + \mu_i}, \qquad i = 1, \ldots, M. \end{aligned}$$

The reason for equations (3.38) and (3.39) are obvious: the cycle time of the Bernoulli machines equals to that of the fastest exponential machine, and the production rate of each exponential machine in isolation equals that of the corresponding Bernoulli machine. The reason behind equation (3.40) is in the following: The buffer in the Bernoulli model can prevent starvation of the downstream machine and blockage of the upstream machine for a number of time slots at most equal to the size of the buffer. In (3.40), $\frac{N_i}{c_{i+1}}$ represents the largest time during which the buffer can protect the downstream machine when the upstream machine is down. Therefore, $\frac{N_i}{c_{i+1}}\mu_i$ is the fraction of average downtime of the upstream machine that can be accommodated by the buffer. Thus, this quantity can be considered as the equivalent Bernoulli buffer size. Analogously, $\frac{N_i}{c_i}\mu_{i+1}$ is the fraction of average downtime of the downstream machine that can be accommodated by the buffer and also can be viewed as the equivalent Bernoulli buffer size. By choosing the worst case, we have the equivalent buffer size in the Bernoulli model given by (3.40).

Note that under the blocked before service convention, the buffer size should be greater than or equal to 1, because the machine itself is viewed as a unit of buffer capacity. To account for this fact, we need the 1 in (3.40).

In principle, the Bernoulli buffer size should be an integer. However, the theory and calculation formulas developed in Part II do work for fractional buffers as well. Therefore, we allow, according to (3.40), fractional buffer sizes.

The accuracy of the exp-B transformation has been assessed numerically by simulating exponential and the corresponding Bernoulli lines, statistically

evaluating TP^{exp} and $TP^{Ber} = c_{max} PR^{Ber}$ and calculating the error

$$\epsilon_{TP} = \frac{|TP^{exp} - TP^{Ber}|}{TP^{exp}} \cdot 100\%.$$

Numerous systems have been investigated. Typical results are illustrated in Tables 3.2 and 3.3 and Figure 3.27 for three- and ten-machine lines, respectively. As one can see from the tables, for exponential lines, which result in $N_i^{Ber} \geq 2$, the error is quite small (less than 4%); for $N_i^{Ber} < 2$, the error is up to 7% - 8%. Similar conclusions follow from Figure 3.27 for ten-machine lines with parameters selected randomly and equiprobably form the following sets:

$$\begin{aligned} e &= [0.7, 0.98], \\ T_{down} &= [5, 20], \\ N &= [5, 40], \\ c &= [1, 2]. \end{aligned} \tag{3.41}$$

Specifically, Figure 3.27 illustrates the following three exponential lines:

Line 1:

$$\begin{aligned} e &= \{0.867, 0.852, 0.925, 0.895, 0.943, 0.897, 0.892, 0.935, 0.903, 0.870\}, \\ T_{down} &= \{14.23, 16.89, 18.83, 16.08, 7.65, 11.09, 19.05, 18.76, 11.15, 18.42\}, \\ N &= \{7.026, 17.350, 33.461, 5.345, 9.861, 12.097, 11.955, 26.133, 14.527\}, \\ c &= \{1.950, 1.231, 1.607, 1.486, 1.891, 1.762, 1.457, 1.019, 1.822, 1.445\}. \end{aligned}$$

Line 2:

$$\begin{aligned} e &= \{0.945, 0.873, 0.911, 0.899, 0.939, 0.926, 0.896, 0.852, 0.932, 0.895\}, \\ T_{down} &= \{14.22, 16.89, 18.83, 16.08, 7.65, 11.09, 19.05, 18.76, 11.15, 18.42\}, \\ N &= \{5.535, 31.138, 20.578, 37.614, 21.310, 19.653, 34.618, 23.380, 12.093\}, \\ c &= \{1.672, 1.838, 1.020, 1.681, 1.380, 1.832, 1.503, 1.709, 1.429, 1.305\}. \end{aligned}$$

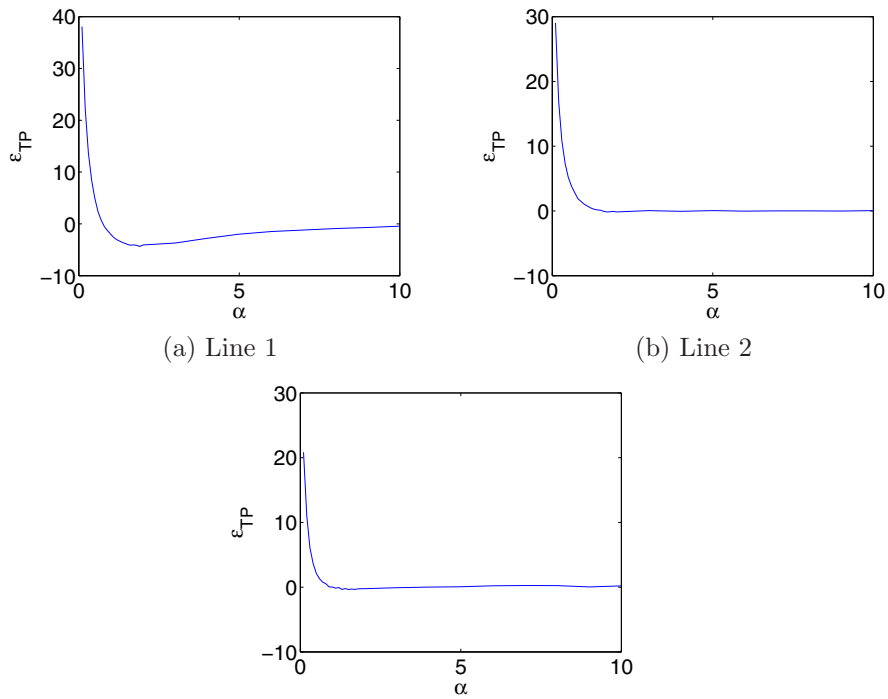
Line 3:

$$\begin{aligned} e &= \{0.869, 0.869, 0.918, 0.880, 0.904, 0.865, 0.920, 0.888, 0.936, 0.935\}, \\ T_{down} &= \{13.91, 12.45, 18.48, 17.33, 14.68, 17.27, 14.90, 10.13, 9.35, 10.12\}, \\ N &= \{26.746, 32.819, 38.490, 23.291, 35.805, 11.054, 39.291, 14.501, 13.832\}, \\ c &= \{1.534, 1.727, 1.309, 1.839, 1.568, 1.370, 1.703, 1.547, 1.445, 1.695\}. \end{aligned}$$

In this figure, the buffer capacity of the exponential lines is selected as

$$\widetilde{N} = \alpha N,$$

where N is given in the definition of each line and $\alpha \in [0.1, 10]$; the error, ϵ_{TP}, is illustrated as a function of α. Again, as one can see, the exp-B transformation is quite accurate for buffers that can accommodate at least one largest downtime of the machines.



(a) Line 1

(b) Line 2

(c) Line 3

Figure 3.27: Accuracy of exp-B transformation for Lines 1 - 3

Table 3.2: Accuracy of the exp-B transformation for identical machines and buffers ($M = 3, \lambda = 0.01, \mu = 0.1$)

N_1	N_2	TP^{exp}	N_1^{Ber}	N_2^{Ber}	TP^{Ber}	$\frac{\lvert TP^{exp} - TP^{Ber} \rvert}{TP^{exp}} \cdot 100\%$
50	50	5.6234±0.0041	1.35	1.35	6.0895	8.3
50	100	5.7491±0.0045	1.35	1.7	6.1727	7.4
50	200	5.8768±0.0046	1.35	2.4	6.2396	6.2
50	400	5.9510±0.0045	1.35	3.8	6.2652	5.3
50	600	5.9641±0.0044	1.35	5.2	6.2677	5.1
100	50	5.8568±0.0041	1.7	1.35	6.2691	7
100	100	5.9811±0.0041	1.7	1.7	6.3565	6.3
100	200	6.1092±0.0043	1.7	2.4	6.4286	5.2
100	400	6.1863±0.0040	1.7	3.8	6.4581	4.4
100	600	6.2005±0.0039	1.7	5.2	6.4612	4.2
200	50	6.1323±0.0039	2.4	1.35	6.4701	5.5
200	100	6.2566±0.0038	2.4	1.7	6.564	4.9
200	200	6.3881±0.004	2.4	2.4	6.644	4
200	400	6.4701±0.0039	2.4	3.8	6.6794	3.2
200	600	6.4864±0.0037	2.4	5.2	6.6836	3
400	50	6.3715±0.0036	3.8	1.35	6.6288	4
400	100	6.5001±0.0035	3.8	1.7	6.7324	3.6
400	200	6.6392±0.0036	3.8	2.4	6.8234	2.8
400	400	6.7311±0.0033	3.8	3.8	6.8662	2
400	600	6.7517±0.0032	3.8	5.2	6.8719	1.8
600	50	6.466±0.0037	5.2	1.35	6.8719	3.3
600	100	6.5998±0.0035	5.2	1.7	6.7911	2.9
600	200	6.7462±0.0034	5.2	2.4	6.889	2.1
600	400	6.8454±0.0032	5.2	3.8	6.9362	1.3
600	600	6.8683±0.0031	5.2	5.2	6.9428	1.1

Table 3.3: Accuracy of the exp-B transformation for non-identical machines and buffers ($c_1 = 7, \lambda_1 = 0.02, \mu_1 = 0.08$), ($c_2 = 10, \lambda_2 = 0.03, \mu_2 = 0.07$), ($c_3 = 9, \lambda_3 = 0.01, \mu_3 = 0.09$)

N_1	N_2	TP^{exp}	N_1^{Ber}	N_2^{Ber}	TP^{Ber}	$\frac{\lvert TP^{exp} - TP^{Ber} \rvert}{TP^{exp}} \cdot 100\%$
50	50	4.4971±0.0041	1.4	1.4	4.3433	7.3
50	100	4.4772±0.0044	1.4	1.7778	4.791	7
50	200	4.5384±0.0044	1.4	2.5556	4.8476	6.8
50	400	4.5577±0.0042	1.4	4.1111	4.8688	6.8
50	600	4.5592±0.0041	1.4	5.6667	4.8708	6.8
100	50	4.7244±0.0036	1.8	1.3889	4.9346	4.4
100	100	4.8001±0.0037	1.8	1.7778	5.0053	4.3
100	200	4.8618±0.004	1.8	2.5556	5.0609	4.1
100	400	4.8837±0.0036	1.8	4.1111	5.083	4.1
100	600	4.8857±0.0036	1.8	5.6667	5.0853	4.1
200	50	5.0632±0.0027	2.6	1.3889	4.9346	2.4
200	100	5.1325±0.0027	2.6	1.7778	5.2498	2.3
200	200	5.1908±0.0028	2.6	2.5556	5.2993	2.1
200	400	5.215±0.0028	2.6	4.1111	5.3197	2
200	600	5.2176±0.0027	2.6	5.6667	5.3197	2
400	50	5.3404±0.003	4.2	1.3889	5.4098	1.3
400	100	5.3954±0.0029	4.2	1.7778	5.4568	1.1
400	200	5.4399±0.0029	4.2	2.5556	5.4907	0.9
400	400	5.4602±0.0029	4.2	4.1111	5.5043	0.8
400	600	5.4629±0.0029	4.2	5.6667	5.5060	0.8
600	50	5.4553±0.0028	5.8	1.3889	5.5031	0.9
600	100	5.4978±0.0028	5.8	1.7778	5.5359	0.7
600	200	5.5296±0.0028	5.8	2.5556	5.557	0.5
600	400	5.5432±0.0028	5.8	4.1111	5.5647	0.4
600	600	5.545±0.0029	5.8	5.6667	5.5657	0.4

3.9.4 The B-exp transformation

The transformation from Bernoulli to the exponential model is accomplished as follows: Suppose a serial Bernoulli production line is given along with the original exponential line from which it has been deduced. Define the parameters of the new exponential line as follows:

$$\mu_i^{tr} = \mu_i, \qquad i = 1, \ldots, M. \tag{3.42}$$

If $\frac{p_i^{Ber} c_{max}}{c_i} < 1$, then

$$c_i^{tr} = c_i, \qquad \text{and} \qquad e_i^{tr} = \frac{p_i^{Ber} c_{max}}{c_i^{tr}}, \qquad i = 1, \ldots, M. \tag{3.43}$$

If $\frac{p_i^{Ber} c_{max}}{c_i} \geq 1$, choose c_i^{tr} such that

$$\frac{p_i^{Ber} c_{max}}{c_i^{tr}} < 1, \qquad \text{and} \qquad e_i^{tr} = \frac{p_i^{Ber} c_{max}}{c_i^{tr}}, \qquad i = 1, \ldots, M. \tag{3.44}$$

From (3.43) and (3.44), obviously

$$0 < e_i^{tr} < 1, \qquad i = 1, \ldots, M. \tag{3.45}$$

Using (3.42), we have

$$e_i^{tr} = \frac{\mu_i}{\lambda_i + \mu_i} = \frac{\mu_i^{tr}}{\lambda_i^{tr} + \mu_i^{tr}}, \qquad i = 1, \ldots, M. \tag{3.46}$$

It follows that

$$\lambda_i^{tr} = \frac{(1 - e_i^{tr})\mu_i^{tr}}{e_i^{tr}}. \tag{3.47}$$

The equation corresponding to (3.40) can be expressed as

$$N_i^{tr} = \max\Big(\frac{N_i^{Ber} - 1}{\mu_i^{tr}} c_{i+1}^{tr}, \frac{N_i^{Ber} - 1}{\mu_{i+1}^{tr}} c_i^{tr}\Big), \qquad i = 1, \ldots, M-1. \tag{3.48}$$

Numerical investigations indicate that the accuracy of the B-exp transformation is roughly the same as that of the exp-B transformation.

Finally, it should be pointed out that the exp-B and B-exp transformations are reversible in the following sense: Transferring an exponential line to a Bernoulli one using (3.38)-(3.40) and then returning to exponential description using (3.42)-(3.48), results in the original exponential line.

PSE Toolbox: The exp-B and B-exp transformations have been implemented as two tools of the toolbox function **Modeling**. For a description and illustration of these tools, see Section 19.2.

3.9.5 Exp-B and B-exp transformations for assembly systems

Exp-B transformaion: This transformation in assembly systems is similar to that in serial lines. Specifically, given the exponential assembly systems with machines $(c_{1i}, \lambda_{1i}, \mu_{1i})$, $(c_{2i}, \lambda_{2i}, \mu_{2i})$, $(c_{0i}, \lambda_{0i}, \mu_{0i})$ and buffers N_{1i}, N_{2i}, N_{0i} (see Figure 1.2), we introduce the Bernoulli assembly system with cycle time τ, machine efficiencies p_{1i}^{Ber}, p_{2i}^{Ber}, p_{0i}^{Ber}, and buffer capacities N_{1i}^{Ber}, N_{2i}^{Ber}, N_{0i}^{Ber} defined as follows:

$$\tau = \frac{1}{c_{max}}, \tag{3.49}$$

$$p_{ji}^{Ber} = \frac{\frac{c_{ji}}{c_{max}} \cdot \frac{1}{\lambda_{ji}}}{\frac{1}{\lambda_{ji}} + \frac{1}{\mu_{ji}}} = \frac{c_{ji} e_{ji}}{c_{max}}, \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j, \tag{3.50}$$

$$N_{ji}^{Ber} = \min\Big(\frac{N_{ji}}{c_{j,i+1}}\mu_{ji}, \frac{N_{ji}}{c_{ji}}\mu_{j,i+1}\Big) + 1, \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j - 1, \tag{3.51}$$

$$N_{jM_j}^{Ber} = \min\Big(\frac{N_{jM_j}}{c_{01}}\mu_{jM_j}, \frac{N_{jM_j}}{c_{jM_j}}\mu_{01}\Big) + 1, \quad j = 1, 2, \tag{3.52}$$

where

$$c_{max} = \max(c_{ji}), \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j,$$
$$e_{ji} = \frac{\mu_{ji}}{\lambda_{ji} + \mu_{ji}}, \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j.$$

B-exp transformation: Similar to B-exp transformation in serial lines, the parameters of the transformed exponential assembly system are defined as

$$\mu_{ji}^{tr} = \mu_{ji}, \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j. \tag{3.53}$$

If $\frac{p_{ji}^{Ber} c_{max}}{c_{ji}} < 1$, then

$$c_{ji}^{tr} = c_{ji}, \quad \text{and} \quad e_{ji}^{tr} = \frac{p_{ji}^{Ber} c_{max}}{c_{ji}^{tr}}, \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j. \tag{3.54}$$

If $\frac{p_{ji}^{Ber} c_{max}}{c_{ji}} \geq 1$, choose c_{ji}^{tr} such that

$$\frac{p_{ji}^{Ber} c_{max}}{c_{ji}^{tr}} < 1, \quad \text{and} \quad e_{ji}^{tr} = \frac{p_{ji}^{Ber} c_{max}}{c_{ji}^{tr}}, \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j. \tag{3.55}$$

From (3.54) and (3.55),

$$0 < e_{ji}^{tr} < 1, \quad j = 0, 1, 2, \ \ i = 1, \ldots, M_j. \tag{3.56}$$

Using (3.53) we have

$$e_{ji}^{tr} = \frac{\mu_{ji}}{\lambda_{ji} + \mu_{ji}} = \frac{\mu_{ji}^{tr}}{\lambda_{ji}^{tr} + \mu_{ji}^{tr}}, \qquad j = 0, 1, 2, \quad i = 1, \ldots, M_j. \tag{3.57}$$

Then, it follows that

$$\lambda_{ji}^{tr} = \frac{(1 - e_{ji}^{tr})\mu_{ji}^{tr}}{e_{ji}^{tr}}, \qquad j = 0, 1, 2, \quad i = 1, \ldots, M_j. \tag{3.58}$$

The buffer capacities of the transformed exponential lines can be expressed as

$$N_{ji}^{tr} = \max\Big(\frac{N_{ji}^{Ber} - 1}{\mu_{ji}^{tr}} c_{j,i+1}^{tr}, \frac{N_{ji}^{Ber} - 1}{\mu_{j,i+1}^{tr}} c_{ji}^{tr}\Big), \qquad j = 0, 1, 2, \quad i = 1, \ldots, M_j - 1, \tag{3.59}$$

$$N_{jM_j}^{tr} = \max\Big(\frac{N_{jM_j}^{Ber} - 1}{\mu_{jM_j}^{tr}} c_{01}^{tr}, \frac{N_{jM_j}^{Ber} - 1}{\mu_{01}^{tr}} c_{jM_j}^{tr}\Big), \qquad j = 1, 2. \tag{3.60}$$

PSE Toolbox: The exp-B and B-exp transformations for assembly systems have been implemented as two tools of the toolbox function **Modeling**. For a description and illustration of these tools, see Section 19.2.

3.10 Case Studies

This section describes modeling of several production systems from the automotive industry. The resulting models are used in Parts II - IV to illustrate methods of analysis, continuous improvement, and design included in this textbook.

3.10.1 Automotive ignition coil processing system

System description and layout: The production system of coils for an automotive ignition unit is shown in Figure 3.28. It consists of 16 operations where the coils are either processed (e.g., Op. 5) or have external parts attached (e.g., Op. 13). Ops. 8 and 12 load the external parts. Ops. 9 and 10 carry out identical operations and are coupled in the sense that one of them being down forces the other down as well; parts are directed to either Op. 9 or 10 automatically. Therefore, Ops. 9 and 10 are aggregated into one machine.

The coils are transported within the system on pallets by a conveyor. The empty pallet buffer is implemented by two devices referred to as elevators. The raw material (i.e., an unfinished coil) is loaded at Op. 1, if an empty pallet is available; if not, Op. 1 is starved for pallets. Op. 16 transfers the finished product to the downstream system and releases the pallet, if the subsequent elevator is not full; otherwise, Op. 16 is blocked.

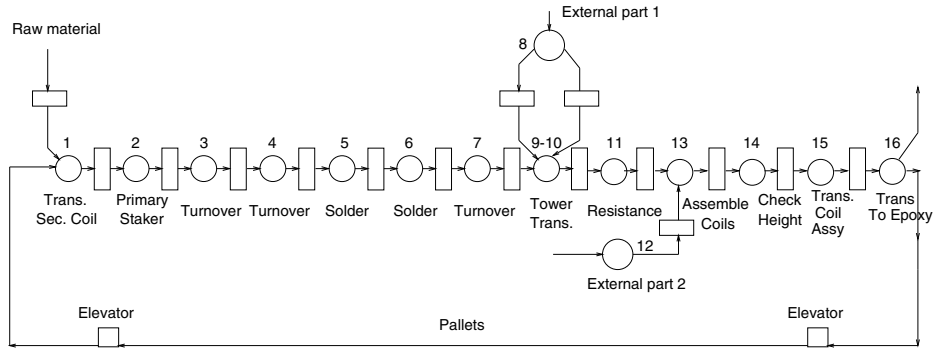


Figure 3.28: Layout of the ignition coil processing system

The system's performance during eight consecutive weeks is characterized in Table 3.4. For the first three weeks, the machines cycle time was set to produce 562.53 parts/hour. The measured performance, however, was substantially lower, amounting to an average loss of 16.1%. For the subsequent five weeks, the system was sped up to produce 592.07 parts/hour, while the average throughput remained practically the same (472.6 parts/hr), with average losses of 20.32%.

Table 3.4: System performance

Week	Nominal throughput (parts/hr)	Actual throughput (parts/hr)	Losses (%)	Average throughput (parts/hr)	Average losses (%)
Week 1	562.53	464	17.52	472	16.10
Week 2	562.53	505	10.23		
Week 3	562.53	447	20.54		
Week 4	593.07	501	15.52	472.6	20.32
Week 5	593.07	454	23.45		
Week 6	593.07	424	28.51		
Week 7	593.07	480	19.07		
Week 8	593.07	504	15.02		

The goal of this case study was to identify reasons for these losses and provide suggestions for system improvement. While the description of the analysis and continuous improvement is given in Part II, its modeling is described below.

Structural modeling: Neglecting the closed nature of the conveyor and the effect of Ops. 8 and 12 (which are quite fast and reliable and, therefore, do not impede the system performance), the model of the coil processing system can be represented as the serial production line shown in Figure 3.29. This

structural model is used for the subsequent modeling, analysis, and continuous improvement.

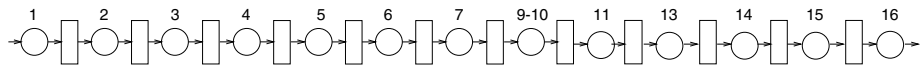


Figure 3.29: Structural model of ignition coil processing system

Modeling and identification of the machines: Assuming that the machines are exponential, their identification amounts to determining the triple $\{\tau_i, \lambda_i, \mu_i\}$, for all operations. Based on Table 3.4, the cycle time for Period 1 (the first three weeks) was 6.4 sec/part and for Period 2 (the last five weeks) 6.07 sec/part. The machines' average up- and downtime were determined from the Weekly Data Report using expression (3.7) and (3.8) (see Table 3.5 and note that Ops. 4 and 13 have zero downtime, and, therefore, are not included in Table 3.5). Based on these data, λ_i and μ_i have been calculated as follows:

$$\begin{aligned} \lambda_i &= \frac{1}{T_{up,i}}, \\ \mu_i &= \frac{1}{T_{down,i}}. \end{aligned}$$

The results are given in Table 3.6.

Table 3.5: Average up- and downtimes (min)

	Op. 1	Op. 2	Op. 3	Op. 5	Op. 6	Op. 7
Uptime	227.79	188.11	504.15	1515.5	572.27	1493.2
Downtime	1.837	1.517	1.517	1.517	16.485	9.013

	Op. 9-10	Op. 11	Op. 14	Op. 15	Op. 16
Uptime	13.98	43.07	74.33	188.11	356.02
Downtime	1.571	1.748	1.517	1.517	2.149

(a). Period 1

	Op. 1	Op. 2	Op. 3	Op. 5	Op. 6	Op. 7
Uptime	141.18	280.31	651.73	438.67	450.37	1974
Downtime	2.15	1.976	3.275	4.879	3.632	1.976

	Op. 9-10	Op. 11	Op. 14	Op. 15	Op. 16
Uptime	16.25	45.11	52.91	168.89	201
Downtime	2.05	2.076	1.976	2.398	2.854

(b). Period 2

Table 3.6: Parameters of the machines (1/min)

	Op. 1	Op. 2	Op. 3	Op. 5	Op. 6	Op. 7
λ_i	0.0044	0.0053	0.0020	0.0007	0.0017	0.0007
μ_i	0.5444	0.6592	0.6592	0.6592	0.0607	0.1110

	Op. 9-10	Op. 11	Op. 14	Op. 15	Op. 16
λ_i	0.0715	0.0232	0.0135	0.0053	0.0028
μ_i	0.6365	0.5721	0.6592	0.6592	0.4653

(a). Period 1

	Op. 1	Op. 2	Op. 3	Op. 5	Op. 6	Op. 7
λ_i	0.0071	0.0036	0.0015	0.0023	0.0022	0.0005
μ_i	0.4651	0.5061	0.3053	0.2050	0.2753	0.5061

	Op. 9-10	Op. 11	Op. 14	Op. 15	Op. 16
λ_i	0.0615	0.0222	0.0189	0.0059	0.0050
μ_i	0.4878	0.4817	0.5061	0.4170	0.3504

(b). Period 2

Modeling and identification of buffers: The buffering capacity of the conveyor between each pair of consecutive operations has been evaluated using the method of Subsection 3.4.2. The results are given in Table 3.7.

Table 3.7: Buffer capacity

Operation	Op. 1	Op. 2	Op. 3	Op. 4	Op. 5	Op. 6
Buffer capacity	3	1	1	4	1	1

Operation	Op. 7	Op. 9-10	Op. 11	Op. 13	Op. 14	Op. 15
Buffer capacity	4	1	6	1	1	2

Overall system model: Based on the data of Tables 3.6 and 3.7, the exponential model of the coil processing system has been identified as shown in Figure 3.30 (where, as before, $e_i = \frac{\mu_i}{\lambda_i+\mu_i}$ is the i-th machine efficiency).

The Bernoulli model of this system can be obtained using the exp-B transformation of Section 3.9. The resulting system is shown in Figure 3.31. In addition, Figure 3.31 includes the effect of the conveyor by multiplying p_1 by the probability that Op. 1 is not starved for pallets. This probability has been identified experimentally during the steady state of system operation. No blockages of Op. 16 by full elevators have been observed and, therefore, p_{16} has not

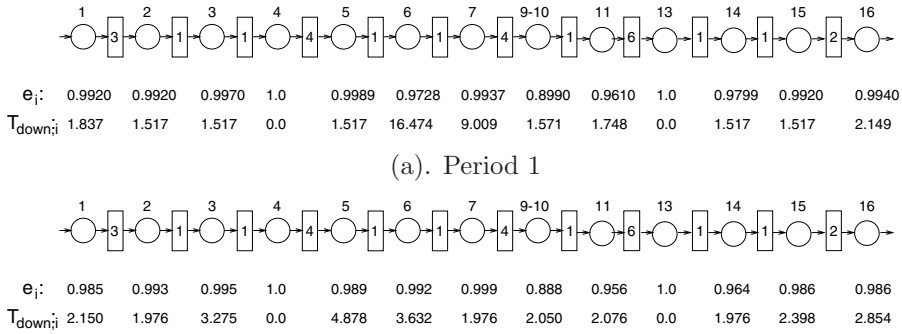


Figure 3.30: Exponential model of the ignition coil processing system

been modified.

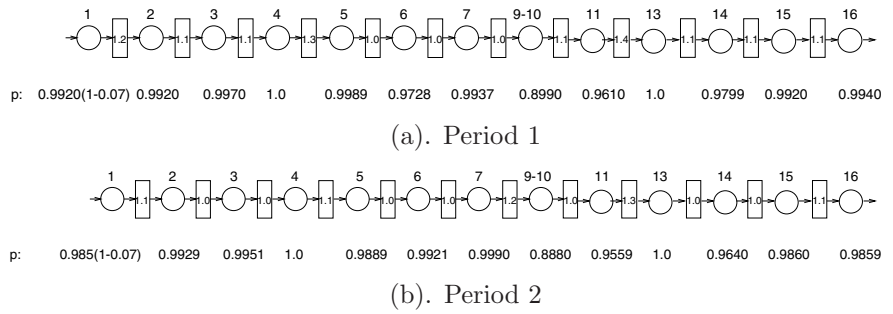


Figure 3.31: Bernoulli model of the ignition coil processing system

The validation, analysis, and continuous improvement of this model are described in Part II.

3.10.2 Automotive paint shop production system

System description and layout: The layout of a paint shop production system at an automotive assembly plant is shown in Figure 3.32. It consists of 11 operations in which the car bodies (referred to as jobs) are cleaned (chemically or physically), sealed (against water leaks), painted, and, finally finessed. Operations 5, 6, and 8 consist of two parallel lines (due to capacity considerations). Operation 10 consists of five parallel painting booths (for both capacity reasons and to ensure color variety).

The jobs within the system are transported by conveyors on two types of carriers. The transfer from one carrier to another occurs after Op. 3. Thus, carriers of type 1 are used in Ops. 1 - 3 and type 2 in Ops. 4 - 11.

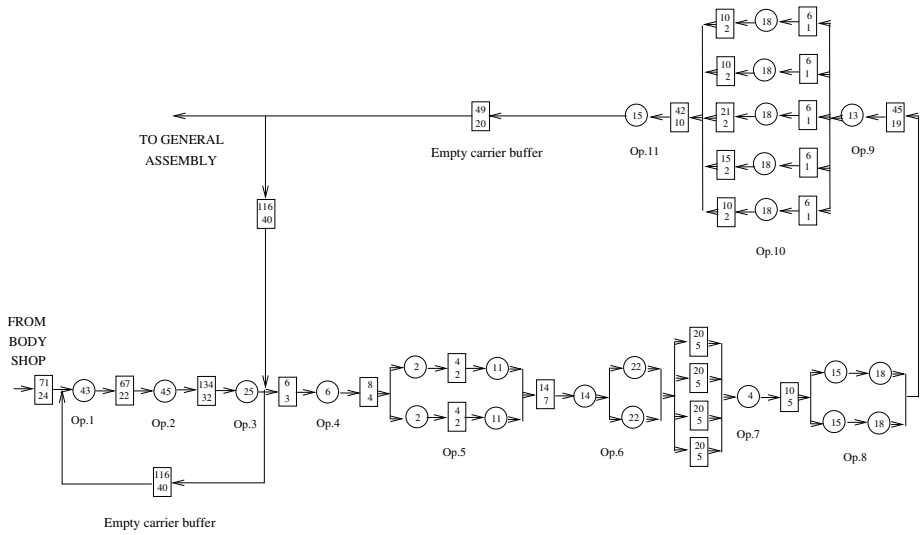


Figure 3.32: Layout of paint shop system

The technological operations are performed while the jobs are moving on their carriers. That is why the parts of the conveyor where the work is being carried out are called operational conveyors, while other parts are referred to as accumulators.

The conveyor as a whole has a modular structure in the sense that each operational conveyor can be stopped without stopping other operational conveyors. Workers typically use push-buttons to stop their operational conveyors in order to complete respective operations with the desired quality. Thus, the downtime is typically due to quality issues rather than machine breakdowns.

The numbers in the circles of Figure 3.32 indicate the number of jobs within the operational conveyors necessary to ensure continuous production. The numbers in rectangles show the minimal occupancy of accumulators to ensure continuous production and the maximal number of jobs that could be contained within an accumulator. Thus, the difference between these two numbers is the buffering capacity of the accumulator.

This system was designed to produce 63 jobs/hour (see the capacity, c_i, of each operation given in Table 3.8). In reality, however, the throughput was much lower, averaging 52.1 jobs/hour (see Table 3.9 where the measured average throughput for five consecutive months is shown). The goal of this case study was to determine reasons for the production losses and to provide recommendations for their elimination. These analyses are described in Part II, while the mathematical model used for this purpose is constructed below.

Structural modeling: The average production losses in Ops. 1 - 11 due to internal reasons (i.e., excluding blockages and starvations) are shown in Table

Table 3.8: Capacity of the machines (jobs/hour)

Ops.	1	2	3	4	5	6	7	8	9	10	11
c_i	63	63	63	63	63	63	63	63	72	100	100

Table 3.9: System performance (jobs/hour)

Period	Month 1	Month 2	Month 3	Month 4	Month 5
TP	53.5	43.81	51.27	54.28	55.89

3.10. Clearly, Ops. 1 and 2 have very low or no losses and, therefore, can be excluded. To accomplish this, we conceptually transfer the common point of the two loops of Figure 3.32 from the output of Op. 3 to its input. This transformation does not lead to reduced accuracy since Op. 3 operates in so-called no-gap mode (i.e., no empty space between consecutive jobs on the operational conveyor is allowed). Therefore, after aggregating the parallel machines of Ops. 5, 6, 8, and 10, we represent the system as shown in Figure 3.33.

Table 3.10: Average production losses (jobs/hour)

Operation	Month 1	Month 2	Month 3	Month 4	Month 5
Op. 1	0	0.05	0.01	0	0
Op. 2	0.05	0.45	0.07	0.03	0.00
Op. 3	2.88	2.15	0.64	2.26	1.35
Op. 4	2.77	2.60	4.45	2.00	2.60
Op. 5	0.23	0.01	0.04	1.07	1.64
Op. 6	0	0	0.001	0.02	0.39
Op. 7	1.09	3.13	1.09	2.68	2.05
Op. 8	1.39	3.42	1.28	2.73	0.41
Op. 9	6.18	7.38	7.01	6.59	6.14
Op. 10	0.35	1.40	1.66	3.09	3.63
Op. 11	0.01	0.01	0.01	0.01	0.01

Modeling and identification of the machines: Since, as it follows from Table 3.10, downtime of each operation is mostly of the same order of magnitude as the cycle time, we adopt the Bernoulli model of machine reliability. The

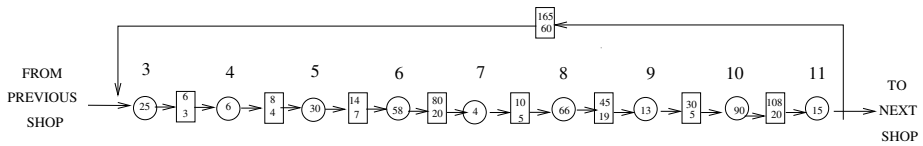


Figure 3.33: Structural model of paint shop system

parameters p_i are calculated as

$$p_i = \min\left\{1, \frac{c_i - L_i}{63}\right\}, \qquad i = 3, \ldots, 11, \tag{3.61}$$

where c_i and L_i are the capacity and the losses of the i-th operation given in Tables 3.8 and 3.10, respectively. These parameters are summarized in Table 3.11.

Table 3.11: Machine parameters (p_i)

Operations	3	4	5	6	7	8	9	10	11
Month 1	0.9543	0.9560	0.9963	1	0.9827	0.9779	1	1	1
Month 2	0.9659	0.9587	0.9998	1	0.9503	0.9457	1	1	1
Month 3	0.9898	0.9294	0.9994	1	0.9827	0.9797	1	1	1
Month 4	0.9641	0.9683	0.9830	0.9997	0.9575	0.9567	1	1	1
Month 5	0.9786	0.9587	0.9740	0.9938	0.9675	0.9935	1	1	1

Modeling and identification of the buffers: As indicated above, the buffering capacity of each accumulator is the difference between its maximal and minimal occupancy. The capacity of the buffers after Ops. 6 and 10 is assumed to be the sum of the capacities of the parallel buffers. The buffers within Op. 5 are omitted. The resulting data on buffer capacity are summarized in Table 3.12.

Table 3.12: Buffer capacity

Operations	3	4	5	6	7	8	9	10
N_i	3	4	7	60	5	26	25	88

Overall system model: Based on the above, the Bernoulli model of the automotive paint shop is represented as shown in Figure 3.34 and the parameters

p_i for each of the five months are given in Table 3.11. Operations 9 - 11 are omitted since their efficiency is 1 and, therefore, they do not affect the system performance. The effect of the closed loop of Figure 3.33 is taken into account using the probability P_{st} that Op. 3 is starved (or, in the original system of Figure 3.32, blocked) by carriers. This probability, evaluated during normal system operation, is given in Table 3.13. In Figure 3.34 this probability is used in the factor $(1 - P_{st})$ multiplying p_3. Thus, a simplified model of the paint shop system is constructed.

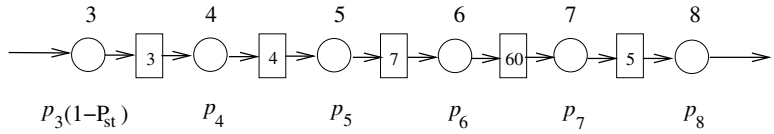


Figure 3.34: Simplified structural model of paint shop system

Table 3.13: Estimated probability of starvation of Op. 3

	Month 1	Month 2	Month 3	Month 4	Month 5
P_{st}	0.0981	0.1171	0.1113	0.1046	0.0975

The validation of this model and its analysis and continuous improvement are described in Part II.

3.10.3 Automotive ignition module assembly system

The description of this system and its structural modeling are given in Section 3.2. Below, the subsequent modeling steps are carried out.

The nominal throughput of the system is 600 parts/hour. The actual performance for six consecutive months is summarized in Table 3.14. As it follows from this table, the average throughput over the six months is 362 parts/hour, i.e., the system operates at 60% of its capacity.

Table 3.14: Actual throughput of the system for six months

Month	May	June	July	Aug.	Sep.	Oct.
TP (parts/hr)	337	347	378	340	384	383

Modeling and identification of the machines: We assume that both the uptime and downtime of the machines are distributed exponentially. Their identification requires to determine the average up- and downtime for each machine,

or their recipricals λ_i and μ_i. The data for this identification have been measured during real-time system operation and are summarized in Tables 3.15 and 3.16. This, along with the cycle time of each operation, 6 sec/part, identifies completely the exponential machines that compromise the system.

Table 3.15: Machine parameters

Operations	May		June		July	
	T_{down} (min)	T_{up} (min)	T_{down} (min)	T_{up} (min)	T_{down} (min)	T_{up} (min)
1	4.8	38.8	8.4	159.6	18.3	286.7
2	3.0	35.4	1.7	55	3.7	42.6
3	4.5	70.5	5.7	136.8	12.4	142.6
4	10.2	68.3	7.1	57.4	12.7	66.7
5	8.9	65.3	6.3	56.7	12.7	66.7
6	2.0	98	6.7	216.6	5.2	59.8
7	1.8	58.2	4.4	142.3	12.2	231.8
8	2.5	47.5	2.4	37.6	5.7	65.6
9	3.9	31.6	7.1	81.7	7.3	65.7
10	2.6	34.5	3.3	33.4	4.0	53.1
11	2.7	31.1	3.4	45.2	4.1	54.5
12	3.3	326.7	0.9	89.1	16.9	224.5
13	3.8	91.2	1.6	38.4	17.9	144.8
14	5.2	98.8	2.5	47.5	10.2	103.1
15	1.8	14.6	2.8	90.5	16.8	223.2
16	2.8	137.2	10.8	529.2	23.4	269.1
17	1.7	55	10.3	504.7	27.7	368
18	2.2	107.8	1.8	20.7	3.4	30.6

Modeling and identification of the buffers: The capacity of the buffers has been identified using the method of Subsection 3.4.2. The resulting capacities are shown in Table 3.17.

Overall system model: Based on the data of Tables 3.15 - 3.17, the exponential model of the coil assembly system has been identified as shown in Figure 3.35 (with the up- and downtime data for the month of May).

For the subsequent analysis, the exponential model of Figure 3.35 and similar models for five other months have been reduced to Bernoulli models, using the exp-B transformation of Section 3.9. The resulting system is shown in Figure 3.36. In this figure, the effect of the closed nature of circular conveyors, i.e., the starvation of Ops. 1 and 9 and the blockage of Op. 18 are taken into account. Specifically, the average fraction of time when Ops. 1 and 9 were starved for

Table 3.16: Machine parameters (cont.)

Operations	August		September		October	
	T_{down} (min)	T_{up} (min)	T_{down} (min)	T_{up} (min)	T_{down} (min)	T_{up} (min)
1	4.4	435.6	3.3	107	9.6	110
2	2.7	267.3	3.7	366	10.7	142
3	2.0	48	1.5	28	2.8	32
4	11.4	92.2	2.2	20	7.7	56
5	10.7	86.6	4.2	48	11.8	95
6	5.9	190.8	3.1	152	5.5	178
7	8.2	401.8	4.1	133	6.1	197
8	2.3	43.7	0.7	34	2.8	67
9	6.5	47.7	4.1	78	2.9	45
10	4.0	46	3.4	65	3.6	86
11	5.1	24.9	2.4	32	1.9	25
12	9.6	310.4	14.7	149	1.6	158
13	6.6	125.4	1.2	39	2.1	50
14	4.5	59.8	2.9	142	2.6	12
15	6.5	58.5	1.0	99	1.2	39
16	5.5	73.1	2.3	22998	1.2	11999
17	4.1	77.9	1.0	9999	4.9	485
18	3.3	38	1.3	64	2.3	113

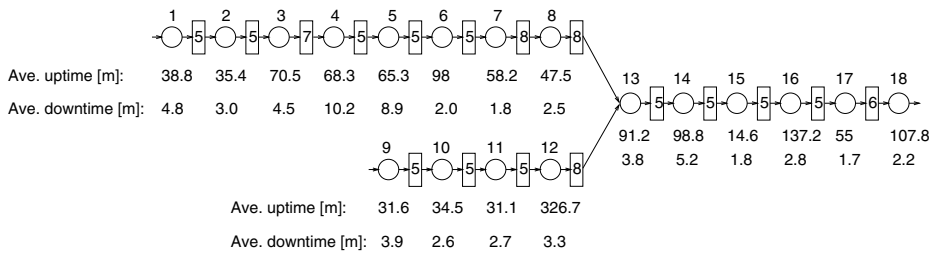


Figure 3.35: Exponential model of the ignition module assembly system

Table 3.17: Buffer capacity

Operations	1	2	3	4	5	6	7	8	9
Buffer capacity	5	5	7	5	5	5	8	8	5

Operations	10	11	12	13	14	15	16	17
Buffer capacity	5	5	8	5	5	5	5	6

pallets and Op. 18 blocked has been identified during real-time operation. The results are summarized in Table 3.18.. To take these losses into account, the efficiencies of Ops. 1, 9 and 18 have been multiplied by a factor (1- average fraction of time when starvation or blockage takes place) (see Figure 3.36).

Table 3.18: Average frequency of starvation and blockage

Month	May	June	July	Aug.	Sep.	Oct.
Average fraction of time when Op. 1 is starved	0.257	0.308	0.285	0.28	0.199	0.252
Average fraction of time when Op. 9 is starved	0.134	0.247	0.199	0.099	0.142	0.089
Average fraction of time when Op. 18 is blocked	0.236	0.256	0.226	0.189	0.11	0.238

The validation, analysis, and continuous improvement of this model are described in Part IV.

3.11 Summary

- The types of production systems considered in this textbook are serial lines and assembly systems.
- The process of mathematical modeling of production systems consists of the following steps:
  - layout investigation
  - structural modeling
  - machine parameter identification
  - buffer parameter identification
  - model validation.

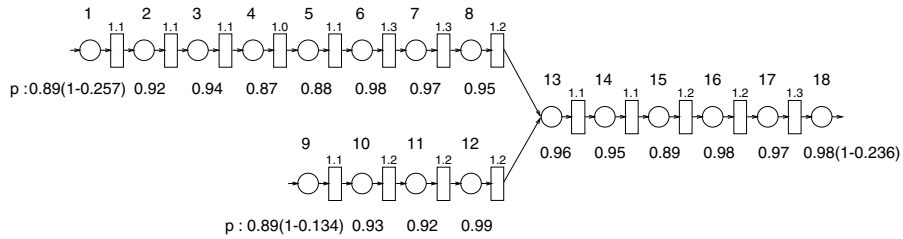


Figure 3.36: Bernoulli model of the ignition module assembly system (based on May data)

- A machine model consists of its
  - cycle time
  - pmf or pdf of its uptime
  - pmf or pdf of its downtime
  - pmf or pdf of parts quality.
- The specific pmf's and pdf's considered in this textbook are:
  - Bernoulli
  - geometric
  - exponential
  - Rayleigh
  - Weibull
  - gamma
  - log-normal
  - general.
- Buffers are modeled by their storing capacity.
- A production system can be considered as operating
  - in slotted or continuous time
  - with time-dependent or operation-dependent failures
  - synchronously or asynchronously
  - as a discrete event system or as a flow system
  - with blocked before service or with blocked after service convention.
- Performance measures that characterize the behavior of production systems are:
  - production rate or throughput, consumption rate, and scrap rate
  - work-in-process
  - finished goods inventory
  - probability of blockages and starvations
  - the level of customer demand satisfaction (or due-time performance)
  - transient properties.
- Mathematical models of production systems described in this chapter are used throughout this textbook for case studies.

3.12 Problems

Problem 3.1 A production system manufactures products A and B. Each product consists of two parts: $A1$ and $A2$ for product A and $B1$ and $B2$ for product B. The processing of A and B require several technological steps. The departments where these steps are carried out are shown in Figure 3.37. The number of each department indicates its order in the technological process. The

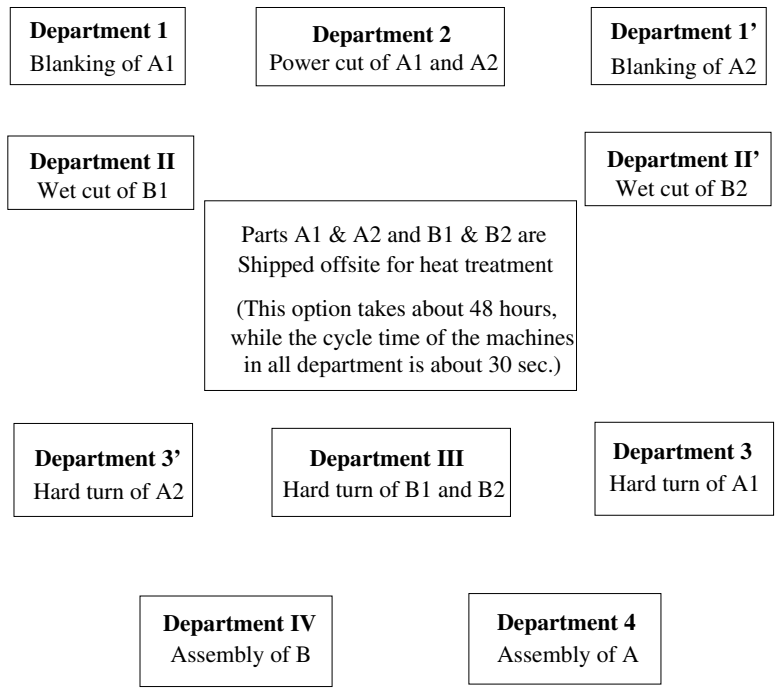


Figure 3.37: Problem 3.1

material handling among the departments is carried out by carts, which are pushed by machine operators from one department to another.

(a) Construct a structural model of this production system.

(b) Describe the data that have to be collected to identify this model.

(c) Describe which steps must be taken to collect these data.

(d) Describe which steps must be taken to validate this model.

Problem 3.2 The layout of a production system for an automotive ignition device is shown in Figure 3.38. It consists of four main operations: Housing Subassembly, Valve Body Assembly, Injector Subassembly, and Injector Final Assembly. In addition, the system contains Shell Assembly, three Welding operations (L.H.W., U.H.W., and Weld), two Overmold operations (O.M.1 and

O.M.2), two Set Stroke operations (Stroke 1 and Stroke 2), one Leak Test operation (L.T.) and one High Potential operation (Hi Pot). Finally, the system includes five buffers positioned as shown in Figure 3.38 and conveyor buffering among all other operations.

Construct a structural model for this system and simplify it to a serial line.

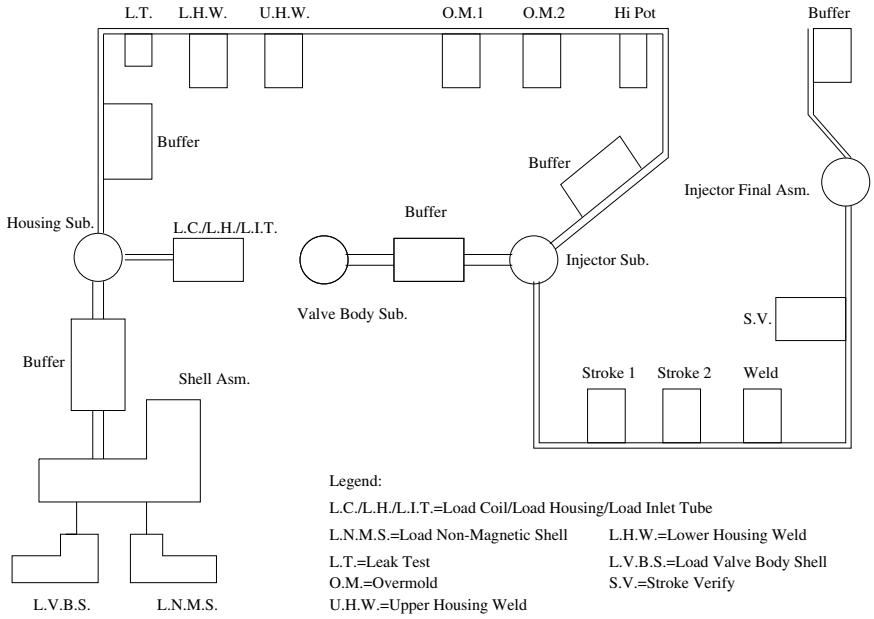


Figure 3.38: Problem 3.2

Problem 3.3 Consider the production system of Figure 3.38 and its model as a serial line obtained in Problem 3.2. Assume that the first five machines are of interest and their parameters are as follows: The cycle time of each machine is 3 sec and the breakdown and repair rates are (in units of 1/min):

$$\begin{array}{llllll}
\lambda\ (\text{S.A}) & = & 0.1075, & \mu\ (\text{S.A.}) & = & 0.5; \\
\lambda\ (\text{H.S.}) & = & 0.3173, & \mu\ (\text{H.S.}) & = & 0.6711; \\
\lambda\ (\text{L.T.}) & = & 0.0051, & \mu\ (\text{L.T.}) & = & 0.5; \\
\lambda\ (\text{L.H.W.}) & = & 0.0051, & \mu\ (\text{L.H.W.}) & = & 0.5; \\
\lambda\ (\text{U.H.W.}) & = & 0.0101, & \mu\ (\text{U.H.W.}) & = & 1.
\end{array}$$

Assume also that the buffers between these operations have the following capacities:

$$\begin{array}{lll}
\text{buffer between S.A. and H.S.} & = & 125; \\
\text{buffer between H.S. and L.T.} & = & 500; \\
\text{buffer between L.T. and L.H.W.} & = & 15; \\
\text{buffer between L.H.W. and U.H.W.} & = & 15.
\end{array}$$

(a) Construct the Bernoulli model of this five-machine exponential serial line.

(b) Using the Simulation function of the PSE Toolbox, investigate the accuracy of the Bernoulli model.

Problem 3.4 A serial production line with five exponential machines is defined as follows:

$$\lambda = [0.0025, 0.0011, 0.0016, 0.0031, 0.0042] \text{ (in units of 1/min)},$$
$$\mu = [0.025, 0.0333, 0.025, 0.0286, 0.05] \text{ (in units of 1/min)},$$
$$c = [1.0714, 1.0714, 0.6667, 0.9375, 1] \text{ (in units of parts/min)},$$
$$N = [26, 10, 28, 30].$$

(a) Construct the Bernoulli model of this five-machine exponential serial line.

(b) Using the Simulation function of the PSE Toolbox, investigate the accuracy of the Bernoulli model.

Problem 3.5 The layout of a production system for an automotive ignition device is shown in Figure 3.39. It consists of 15 operations, separated by buffer-conveyors. Construct a structural model for this system and simplify it to a serial line.

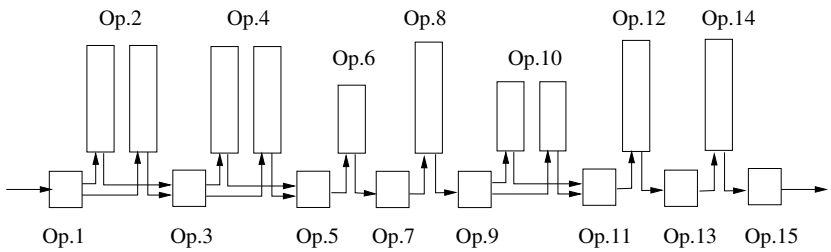


Figure 3.39: Problem 3.5

Problem 3.6 The layout of an automotive camshaft production line is shown in Figure 3.40. Construct a structural model for this system and simplify it into two parallel serial lines.

3.13 Annotated Bibliography

Various aspects of production systems modeling can be found in

[**3.1**] J.A. Buzacott and J.G. Shantikumar, *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ, 1993.

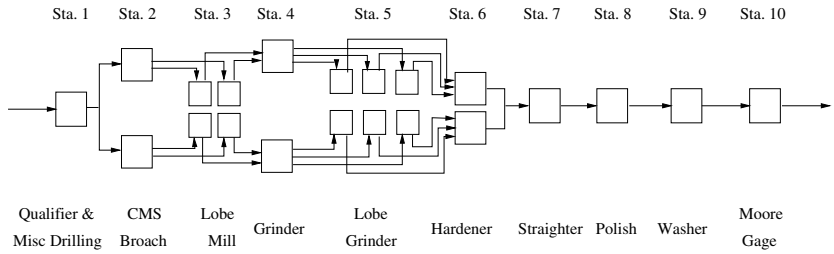


Figure 3.40: Problem 3.6

[**3.2**] H.T. Papadopoulos, C. Heavey and J. Browne, *Queueing Theory in Manufacturing Systems Analysis and Design*, Chapman & Hall, London, 1993.

[**3.3**] S.B. Gershwin, *Manufacturing Systems Engineering*, Prentice Hall, Englewood Cliffs, NJ, 1994.

[**3.4**] T. Altiok, *Performance Analysis of Manufacturing Systems*, Springer, New York, 1997.

A detailed discussion of different types of machine blocking is given in

[**3.5**] H.G. Perros, *Queueing Networks with Blocking*, Oxford University Press, New York, 1994.

The exp-B and B-exp transformations are introduced in

[**3.6**] C.-T. Kuo, *Bottlenecks in Production Systems: A Systems Approach*, Ph.D Thesis, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 1996.

The accuracy of parallel machine aggregation (Subsection 3.3.5) has been investigated by F. Xu in the framework of his Research Experience for Undergraduates at the University of Michigan. He also produced the data for Figure 3.27.

The case studies reported in Section 3.10 have been carried out by S.-Y. Chiang, D. Jacobs, C.-T. Kuo, J. Li, J.-T. Lim, S. M. Meerkov, F. Top, and L. Zhang. Some of them can be found in

[**3.7**] J.-T. Lim, S.M. Meerkov and F. Top, "Homogeneous, Asymptotically Reliable Serial Production Lines: Theory and a Case Study," *IEEE Transactions on Automatic Control*, vol. 35, pp. 524-534, 1990.

[**3.8**] D.A. Jacobs and S.M. Meerkov, "A System-Theoretic Property of Serial Production Lines: Improvability," *International Journal of Systems Science*, vol. 26, pp. 755-785, 1995.

[**3.9**] J.-T. Lim and S.M. Meerkov, "On Asymptotically Reliable Serial Production Lines," *Control Engineering Practice*, vol. 1, 147-152, 1993.

[**3.10**] C.-T. Kuo, J.-T. Lim and S.M. Meerkov, "Bottlenecks in Serial Production Lines: A System-Theoretic Approach," *Mathematical Problems in Engineering*, vol. 2, pp. 233-276, 1996.

[**3.11**] S.-Y. Chiang, C.-T. Kuo and S.M. Meerkov, "DT-Bottlenecks in Serial Production Lines: Theory and Applications," *IEEE Transactions on Robotics and Automation*, vol. 16, pp. 567-580, 2000.

[**3.12**] S.-Y. Chiang, C.-T. Kuo and S.M. Meerkov, "Improvability of Assembly Systems II: Improvability Indicator and Case Study," *Mathematical Problems in Engineering*, vol. 5, pp. 359-393, 2000.

[**3.13**] S.-Y. Chiang, C.-T. Kuo and S.M. Meerkov, "c-Bottlenecks in Serial Production Lines: Identification and Application," *Mathematical Problems in Engineering*, vol. 7, pp. 543-578, 2001.

[**3.14**] J. Li and S.M. Meerkov, "Customer Demand Satisfaction in Production Systems: A Due-Time Performance Approach," *IEEE Transactions on Robotics and Automation*, vol. 17, pp. 472-482, 2001.