Chapter 3
Single Workstation Factory Models

Throughout the analyses given in this textbook, emphasis is on the development of steady-state system measures such as the expected number of jobs in the system (WIP) and their mean cycle times (CT). For these analyses, it is often useful to obtain the probability mass function (pmf) of the steady-state number of jobs in the system. From these pmf's, the measures of system effectiveness can often be developed. For notational purposes, define the random variable N as the number of jobs in the system and define p_n as the probability that the number of jobs in the system is n; namely, $p_n = \Pr\{N = n\}$. In the first section, a method is developed for deriving equations that determine the steady-state probabilities p_n for $n = 0, 1, \cdots$. The initial models will include probabilistic behavior for the arrival process and processing times, and the early models will restrict these two probability laws to the exponential distribution.

Important assumptions on the operating characteristics of the system are also made. It is assumed that job inter-arrival times are independent of the status of the system. Another operating assumption is that the server will never be idle when there is a job in the system that can be served. That is, if it is allowed for the processor to serve a job, then no delay occurs between the time that one job leaves the server and the next job begins processing on the server. Here the assumption is made that the server is always busy processing jobs when there are jobs available for service. Thus, the server will only be idle when there are no jobs available. In later models, nonproductive times will be incorporated into the model. For example, in order to have realistic models for many systems, machine breakdowns will need to be incorporated.

3.1 First Model

Consider a single server with a limited waiting area for $n_{\max} - 1$ jobs and one in the server position for a maximum of $n_{\max}$ jobs in the system. Jobs arrive to the system one at a time with exponentially distributed inter-arrival times. Denoting the mean

G.L. Curry, R.M. Feldman, *Manufacturing Systems Modeling and Analysis*, 2nd ed.,
DOI 10.1007/978-3-642-16618-1_3,

arrival rate as λ, the mean inter-arrival time is then $1/\lambda$. If the system is full, the arriving job is rejected (and lost to another factory). If there is room in the waiting area, the arriving job is accepted and processed in a first-come-first-serve order (this sequence is denoted by FIFO which stands for first-in first-out). The processing time is also assumed to be exponentially distributed, with mean rate μ (the mean service time is $1/\mu$).

Since this system can have at most $n_{\max}$ jobs, there are $n_{\max}+1$ possible states, $\{0,1,\cdots,n_{\max}\}$, representing the number of jobs in the system. Interest is in developing the steady-state distribution of the number of jobs in the system. Assuming that a steady-state exists, then the flow into and out of each state must balance. This balance is the key property used to establish the steady-state probability of being in each possible system state.

Let p_n denote the steady-state probability of n jobs in the system for $n = 0,\cdots,n_{\max}$. The flow *into* an intermediate state n $(0 < n < n_{\max})$ is made up of two components: (1) the arrival of a new job to the system when the system has exactly $n-1$ jobs, and (2) the completion of a job's service when the system has exactly $n+1$ jobs. The steady-state flow *out* of an intermediate state n $(0 < n < n_{\max})$ is also made up of two components: (1) the completion of a job's service when the system has exactly n jobs, and (2) the arrival of a new job to the system when there are exactly n jobs in the system prior to the arrival event.

The resulting flow balance equation for state n is made up of the above four components. The mean arrival rate of jobs into the system is λ and the mean service rate of jobs when there is at least one job in the system is μ. The flow into state n occurs at the rate λ times the probability that the system is in state $n-1$ plus the rate μ times the probability that the system is in state $n+1$. Similarly, the flow out of state n occurs with rate $(\lambda+\mu)$ times the probability that the system is in state n. Thus, the steady-state flow-balance equation for an intermediate state n is

$$\lambda p_{n-1} + \mu p_{n+1} = (\lambda+\mu)p_n \text{ for } n = 1,\cdots,n_{\max}, \tag{3.1}$$

where the left-hand-side is the inflow and the right-hand-side is the outflow.

States 0 and $n_{\max}$ have different equations since some of the terms of the intermediate states equation are not valid for these boundary states. For example, the service rate is zero if there are no jobs in the system (state 0) nor can the system reside in state -1 so that an arrival event will put it into state 0. Also if the system is full (state $n_{\max}$), then no service from state $n_{\max}+1$ can occur and no new jobs are allowed to enter the system. The two special flow-balance equations (for states 0 and $n_{\max}$) are

$$\mu p_1 = \lambda p_0 \tag{3.2}$$

and

$$\lambda p_{n_{\max}-1} = \mu p_{n_{\max}}. \tag{3.3}$$

These three equations (namely, 3.1, 3.2, and 3.3) specify $n_{\max}+1$ equations connecting the state probabilities p_n. In addition, it is also known that the sum of these probabilities must add to one. Thus, there exists the additional equation, called the

norming equation, written as

$$\sum_{n=0}^{n_{\max}} p_n = 1 \,. \tag{3.4}$$

It turns out that the system is over-specified; that is, Eqs. (3.1–3.4) contain more equations than unknowns. To solve the system, any one of the equations can be omitted *except* for the norming equation. (The reader is asked to consider this point further in Problem 3.6.) After (arbitrarily) eliminating one equation from the system comprised of (3.1–3.3), there will be a total of $n_{\max}+1$ linear equations in $n_{\max}+1$ unknowns from the system defined by (3.1–3.4).

Given the mean arrival rate λ, the mean service rate μ and a system limit of $n_{\max}$, the resulting $n_{\max}+1$ linear equations can be solved by standard numerical methods. If $n_{\max}$ is not large, the equations can be written explicitly and solved for the specified values of λ and μ. However, because the system (3.1–3.4) has a fairly simple structure, it can be also be solved in general by a recursive substitution scheme and a closed form solution obtained. Not all systems that we develop in this text will have a structure leading to a general solution, but when this can be accomplished, it is the preferred method since the values of the parameters λ, μ and $n_{\max}$ need not be specified and a parametric solution for all values (or acceptable ranges of these parameter values) is obtained when solving the general system. For illustrative purposes, the system (3.1–3.4) is solved by both methods.

Example 3.1. **Specific Solution.** Consider a facility with a single machine that is used to service only one type of job. The company policy is to limit the number of orders accepted at any one time to 3. The mean arrival rate of orders, λ, is 5 jobs per day, and the mean processing time for a job is 1/4 day (thus, the processing rate is $\mu = 4/\text{day}$). Both the processing and inter-arrival times are assumed to be exponentially distributed. These assumptions lead to the system of equations

$$\begin{aligned} 4p_1 - 5p_0 &= 0 \\ 5p_0 + 4p_2 - (5+4)p_1 &= 0 \\ 5p_1 + 4p_3 - (5+4)p_2 &= 0 \\ 5p_2 - 4p_3 &= 0 \\ p_0 + p_1 + p_2 + p_3 &= 1 \,. \end{aligned}$$

We ignore the fourth equation and only use the first three equations plus the fifth (norming) equation to obtain

$$(p_0, p_1, p_2, p_3) = (0.173, 0.217, 0.271, 0.339) \,.$$

(See the appendix for using Excel to solve linear systems of equations.) The number of lost jobs per hour (i.e., those arriving to a full system) is given by $\lambda p_3 = 5 \times 0.339 = 1.695$. The server is idle when the system is empty, so the percentage of server idle time is 17.3%. Because the system is at steady-state, the throughput is equal to the number of jobs that enter the system per unit time (those jobs that actually get into the system, called the effective arrival rate). Thus, throughput rate

equals the arrival rate minus the loss rate; namely, 5 - 1.695 = 3.305 jobs/day. Note that

$$WIP = E[N] = \sum np_n = 1 \times 0.217 + 2 \times 0.271 + 3 \times 0.339 = 1.776 \text{ jobs} ,$$
$$CT = WIP/th = WIP/(\lambda(1-p_3)) = 1.776/3.305 = 0.537 \text{ days} .$$

□

Example 3.2. **General Solution.** To illustrate the more general solution approach, this system of equations is solved using the parameters rather than their actual values. The system to be solved is

$$\begin{aligned} \mu p_1 - \lambda p_0 &= 0 \\ \lambda p_0 + \mu p_2 - (\lambda + \mu) p_1 &= 0 \\ \lambda p_1 + \mu p_3 - (\lambda + \mu) p_2 &= 0 \\ \lambda p_2 - \mu p_3 &= 0 \\ p_0 + p_1 + p_2 + p_3 &= 1 . \end{aligned}$$

As before, the first three equations and the fifth equation will be used. The solution procedure is a two-step process. First, all variables are expressed in terms of p_0 by use of the first three equations. This is accomplished through a series of successive substitutions. Second, the value of p_0 is obtained by the use of the norming equation. Specifically, the first equation yields p_1 in terms of p_0 by

$$\begin{aligned} \mu p_1 &= \lambda p_0 \\ p_1 &= \frac{\lambda}{\mu} p_0 . \end{aligned}$$

The variable p_2 is obtained as a function of p_0 by substituting the expression for p_1 into the second equation as

$$\begin{aligned} \lambda p_0 + \mu p_2 &= (\lambda + \mu) p_1 \\ \mu p_2 &= (\lambda + \mu) p_1 - \lambda p_0 \\ p_2 &= (\lambda + \mu) \frac{\lambda}{\mu^2} p_0 - \frac{\lambda}{\mu} p_0 \\ p_2 &= \left(\frac{\lambda}{\mu}\right)^2 p_0 . \end{aligned}$$

Similarly, the third equation is used to obtain p_3 as a function of p_0 by substituting the expressions for the previously obtained p_1 and p_2; namely,

$$\begin{aligned} \lambda p_1 + \mu p_3 &= (\lambda + \mu) p_2 \\ p_3 &= (\lambda + \mu) \frac{\lambda^2}{\mu^3} p_0 - \left(\frac{\lambda}{\mu}\right)^2 p_0 \end{aligned}$$

$$p_3 = \left(\frac{\lambda}{\mu}\right)^3 p_0 \,.$$

The conclusion from the first step is that all probabilities are now in terms of p_0; namely,

$$p_1 = \left(\frac{\lambda}{\mu}\right) p_0, \qquad p_2 = \left(\frac{\lambda}{\mu}\right)^2 p_0, \qquad p_3 = \left(\frac{\lambda}{\mu}\right)^3 p_0 \,. \tag{3.5}$$

The final step is to substitute these expressions into the norming equation as follows:

$$\begin{aligned} 1 &= p_0 + p_1 + p_2 + p_3 \\ &= \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3\right] p_0 = 1 \end{aligned}$$

thus

$$p_0 = \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3\right]^{-1} \,. \tag{3.6}$$

From here we can develop the measures of $WIP = p_1 + 2p_2 + 3p_3$, $th = \lambda(p_0 + p_1 + p_2)$, and $CT = WIP/th$. □

Before moving to the remainder of the chapter, it is beneficial to formally define the effective arrival rate and comment on Little's Law. Whenever the system is finite, there is the possibility that the system will be full and arriving jobs will be lost; hence, the actual rate of jobs that enter the system, λ_e may not be the same as the arrival rate, λ.

Definition 3.1. The *effective arrival rate* for a system is the rate at which jobs enter the system. For a workstation with constant arrival rate, λ, and with a maximum number of jobs at the workstation limited to $n_{\max}$, the effective arrival rate is given by

$$\lambda_e = \lambda(1 - p_{n_{\max}})$$

where $p_{n_{\max}}$ is the probability that the workstation is full.

A system at steady-state will have its system throughput rate equal to the effective arrival rate; that is, $th = \lambda_e$, and the use of Little's Law (Property 2.1) must always use λ_e and not λ for the throughput.

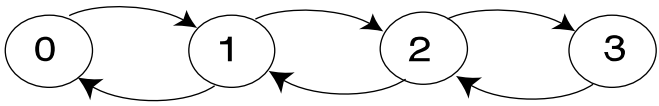
- *Suggestion: Do Problem 3.1.*

3.2 Diagram Method for Developing the Balance Equations

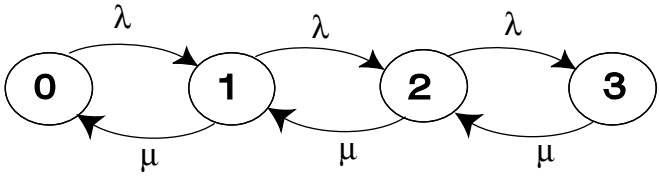
There is a relatively straightforward method for developing the balance equations for essentially any system in steady-state whose inter-arrival and service times are

exponentially distributed. The approach is to start by listing all of the states as nodes in a network. For the single-server problem, a sequential listing is the best. As one develops an understanding of this approach, a suitable layout will be apparent. The node listing is

Now directional arcs are added to the network to represent possible flows between nodes (states). For instance, node 0 is connected to node 1 to represent the flow from state 0 to 1 when an arrival occurs and the system is in state 0. Similarly, node 1 is connected to node 0 to represent the flow when a service occurs with the system in state 1 (a service results in an empty system or state 0). States 1 and 2 are connected, with a directed arc from 1 to 2, by an arrival event while in state 1. Conversely, states 1 and 2 are connected by a service event while in state 2; thus, the directed arc is from 2 to 1. The same logic connects states 2 and 3. So the following directed network is obtained. Note that an arrival into the system cannot occur when the system is in state 3 (i.e., when the system is full).

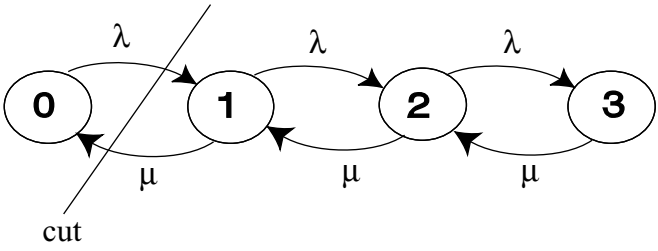


Now that the appropriately directed arc network of the system being modeled has been developed, the actual flow rates can be displayed on theses arcs. These rates are relatively straightforward to determine. Since the system has an arrival process that does not depend on the state of the system (excluding when it is full and so no arrivals can occur), the upward movements among the states all occur at a rate λ times the probability of being in that state, p_n. That is, the conditional arrival rate given that the system is in state n is λ and the net upward rate from state n is λp_n. The downward movements all occur when a service has been completed and these have rates that are μ times the probability of being in the particular state, p_n. Thus, the conditional service rate given that there is a job in the system to be serviced is μ. The resulting downward rates from state n is μp_n. The similarity of the service rates is again due to the assumption about the system. There is a single server and the service rate is independent of the state of the system. That is, the server works at the same rate without regard to the number of jobs in the queue. The standard method of graphically depicting the flow between states is to label the flow (arrows) with the conditional rates for that state.



This completed directed network can now be used to derive the steady-state balance equations previously analyzed. The logic goes as follows. Partition the nodes into two subsets of nodes, then establish values for the appropriate steady-state probabilities to balance the flow between the two subsets. Partitions are redrawn at $n-1$ different locations to obtain $n-1$ equations. These balance equations are then combined with the norming equations to yield a system of equations similar to the system of (3.1–3.4).

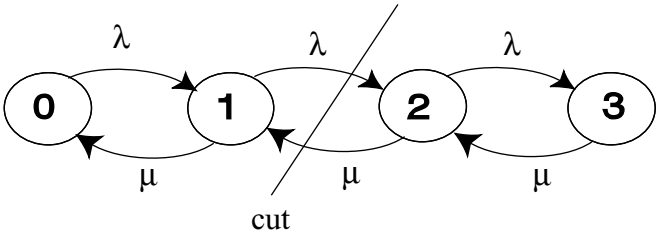
Consider the two subsets of nodes formed when a cut is made between nodes 0 and 1 as is illustrated below.



The balance equation associated with this initial cut is

$$\lambda p_0 = \mu p_1 \, .$$

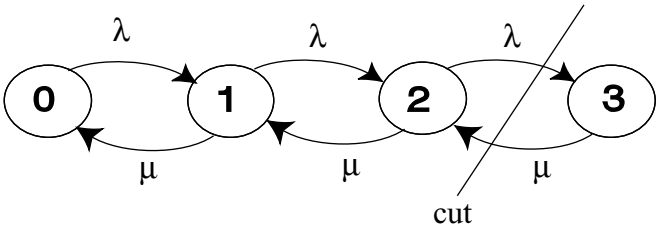
The second cut is between states 1 and 2.



The resulting balance equation associated with this cut is

$$\lambda p_1 = \mu p_2 \, .$$

The final cut is between states 2 and 3 as depicted below.



Thus the third balance equation is

$$\lambda p_2 = \mu p_3 \, .$$

These three-balance equations and the norming equation yield another representation for our modeled system as

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ \lambda p_1 &= \mu p_2 \\ \lambda p_2 &= \mu p_3 \\ \sum_{n=0}^{3} p_n &= 1. \end{aligned} \tag{3.7}$$

The system (3.7) obviously has the same relationships between the probabilities as (3.1–3.4); however, there is usually less work in obtaining this system using the flow balance approach. Successive substitution can then be used with (3.7) to obtain (3.5) and the norming equation yields the value for p_0 as was accomplished with (3.6).

Another subset partition that leads to the same system of equations is obtained by separating each node into its own singleton subset. The other subset contains all the other nodes of the network. The associated balance equations for each node arise when considering the input arcs to the node and balancing those rates with the outflow arcs. The development of this set of balance equations parallels the discussion in Sect. 3.1 and is left as an exercise for the reader (Problem 3.2).

The labeled directed arc network and partitioning method is a powerful methodology for deriving balance equations for queueing systems with exponentially distributed inter-arrival and service times. It is a useful method that helps one visualize the relationships in the system and keep track of the associated derived balance equations as they are being developed. Extensive use is made in this textbook of the labeled-directed arc-diagram approach for studying factory models.

3.3 Model Shorthand Notation

The models studied to this point all assumed exponentially distributed inter-arrival and service mechanisms. There is a notational shorthand due to Kendall [6] for characterizing queueing models that is quite useful. With essentially one word, the model assumptions and system behavior can be summarized. This notation, or variants of it, frequently appear in the queueing theory literature, particularly in paper titles. This system does not encompass all model variations imaginable, but it does present a great deal of information about the system in concise notation. The Kendall notation for queues is a list of characters each separated by a "/". The first element in the list specifies the inter-arrival time distribution assumption. The symbol M (for Markovian) depicts exponentially distributed times. The second element in the list denotes the service time distribution assumption. The third element in the list specifies the number of servers and the fourth element is the maximum number of jobs

allowed in the system at one time. An optional fifth element specifies the assumption for the queueing discipline. The general form for Kendall's notation is

$$\left(\begin{array}{c}\text{arrival}\\ \text{process}\end{array}\Big/\begin{array}{c}\text{service}\\ \text{process}\end{array}\Big/\begin{array}{c}\text{number}\\ \text{of servers}\end{array}\Big/\begin{array}{c}\text{maximum}\\ \text{possible}\\ \text{in system}\end{array}\Big/\begin{array}{c}\text{queue}\\ \text{discipline}\end{array}\right)$$

with Table 3.1 providing a summary of the commonly used abbreviations. Thus, the example queueing system just studied is denoted as an $M/M/1/3$ system. The two server model of Problem 3.3 is denoted by $M/M/2/3$. If the system has no effective limit on the number of jobs allowed, then the fourth parameter would be infinity. Most often the fourth parameter is omitted when it is not finite, so that such a model would often be written as $M/M/1$ instead of $M/M/1/\infty$.

Table 3.1 Queueing symbols used with Kendall's notation

Symbols	Explanation
M	Exponential (Markov) inter-arrival or service time
D	Deterministic inter-arrival or service time
E_k	Erlang type k inter-arrival or service time
G	General inter-arrival or service time
$1, 2, \cdots, \infty$	Number of parallel servers or capacity
FIFO	First in, first out queue discipline
LIFO	Last in, first out queue discipline
SIRO	Service in random order
PRI	Priority queue discipline
GD	General queue discipline

As the need arises, other parameter designations will be defined such as D for a deterministic time and G for a general distribution. To illustrate this notation, some of the most fundamental results needed for studying factory performance are the $G/G/1$ model approximations that are taken up at the end of this chapter.

- *Suggestion: Do Problems 3.2–3.6.*

3.4 An Infinite Capacity Model ($M/M/1$)

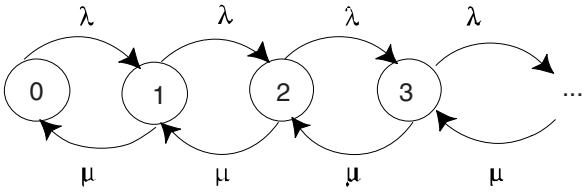
The finite capacity limitation on the $M/M/1/3$ model just studied is easily dropped, and the removal of this limitation has some interesting consequences. First note that the system of equations derived above (i.e., with a finite capacity) has a solution regardless of the relationship between the arrival rate and the system service rate. If the arrival rate of jobs to the system is larger than the system service capacity, the system is full a relatively high proportion of the time. This in turn leads to more jobs being turned away because of the full system. In fact, the effective arrival rate (those jobs getting into the system) will necessarily be less than the system's service

capacity. Let's consider a few cases for the above example that illustrate this point. Suppose that the mean arrival rate is equal to the mean service rate, $\lambda = \mu$ for the $M/M/1/3$ system. With $\lambda = \mu$, each probability is equal so that $p_0 = \cdots = p_3 = 1/4$. The effective arrival rate is, thus, given by $\lambda_e = \lambda(1 - p_3) = (3/4)\lambda < \mu$. If the mean arrival rate is twice the mean service rate, $\lambda = 2\mu$, then the effective arrival rate becomes $\lambda_e = (7/15)\lambda < \mu$. For a mean arrival rate that is three times the mean service rate, $\lambda = 3\mu$, the effective arrival rate becomes $\lambda_e = (13/40)\lambda < \mu$. Note that as the ratio of λ/μ becomes larger, the effective arrival rate approaches the inverse of this ratio but never reaches it. The reader is asked to compute these effective rates in Problem 3.5.

One of the lessons to be learned from the finite capacity model is that these systems have a built-in mechanism to adjust the arrival rate (called the effective arrival rate) to a level that can be handled by the system service capacity. If a system that has no realistic limit on the number of jobs allowed is considered, then mathematically, these systems can be put in a situation where the mean arrival rate exceeds the mean service rate and no steady-state exists. It is unreasonable to assume that jobs continue to arrive when there is essentially an infinite queue and the expected cycle time is also infinite. Of course, one would like to operate well below the blowup point with respect to the arrival and service capacity ratio. The analyses of the unlimited queueing models result in conditions that establish the existence of the steady-state behavior for these models.

The formulation of the unlimited-jobs system is very analogous to the finite capacity model formulation. The solution procedure is considerably different in that an infinite number of states exist and, correspondingly, an infinite number of descriptive equations result. Thus, standard numerical solutions for linear equations cannot be used. One is forced to solve these systems in a fashion analogous to the parametric solution approach illustrated for the finite capacity systems. This method is essentially substitution and formulation of a recursive relationship for the general solution structure.

The set of equations for the $M/M/1$ system is the same as the equations for the finite system capacity case except that the system does not have a final equation. Thus, an infinite system of equations exists. The diagram for this system is depicted below.



Using the cut partitioning method for obtaining the system of equations needed in defining the steady-state probabilities, the following is obtained:

$$\begin{aligned}
\lambda p_0 &= \mu p_1 \\
\lambda p_1 &= \mu p_2 \\
\lambda p_2 &= \mu p_3 \\
&\vdots \\
\lambda p_n &= \mu p_{n+1} \\
&\vdots \\
\sum_{n=0}^{\infty} p_n &= 1 \, .
\end{aligned}$$

The above system can be rewritten to obtain the following equivalent system.

$$\begin{aligned}
p_1 &= \tfrac{\lambda}{\mu} p_0 \\
p_2 &= \tfrac{\lambda}{\mu} p_1 \\
p_3 &= \tfrac{\lambda}{\mu} p_2 \\
&\vdots \\
p_n &= \tfrac{\lambda}{\mu} p_{n-1} \\
&\vdots
\end{aligned}$$

Using a successive substitution procedure, each p_n term can be written as a function of p_0 to obtain

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \;\text{ for } n = 0, 1, \cdots \, . \tag{3.8}$$

The final step is to substitute (3.8) into the norming equation yielding

$$p_0 + \left(\frac{\lambda}{\mu}\right) p_0 + \left(\frac{\lambda}{\mu}\right)^2 p_0 + \cdots + \left(\frac{\lambda}{\mu}\right)^n p_0 + \cdots = 1 \, ,$$

which can be solved to obtain an expression for p_0 as

$$p_0 = \frac{1}{\left(1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \cdots + \left(\frac{\lambda}{\mu}\right)^n + \cdots\right)} \, .$$

The denominator is a geometric series[1] that has a finite value if $\lambda/\mu < 1$. Under the condition that $\lambda < \mu$, this series sums to

$$p_0 = 1 - \frac{\lambda}{\mu} \, , \tag{3.9}$$

[1] The geometric series is $\sum_{n=0}^{\infty} r^n = 1/(1-r)$ for $|r| < 1$. Taking the derivative of both sides of the geometric series yields another useful result, $\sum_{n=1}^{\infty} n r^{n-1} = 1/(1-r)^2$ for $|r| < 1$.

and the general solution to the steady-state probabilities is (given that $\lambda/\mu < 1$)

$$p_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \quad \text{for } n = 0, 1, \cdots . \tag{3.10}$$

The throughput rate per unit time for this system is λ. (The reader is asked to develop this result in Problem 3.10.) The utilization factor u for the server is obtained from

$$u = 0p_0 + 1\left(\sum_{n=1}^{\infty} p_n\right) = 1 - p_0 = 1 - \left(1 - \frac{\lambda}{\mu}\right) = \frac{\lambda}{\mu} .$$

The expected number of jobs in the system in steady-state is obtained by using the derivative of the geometric series as follows:

$$\begin{aligned} WIP_s = E[N] &= \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \\ &= \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)\sum_{n=1}^{\infty} n\left(\frac{\lambda}{\mu}\right)^{n-1} \\ &= \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)\left(\frac{1}{1-\frac{\lambda}{\mu}}\right)^2 \\ &= \frac{\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)}{\left(1-\frac{\lambda}{\mu}\right)^2} = \frac{\frac{\lambda}{\mu}}{\left(1-\frac{\lambda}{\mu}\right)} = \frac{u}{1-u} \end{aligned} \tag{3.11}$$

where N is a random variable denoting the number of jobs in the system. Using Little's Law (Property 2.1), the expected time in system (the cycle time) CT_s is given by

$$CT_s = \frac{WIP_s}{\lambda} = \frac{1}{\lambda}\frac{\frac{\lambda}{\mu}}{(1-\frac{\lambda}{\mu})} = \frac{1}{\mu - \lambda} . \tag{3.12}$$

Example 3.3. Consider a single server system with exponentially-distributed interarrival times and exponentially-distributed service times (thus, this is an $M/M/1$ system). If 4 jobs per hour arrive for service ($\lambda = 4$) and the mean service time is 1/5 hour ($\mu = 5$), then the utilization factor u ($u = \lambda/\mu$) equals 0.8. The expected number of jobs in the system, WIP_s from (3.11) is

$$WIP_s = \frac{0.8}{(1-0.8)} = 4 .$$

The cycle time in the system, CT_s, is given by (3.12) and is

$$CT_s = \frac{1}{5-4} = 1 \text{ hr} .$$

The cycle time in the system is the sum of the cycle time in the queue plus the service time. Hence, $CT_q = 1 - 0.2 = 0.8$ hr. The probability that the server is idle, of course, equals the probability that the system is empty, p_0. This probability is

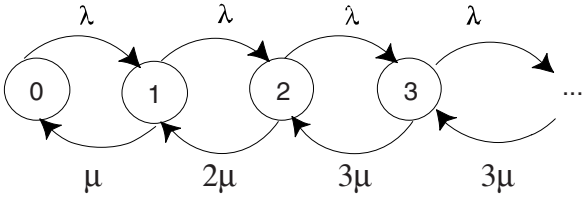
$$p_0 = 1 - \frac{\lambda}{\mu} = 0.2 \, .$$

The steady-state probability that there are n jobs in the system is given by

$$p_n = 0.2 \times 0.8^n \text{ for } n = 0, 1, \cdots \, .$$

□

A workstation may consist of multiple machines; however, in most models, server or machine distinctions are not usually made. That is, if there are two machines available, then for ease of modeling it is usually assumed that these are identical machines and that jobs are not split, but processed completely on one machine. Under the assumption of identical machines, if one machine operates at a rate of μ, then n machines operate at a rate of $n\mu$, and the state diagram must be adjusted accordingly. For example, suppose a workstation has three machines, then the service rate when two machines are busy is 2μ and whenever all machines are busy the service rate is 3μ; thus, the rate diagram is as below.



- *Suggestion: Do Problems 3.7–3.14.*

3.5 Multiple Server Systems with Non-identical Service Rates

The assumptions of identical machines may not be accurate, and if there is a significant difference in the operating characteristics of the machines associated with a single workstation, more complex models will result. To provide some exposure to the complexity involved in modeling non-identical machines within a single workstation, a simple non-identical servers model is considered and the associated defining equations for the steady-state probabilities are developed. The structure of this system is that it has two non-identical servers and a limit of four jobs in the system at one time. Inter-arrival and service times are all assumed to be exponentially distributed with a mean arrival rate of λ and mean service rates of μ and γ for the two distinct machines. Let $\gamma < \mu$, so that the μ machine is faster and, therefore,

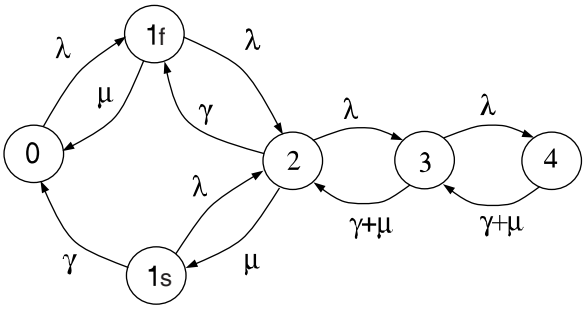


Fig. 3.1 State diagram for an $M/M/2/4$ system with non-identical servers, where μ denotes the rate of the faster machine and γ is the rate of the slower machine

preferred. The system operating policy is such that when the system is empty, an arriving job is always assigned to the faster machine. If a job arrives to the system and finds that only one machine is busy, the job is assigned to the idle machine immediately regardless of the speed of the machine or how long the other machine has been busy. This same logic is applied when a machine completes service and there is a queue of waiting jobs. The next job in line is immediately allocated to the idle machine; thus, machines can never be idle when there is a queue of waiting jobs. A final assumption is that once a job is assigned to a machine for processing, it remains on that machine until its processing is complete. Hence, jobs cannot be split and processed on both machines nor can a job be moved from the slower to the faster machine.

As before, $n_{\max}$ is the maximum number of jobs allowed in the system (here $n_{\max} = 4$) so that there will be a total of $n_{\max} + 2$ possible states for this model. In the identical server model, there were $n_{\max} + 1$ possible states. The extra state arises because we must know which machine is busy when there is only one job at the workstation in order to know the service rate associated with the job in process. When there are two or more jobs in the system, both machines are busy and no distinction about the state needs to be made. Denoting the state (i.e., the number of jobs at the workstation) by n, one possible state space is the set $\{0, 1\text{f}, 1\text{s}, 2, 3, 4\}$, where $n = 1\text{f}$ indicates that one job is in the system and that job is being processed on the fast machine and $n = 1\text{s}$ indicates that one job is in the system and is being processed on the slow machine. Given these operational rules and notation, the state diagram of this system is displayed in Fig. 3.1.

The transition rates shown in the diagram of Fig. 3.1 are explained as follows. In any state (other than the maximum), the arrival of a job takes the system to the next higher state number. Both states 1f and 1s move to state 2 with a job arrival. An arrival to an empty system moves the state from 0 to state 1f because of the assumption that the faster machine is preferred. From state 2, the next state depends on which machine finishes first. If the faster machine finishes before the slower machine, the system has one job remaining and this job continues being processed on the slower machine; thus, the system ends up in state 1s. This occurs with rate

μp_2. With similar reasoning, it should be clear that if the slower machine completes its processing first, the system transitions to state 1f. The transition from 2 to 1f occurs at a rate of γp_2. Notice that the downward movement from state 2 occurs with rate $(\mu+\gamma)p_2$. Downward movement from state 3 to state 2 occurs with rate $(\mu+\gamma)p_3$ and, similarly, from state 4 to state 3 with rate $(\mu+\gamma)p_4$.

The defining equations for the steady-state probabilities are determined by taking cuts between states. A slight problem exists with defining a cut between states due to the multiplicity of state 1 (i.e., 1f and 1s). The general idea of a cut is to isolate a set of states from the remaining states. In a serial system this cut process is easily defined and leads to the number of equations necessary for uniquely defining the probabilities when combined with the norming equation. The diagram (Fig. 3.1) for this non-identical server system is non-serial and thus there are several more possibilities for the cuts. The actual cuts that are used in the final analysis must be chosen wisely so that all probabilities are defined. For our set, we shall establish five cuts such that a cut is placed immediately to the right of each node subset contained within the following set:

$$\{\ \{0\},\{0,1\mathrm{f}\},\{0,1\mathrm{f},1\mathrm{s}\},\{0,1\mathrm{f},1\mathrm{s},2\},\{0,1\mathrm{f},1\mathrm{s},2,3\}\ \}$$

thus producing the following five equations:

$$\begin{aligned}
\lambda p_0 &= \mu p_{1\mathrm{f}} + \gamma p_{1\mathrm{s}} \\
\lambda p_{1\mathrm{f}} &= \gamma p_2 + \gamma p_{1\mathrm{s}} \\
\lambda p_{1\mathrm{f}} + \lambda p_{1\mathrm{s}} &= (\gamma+\mu) p_2 \\
\lambda p_2 &= (\gamma+\mu) p_3 \\
\lambda p_3 &= (\gamma+\mu) p_4\ .
\end{aligned} \tag{3.13}$$

These equations, plus the norming equation,

$$p_0 + p_{1\mathrm{f}} + p_{1\mathrm{s}} + p_2 + p_3 + p_4 = 1$$

are six equations that can be solved to obtain the steady-state probabilities for this system.

Example 3.4. An overhaul facility for helicopters is open 24 hours a day, seven days a week and helicopters arrive to the facility at an average rate of 3 per day according to a Poisson process (i.e., exponential inter-arrival times). One of the areas within the facility is for degreasing one of the major components. There is only room in the facility for 4 jobs at any one time and there are two machines that do the degreasing. The newer of the two degreasing machines takes an average of 8 hours to complete the degreasing and the older machine takes 12 hours for the degreasing operation. Because of the large variability in helicopter conditions, all times are exponentially distributed. Thus, we have $\lambda = 3$ per day, $\mu = 3$ per day, and $\gamma = 2$ per day. The system of equations given by (3.13) become

$$\begin{aligned} 3p_0 - 3p_{1\mathrm{f}} - 2p_{1\mathrm{s}} &= 0 \\ 3p_{1\mathrm{f}} - 2p_2 - 2p_{1\mathrm{s}} &= 0 \\ 3p_{1\mathrm{f}} + 3p_{1\mathrm{s}} - 5p_2 &= 0 \\ 3p_2 - 5p_3 &= 0 \\ 3p_3 - 5p_4 &= 0 \\ p_0 + p_{1\mathrm{f}} + p_{1\mathrm{s}} + p_2 + p_3 + p_4 &= 1 \,. \end{aligned}$$

The solution to this system of equations is

$$p_0 = 0.288,\ p_{1\mathrm{f}} = 0.209,\ p_{1\mathrm{s}} = 0.118,\ p_2 = 0.196,\ p_3 = 0.118,\ p_4 = 0.071 \,.$$

The average number in the system is obtained by using the definition of an expected value; namely,

$$WIP_s = p_{1\mathrm{f}} + p_{1\mathrm{s}} + 2p_2 + 3p_3 + 4p_4 = 1.356$$

and the average number in the queue is obtained similarly,

$$WIP_q = p_3 + 2p_4 = 0.259 \,.$$

Note that for the average number in the queue, p_3 is multiplied by 1 because when there are 3 in the system, there is only 1 in the queue. Also, p_4 is multiplied by 2 because when there are 4 in the system, there are 2 in the queue. Average cycle times are obtained through Little's Law as

$$\begin{aligned} CT_s &= \frac{WIP_s}{\lambda_e} = \frac{1.356}{3 \times (1 - 0.071)} = 0.486 \text{ day} \\ CT_q &= \frac{WIP_q}{\lambda_e} = \frac{0.259}{3 \times (1 - 0.071)} = 0.093 \text{ day} \,. \end{aligned}$$

A couple of other measures that are sometimes desired by management are the number of busy processors (i.e., degreasers) and their utilization. The expected number of busy servers, $E[BS]$, is 1.097, and is obtained as

$$E[BS] = 1p_{1\mathrm{f}} + 1p_{1\mathrm{s}} + 2p_2 + 2P_3 + 2p_4 = 1.097 \,.$$

The system utilization factor u is the expected number of busy servers divided by the number of machines available

$$u = \frac{E[BS]}{2} = 0.5485 = 54.85\% \,.$$

Our final calculation is to obtain the average time needed for degreasing. Because of the preference given to using the faster machine, we would expect the average time to be closer to 8 hours than to 12 hours. To get an exact value, we take advantage of the fact that the time in the system equals the time in the queue plus

service time (Eq. 2.1); thus

$$E[T] = CT_s - CT_q = 0.486 - 0.093 = 0.393 \text{ days} = 9.4 \text{ hr}.$$

□

- *Suggestion: Do Problems 3.15–3.20.*

3.6 Using Exponentials to Approximate General Times

The exponential distribution is an extremely powerful modeling tool because of its lack of memory (Eq. 1.16 and Problem 1.24). That is, the rate of completion of the process does not change with elapsed time. So for systems with exponential times, it is not necessary to keep track of the elapsed inter-arrival time nor the elapsed service time. This allows the steady-state modeling approach to be used. To model more general systems, one fruitful approach is to approximate the general times by combinations of exponentials. Then the exponential rate modeling approach can still be applied by developing more complex state representations of the system.

The Erlang-k distribution (see p. 18 for a review of the Erlang) provides an excellent distribution to use for the expanded state modeling approach. The Erlang-k distribution is the sum of k independent and identical exponential distributions, so that it can be modeled as a serial k-node system, with each node referring to identical exponentials. Since the Erlang-k has a squared coefficient of variation given by $C^2 = 1/k$, it also allows modeling of processes that have less variation than the exponential distribution.

3.6.1 Erlang Processing Times

To illustrate the expanded state modeling approach, consider a single server system with exponential inter-arrival times having a mean rate λ and a processing time that is described by an Erlang-2 distribution with mean rate μ and thus mean time $1/\mu$. This Erlang-2 distribution will be modeled using two exponential nodes (or *phases*), where each node has a mean rate of 2μ. Since rates and times are reciprocals, the mean time spent in each node is $1/(2\mu)$. This gives the total time spent in the two nodes as $1/\mu$ (i.e., the sum of the two means) which is equal to the average time of the Erlang-2 processing time distribution. To further simplify this example, the number of jobs allowed into the system will be limited to three. Thus, we are interested in analyzing an $M/E_2/1/3$ system.

The idea of the expanded state space approach is to represent the non-exponential process by more than one node, where each individual node is exponential. Therefore, the service process will have two nodes representing the two phases of the Erlang-2 distribution. When a job begins its processing, it enters the node represent-

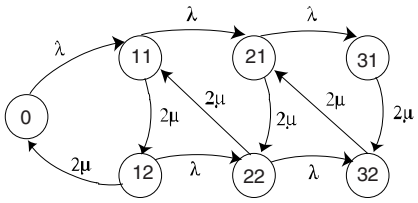


Fig. 3.2 Diagram for an $M/E_2/1/3$ model where the state (n,i) indicates that there are n jobs in the system with the i^{th} service phase busy

ing phase 1 and stays in phase 1 for an exponential length of time. When the job has been completed its phase 1 service, the job moves to the node representing phase 2. As long as the job is in either phase, it is considered to be continuing its processing and a new job is not allowed into service. When the job is finished with phase 2, it is considered to be finished with its processing and it leaves the system, and at this point in time, a new job can enter phase 1 to begin its service. A convenient representation for the state space is to use ordered pairs. In other words, (n,i) denotes a state of the system, where n is the number of jobs in the system and i is the service phase being occupied by the job being processed. The $M/E_2/1/3$ state diagram is displayed in Fig. 3.2.

There are $2n_{\max}+1$ states, where $n_{\max}$ is the maximum number of jobs allowed into the system (here $n_{\max}=3$). To obtain the steady-state probabilities for this system, six cuts are placed so that the following node sets are isolated on one side of the cut

$$\{\ \{0\},\{0,(1,2)\},\{0,(1,1)\},\{0,(1,1),(1,2)\},\{(3,1),(3,2)\},\{(3,2)\}\ \}$$

which together with the norming equation yields the following system of equations,

$$\begin{aligned}
\lambda p_0 - 2\mu p_{12} &= 0\\
\lambda p_0 + \lambda p_{12} - 2\mu p_{11} &= 0\\
(\lambda+2\mu)p_{11} - 2\mu p_{12} - 2\mu p_{22} &= 0\\
\lambda p_{11} + \lambda p_{12} - 2\mu p_{22} &= 0\\
\lambda p_{21} + \lambda p_{22} - 2\mu p_{32} &= 0\\
\lambda p_{22} + 2\mu p_{31} - 2\mu p_{32} &= 0\\
p_0 + p_{11} + p_{12} + p_{21} + p_{22} + p_{31} + p_{32} &= 1\ .
\end{aligned}$$

The performance measures of work-in-process, cycle time and throughput are computed from

$$\begin{aligned}
WIP_s &= \sum_{n=1}^{4} n(p_{n1}+p_{n2})\\
th &= \lambda_e = \lambda(1-p_{31}-p_{32})\\
CT_s &= WIP_s/\lambda_e\ .
\end{aligned}$$

3.6.2 Erlang Inter-Arrival Times

If the inter-arrival process is an Erlang distribution then the state-space scheme is slightly different from that used for Erlang service. The same concept of breaking the service process into phases is used for the arrival process; however, the state space will be slightly different. We illustrate the expanded state space process applied to arrivals by assuming an Erlang-2 inter-arrival time process. The arrivals will be processed one-at-a-time at a single workstation with exponentially distributed service times with a limit of three jobs in the system, in other words, we consider an $E_2/M/1/3$ system.

Conceptually, an arriving job is always in one of two phases, and each phase has a mean rate of 2λ or a mean sojourn time of $1/(2\lambda)$. As long as a job is in one of the arrival phases, it is not yet considered part of the system. The arriving job begins in phase 1. After an exponentially distributed length of time, the job transitions to phase 2. After another exponential length of time, two events occur simultaneously: the job leaves phase 2 and enters the system and another jobs enters phase 1. (Note that for a model of phased arrivals, one of the arrival phases is always occupied and the other phases are empty.)

The slight difference in the state space for the Erlang inter-arrival time model versus the Erlang service time model occurs due to the situation that the arrival process has two phases regardless of the number of jobs in the system. So when the system is empty, there are still two phases that the arriving job must complete before it becomes an active job attempting to enter the system. The state-space notation used is (i,n) where as before i is the phase and n is the number of jobs in the system. Note that the order has been reversed from the Erlang service model to help keep in mind that the phases are for the arrival process. The states needed to model the $E_2/M/1/3$ system are: $\{$(1,0), (2,0), (1,1), (2,1), (1,2), (2,2), (1,3), (2,3)$\}$. The diagram of this model is given in Fig. 3.3. Note also that there is a different situation for blocked jobs for this model. A job is not blocked until it arrives to a full system which occurs from state (2,3) with rate 2λ. Then the arrival process starts over in state (1,3) rather than staying at state (2,3). That is, the arriving job is rejected and the arrival process starts over at state (1,3) for the next job creation. Thus, there is an arc between (2,3) and (1,3) with rate 2λ in Fig. 3.3 to represent this transition.

Instead of using cuts to derive the equations of state, we use the single-node isolation method for generating the equations that define the steady-state probabilities. The following system of equations (all eight equations are given but only seven are used since the norming equation is also required) are generated for the states in the order that they appear in the above state list.

$$\begin{aligned} 2\lambda p_{10} &= \mu p_{11} \\ 2\lambda p_{20} &= 2\lambda p_{10} + \mu p_{21} \\ (2\lambda+\mu) p_{11} &= 2\lambda p_{20} + \mu p_{12} \\ (2\lambda+\mu) p_{21} &= 2\lambda p_{11} + \mu p_{22} \end{aligned}$$

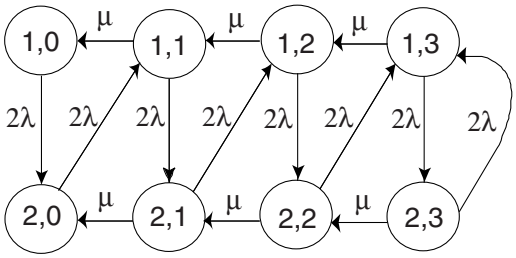


Fig. 3.3 Diagram for an $E_2/M/1/3$ model where the state (i,n) indicates that the arrival process is in phase i and there are n total jobs in the system

$$
\begin{aligned}
(2\lambda+\mu)p_{12} &= 2\lambda p_{21} + \mu p_{13} \\
(2\lambda+\mu)p_{22} &= 2\lambda p_{12} + \mu p_{23} \\
(2\lambda+\mu)p_{13} &= 2\lambda p_{22} + 2\lambda p_{23} \\
(2\lambda+\mu)p_{23} &= 2\lambda p_{13}
\end{aligned}
$$

and

$$p_{10} + p_{20} + p_{11} + p_{21} + p_{12} + p_{22} + p_{13} + p_{23} = 1\,.$$

Example 3.5. Since this system consists of only 8 unknowns, it is easily solved using the matrix formulas in Excel (see the appendix to this chapter). Let $\lambda = 5$ jobs/hr and $\mu = 5$ jobs/hr, and the solution to the $E_2/M/1/3$ system of equations is

$$
\begin{aligned}
p_{10} &= 0.0687\,, & p_{20} &= 0.1358\,, \\
p_{11} &= 0.1374\,, & p_{21} &= 0.1342\,, \\
p_{12} &= 0.1406\,, & p_{22} &= 0.1278\,, \\
p_{13} &= 0.1534\,, & p_{23} &= 0.1022\,.
\end{aligned}
$$

Some of the system performance measures are

$$
\begin{aligned}
WIP_s &= 0(p_{10}+p_{20}) + 1(p_{11}+p_{21}) + 2(p_{12}+p_{22}) + 3(p_{13}+p_{23}) = 1.5751 \\
u &= p_{11}+p_{21}+p_{12}+p_{22}+p_{13}+p_{23} = 1-(p_{10}+p_{20}) = 79.55\% \\
th &= \lambda_e = \lambda - 2\lambda p_{23} = \mu \times u = 3.978 \text{ jobs/hr} \\
CT_s &= WIP_s/th = 0.3960 \text{ hr}\,.
\end{aligned}
$$

Notice that the throughput can be calculated in a couple of different but equivalent ways. The expression $\lambda - 2\lambda p_{23}$ arrises by observing that arrivals are blocked from entering the system whenever the system is in the $(2,3)$ state and then the rate at which jobs leave state $(2,3)$ and try to enter the system is 2λ. Alternately, the throughput can be determined by multiplying the service rate μ times the probability that the server is busy, i.e., the utilization. □

- *Suggestion: Do Problems 3.21–3.24, and 3.33–3.36.*

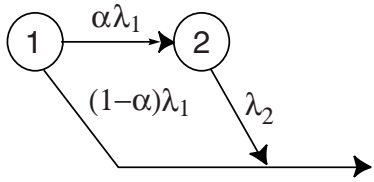


Fig. 3.4 A generalized Erlang with two phases, where the first phase always occurs and has a mean rate λ_1 and the second phase occurs with probability α and has a mean rate λ_2

3.6.3 Phased Inter-arrival and Processing Times

The improved modeling generality gained from the phased-service time model is frequently worth the notational inconvenience. For a phased-service time model, the state space is expanded essentially by a multiple of the number of phases. The state space for an $M/M/1/3$ system has four states $(n_{\max}+1)$, while its extension to the $M/E_2/1/3$ system has seven states $(2n_{\max}+1)$. The inter-arrival time process can also be broken into phases at the same time that the service times have phases to allow for even greater modeling flexibility, and the phases can be structured so as to be more general than the standard Erlang model. To illustrate the approach, the previous $M/E_2/1/3$ model is extended in this section to have a generalized Erlang-2 arrival process. There are two generalizations in the Erlang process that allow for a broader range of squared coefficients of variation, C^2, values while maintaining the essential exponential nature of individual nodes. The first generalization is to allow for non-identical phases and second is to give a probability that the process is complete at the end of each phase. Such a phased process is called a Generalized Erlang, *GE*, or a Coxian distribution. A *GE* with two phases is diagramed in Fig. 3.4.

A two-phase *GE* will be denoted by GE_2. Thus, the system of interest is an $GE_2/E_2/1/3$ model. The purpose of illustrating this generalization is to develop modeling skills that have more flexibility in the range of inter-arrival and service time distributions that can be studied. The distribution resulting from the GE_2 process illustrated in Fig. 3.4 can result in a squared coefficient of variation C^2 in the range $[0.5,\infty)$. Thus, the parameters of an GE_2 distribution can be selected to fit any finite mean and C^2 values needed, given that $C^2 \geq 1/2$. Notice that we have three parameters for the GE_2 distribution; namely, λ_1, λ_2, and α. It is possible to fix those three parameters to match a given mean, variance, and skewness for a distribution provided the skewness coefficient is not too large [2, p. 53]. However, it is more common to have only the mean and variance for a distribution. Parametric values for the GE_2 distribution have been suggested by Altiok [2, p. 54–56] when fitting the parameters to two moments. These are

$$\lambda_1 = \frac{2}{E[X]}\,, \quad \lambda_2 = \frac{1}{E[X]C^2[X]}\,, \quad \alpha = \frac{1}{2C^2[X]} \quad \text{for } C^2[X] > 1\,; \tag{3.14}$$

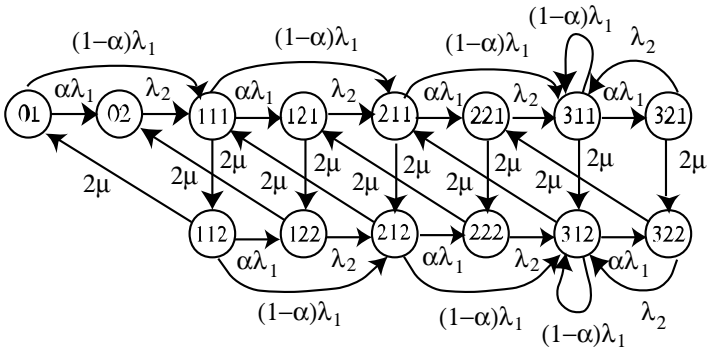


Fig. 3.5 State diagram for an $GE_2/E_2/1/3$ model, where a (n,i,j) indicates that there are n jobs in the system with one job in arrival phase i and one job is service phase j

$$\lambda_1 = \frac{1}{E[X]C^2[X]}\ ,\quad \lambda_2 = \frac{2}{E[X]}\ ,\quad \alpha = 2(1-C^2[X])\quad \text{for } \frac{1}{2} \le C^2[X] \le 1\ . \quad (3.15)$$

Note that matching two parameters of a distribution does not always characterize the distribution. Some distributions require three or more parameters for proper characterization, while the exponential distribution only requires one parameter (the mean rate λ or mean time $1/\lambda$).

Modeling with the GE_2 distribution causes these systems to quickly become quite complex. The $GE_2/E_2/1/3$ model, illustrated in Fig. 3.5, has 14 states, two states for each of the proceeding $M/E_2/1/3$ system states including the 0 state. The system empty state, state 0, now must be expanded so that the phase of the arriving job is represented. As one can readily see from the state diagram (Fig. 3.5) for this system, exponential-based generalizations for system times can be accomplished; however, these generalizations yield complex, and often intractable, models. The next section develops another approach for approximating general system time distributions (inter-arrival and service times).

- *Suggestion: Do Problems 3.25–3.29.*

3.7 Single Server Model Approximations

There are a variety of single facility generalizations that are standard in the queueing literature. Our concern is mainly with the assumptions regarding the inter-arrival and service time distributions. To use these models in a factory setting, more general assumptions on these distributions are needed. Rather than giving the general $G/G/1$ approximation model directly, a more circumspect route is taken that, hopefully, illuminates why and where the approximation arose. The model considered

next is the exact result for the $M/G/1$ queue, that in a proper form, suggests the structure of the general approximation result.

3.7.1 General Service Distributions

Consider a single-server system with exponential inter-arrival times, with mean rate λ, and a general service time distribution having mean time $1/\mu$ and variance σ_s^2. The state-diagram approach can no longer be used to develop equations that define the steady-state probabilities since these diagrams are tied to the exponential distribution or Markovian property. Variations such as Erlang service times can be developed using the state-diagram approach because the Erlang continues with the exponential assumption for the individual phases. The point of view taken for a general service process is to observe the system only at service completion times. This allows us to model, using the Markovian properties of the arrival process, the steady-state system size probabilities at departure points. It turn out that for this $M/G/1$ system, the steady-state probabilities at departure points are the same as the steady-state probabilities at an arbitrary point in time [4, p. 221]. The derivation of these probabilities is beyond the scope of this text and involves developing the generating function transform for the departure point probabilities. The development of the mean values for the number of jobs in the system was initially obtained independently in 1932 by Pollaczek and Khintchine and is now considered a standard property for general service time queueing systems.

Property 3.1. *The Pollaczek and Khintchine, or "P-K", formula for WIP in an $M/G/1$ queueing system is given by*

$$WIP_s = E[N] = \frac{\lambda}{\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^2 + \lambda^2\sigma_s^2}{2\left(1-\frac{\lambda}{\mu}\right)}$$

where N is the number of jobs in the system, λ is the mean arrival rate, and the service distribution has mean and variance given by $1/\mu$ and σ_s^2, respectively.

The notation used in the above property is common throughout this text. The subscript s used with WIP is to emphasize that the mean work-in-process is over the entire system; the subscript s used with the variance is to emphasize that the parameter refers to the service time distribution and is frequently used to differentiate the service distribution parameters from the inter-arrival parameters.

One implication of Little's Law is that for workstations that have one-at-a-time processing, the relationship between the average number in the system and the average number in the queue is given by $WIP_s - WIP_q = \lambda_e/\mu$. Since $\lambda_e = \lambda$ for $M/G/1$ systems, the expected number of jobs waiting for the processing, $E[N_q]$, is

$$WIP_q = E[N_q] = \frac{\left(\frac{\lambda}{\mu}\right)^2 + \lambda^2\sigma_s^2}{2\left(1-\frac{\lambda}{\mu}\right)} .$$

Using Little's Law one more time, the following important property is obtained, and this property will be used to develop approximations for more complicated systems.

Property 3.2. *The P-K formula for the queue cycle time in an $M/G/1$ system is given by*

$$CT_q = E[T_q] = \frac{WIP_q}{\lambda} = \frac{\left(\frac{\lambda}{\mu}\right)^2 + \lambda^2\sigma_s^2}{2\lambda\left(1-\frac{\lambda}{\mu}\right)}$$

where T_q is a random variable denoting the time a job spends in the queue, λ is the mean arrival rate, and the service distribution has mean and variance given by $1/\mu$ and σ_s^2, respectively.

The goal is now to rearrange this formula into a form that will be utilized a great deal in the development of more realistic factory models. First recall from (1.11) that the squared coefficient of variation is defined by

$$C^2[T] = \frac{V[T]}{E[T]^2}$$

so that in terms of service time distribution parameters, we can write

$$C_s^2 = \mu^2\sigma_s^2 .$$

Recall from (3.11) and (3.12) that the results for the $M/M/1$ model are

$$WIP_s(M/M/1) = \frac{u}{1-u} , \text{ and}$$
$$CT_s(M/M/1) = \frac{1}{\mu-\lambda}$$

where u is the server utilization factor and is equal to λ/μ. Here we have introduced a notational convention of writing the model assumptions (i.e., $M/M/1$) explicitly in the formula. This convention will be used whenever the context does not make the model clear. It should not be difficult to show (hint: use (2.1)) the following:

$$WIP_q(M/M/1) = \frac{u^2}{1-u} , \text{ and}$$
$$CT_q(M/M/1) = \frac{u}{1-u}E[T_s] \tag{3.16}$$

where T_s is a random variable denoting the time a job spends in the server.

The P-K formula for cycle time in the queue (Property 3.2) can be rewritten as

$$\begin{aligned} CT_q &= \frac{\left(\frac{\lambda}{\mu}\right)^2 + \lambda^2\sigma_s^2}{2\lambda\left(1-\frac{\lambda}{\mu}\right)} \\ &= \frac{\left(\frac{\lambda}{\mu}\right)^2 + \lambda^2\frac{C_s^2}{\mu^2}}{2\lambda\left(1-\frac{\lambda}{\mu}\right)} \\ &= \left(\frac{1+C_s^2}{2}\right)\left(\frac{u}{1-u}\right)E[T_s]\,. \end{aligned}$$

Thus, we have an extremely important (exact) relationship between the $M/G/1$ and the $M/M/1$ models; namely,

$$CT_q(M/G/1) = \left(\frac{1+C_s^2}{2}\right)CT_q(M/M/1)\,. \tag{3.17}$$

3.7.2 Approximations for $G/G/1$ Systems

The P-K mean queue cycle time result (3.17) is based on the assumption of exponential inter-arrival times. Since the coefficient of variation for the exponential distribution is one, the P-K result could just as accurately have been written as

$$CT_q(M/G/1) = \left(\frac{C_a^2+C_s^2}{2}\right)CT_q(M/M/1)\,,$$

where C_a^2 refers to the squared coefficient of variation for the inter-arrival times. This form suggests that the relationship might be a reasonable approximation for the general $G/G/1$ system. In fact, Kingman [7] looked at various approximations in heavy-traffic conditions (i.e., for utilization factors close to 1) and obtained a similar result. Therefore, our first approximation is named after Kingman.

Property 3.3. *The Kingman diffusion approximation for the $G/G/1$ queueing system is*

$$CT_q(G/G/1) \approx \left(\frac{C_a^2+C_s^2}{2}\right)CT_q(M/M/1)\,,$$

where C_a^2 and C_s^2 are the squared coefficients of variation for the inter-arrival distribution and the service time distribution, respectively.

There have been extensive studies using the Kingman diffusion approximation and it has been shown to be an upper bound on the actual mean queue cycle time.

An improved approximation was developed by Kraemer and Langenbach-Belz [8] and studies by Whitt [10] have shown that it is good when the inter-arrival time variability is less than the exponential distribution. Whitt's conclusion is to extend the approximation by adding another multiplicative term resulting in the following:

$$CT_q(G/G/1) \approx g(u, C_a^2, C_s^2) \times \left(\frac{C_a^2 + C_s^2}{2}\right) CT_q(M/M/1)\,, \tag{3.18}$$

where g is a function of server utilization and the two squared coefficients of variation defined as

$$g(u, C_a^2, C_s^2) = \begin{cases} \exp\{-\frac{2(1-u)}{3u}\,\frac{(1-C_a^2)^2}{C_a^2+C_s^2}\} & \text{for } C_a^2 < 1\,, \\ 1 & \text{for } C_a^2 \geq 1\,. \end{cases}$$

For the remainder of this textbook, the simple form of Kingman's diffusion approximation (Property 3.3) is used with the understanding that improvements are possible using Whitt's extension (3.18). Since the time in the system equals the time in the queue plus the processing time, we also have a good approximation for the system mean cycle time as

$$CT_s(G/G/1) \approx \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u}{1-u}\right) E[T_s] + E[T_s]\,. \tag{3.19}$$

Example 3.6. Consider again Example 3.3 illustrating an $M/M/1$ system. For this model, $\lambda = 4$/hr and $\mu = 5$/hr yielding a utilization factor $u = 0.8$. Since this was an exponential system, we had $C_a^2 = C_s^2 = 1$ and $E[T_s] = 0.2$ hr. Thus, the $G/G/1$ approximation is

$$CT_q(G/G/1) = \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u}{1-u}\right) E[T_s] = \left(\frac{1+1}{2}\right)\left(\frac{0.8}{0.2}\right) 0.2 = 0.8 \text{ hr}\,.$$

Whenever the Kingman approximation (Property 3.3) is applied to an $M/M/1$ or $M/G/1$ system, it is exact and not an approximation. We observe that the above result of 0.8 hr for the waiting time agrees exactly with CT_q as calculated in Example 3.3. (It is always nice to have consistency in mathematics!) □

Example 3.7. Consider a $G/G/1$ system with inter-arrival times distributed according to a gamma distribution with mean 15 minutes and standard deviation 30 minutes, and with service times distributed according to an Erlang-4 distribution with mean 12 minutes. Since the distribution of service times is Erlang, the initial temptation may be to use the methodology of Sect. 3.6.1; however, because the arrival times are not exponential, we are left with the $G/G/1$ results. The given data yields the following parameters: $\lambda = 4$/hr, $\mu = 5$/hr, $C_a^2 = 4$, and $C_s^2 = 0.25$. Thus, this example has the same mean characteristics of Example 3.6 yielding a utilization of $u = 0.8$, but the arrival process has more variability and the processing times are less variable. Using the Kingman diffusion approximation (Property 3.3), we have

$$CT_q(G/G/1) \approx \left(\frac{C_a^2+C_s^2}{2}\right)\left(\frac{u}{1-u}\right)E[T_s] = \left(\frac{4+0.25}{2}\right)\left(\frac{0.8}{0.2}\right)0.2 = 1.7 \text{ hr}\,.$$

This cycle time is over twice a large as the exponentially distributed system result; thus, the variability associated with non-exponential distributions can have a significant impact on the expected cycle time.

The queue waiting times for single-server queueing systems can be easily simulated with a spreadsheet model (see the Appendix); thus to check the accuracy of the approximation, we simulated the $G/G/1$ system using Excel as discussed in the appendix. (Also refer to the appendix for the importance of reporting confidence intervals along with simulation results.) The simulation yielded a mean waiting time of 1.89 hours with a half-width of ± 2 minutes for the 95% confidence interval. It is interesting that when a Weibull distribution with the same mean and variance was used instead of the Gamma distribution, the simulated mean waiting time was 1.71 hours with a half width of ± 1.5 minutes for the 95% confidence interval. □

3.7.3 Approximations for $G/G/c$ Systems

There are many generalizations of the $G/G/1$ approximations to account for multiple server systems in the literature. Allen and Cunneen [1] have one of the first commonly used approximation based on the Kingman diffusion approximation. Their approximation was later adjusted by Hall [3] to be a simple extension of Property 3.3 and is given as

$$CT_q(G/G/c) \approx \left(\frac{C_a^2+C_s^2}{2}\right)CT_q(M/M/c)\,. \tag{3.20}$$

This form of the multiple server approximation is particularly appealing and will be used herein since it reduces to the form of the single-server approximation when $c = 1$. In addition, it is not too difficult to obtain WIP and CT for an $M/M/2$ system (see Problem 3.9) and the $M/M/3$ system; thus, we have the following two properties.

Property 3.4. *The Kingman diffusion approximation extended for a two-server system is*

$$CT_q(G/G/2) \approx \left(\frac{C_a^2+C_s^2}{2}\right)\left(\frac{u}{1-u}\right)\left(\frac{u}{1+u}\right)E[T_s]\,,$$

where $u = \lambda E[T_s]/2$ is server utilization. This approximation is exact for the $M/M/2$ system.

Property 3.5. *The Kingman diffusion approximation extended for a three-server system is*

$$CT_q(G/G/3) \approx \left(\frac{C_a^2+C_s^2}{2}\right)\left(\frac{u}{1-u}\right)\left(\frac{3u^2}{2+4u+3u^2}\right)E[T_s]\,,$$

where $u = \lambda E[T_s]/3$ is server utilization. This approximation is exact for the $M/M/3$ system.

An approximation proposed in Hopp and Spearman [5] uses the following approximation for a Markovian multiple server system from [9]

$$CT_q(M/M/c) = \left(\frac{u^{\sqrt{2c+2}-2}}{c}\right)CT_q(M/M/1)\,.$$

The resulting approximation of Hopp and Spearman yields a general extension as:

Property 3.6. *The Kingman diffusion approximation extended for a multi-server system is*

$$CT_q(G/G/c) \approx \left(\frac{C_a^2+C_s^2}{2}\right)\left(\frac{u^{\sqrt{2c+2}-1}}{c(1-u)}\right)E[T_s]\,,$$

where $u = \lambda E[T_s]/c$ is server utilization.

Finally, we repeat the obvious rule for system cycle time (3.19) extended to a multiple-server system that holds whenever service is one-at-a-time:

$$CT_s(G/G/c) = CT_q(G/G/c) + E[T_s]\,. \tag{3.21}$$

Example 3.8. Consider again the system of Example 3.7 except for a two-server system and with a mean service time of 24 minutes. Thus, server utilization stays the same (namely, $u = 0.8$) and the squared coefficients of variation are still given as $C_a^2 = 4$ and $C_s^2 = 0.25$. Then the expected system cycle time using the approximation of Property 3.6 is

$$CT_q(G/G/2) \approx \left(\frac{4+0.25}{2}\right)\left(\frac{(0.8)^{\sqrt{6}-1}}{2(1-0.8)}\right)0.4$$

$$= 1.54 \text{ hr}\,.$$

If we use Property 3.4, the approximation becomes

$$CT_q(G/G/2) \approx \left(\frac{4+0.25}{2}\right)\left(\frac{0.8}{1-0.8}\right)\left(\frac{0.8}{1+0.8}\right)0.4$$
$$= 1.51 \text{ hr} .$$

A simulation of this system yielded a mean cycle time in the queue of 1.63 hr with a half-width of ± 0.01 hr for the 95% confidence interval. □

A comparison of the analytical result and the simulation result in the above example illustrates that these approximations are adequate but certainly not exact. Throughout the next four chapters, we will utilize these approximations extensively as we build approximations for more general factory models.

- *Suggestion: Do Problems 3.30–3.32.*

Appendix

In this appendix, we discuss using Excel to solve linear systems of equations and the use of confidence intervals within a simulation. We also present a very simple method for simulating a single-server queueing system with a FIFO queueing discipline.

Solutions to Linear Systems of Equations. Linear systems can always be written in matrix form as

$$A\mathbf{x} = \mathbf{b} ,$$

where A is an $m \times n$ matrix of the coefficients, $\mathbf{x}$ is a vector of n unknowns, and $\mathbf{b}$ is an m dimensioned vector of the right-hand-side constants. If the system has the same number of equations as unknowns (namely, $m = n$) and if the matrix A has an inverse, the solution to this system is

$$\mathbf{x} = A^{-1}\mathbf{b} ,$$

where A^{-1} denotes the inverse of the matrix. Excel has functions for both the matrix inverse and for matrix multiplication. The key to using an Excel function that has an array for the answer, is to highlight the area of the answer and use `<ctrl-shift-enter>` when executing the function. For example, suppose we wish to solve the following system:

$$\begin{aligned} 3x_1 + 4x_2 + 5x_3 &= 4 \\ 2x_1 + 2x_2 + 5x_3 &= 3 \\ 1x_1 + 6x_2 - 2x_3 &= 1 . \end{aligned}$$

Using Excel, type the coefficient matrix, A, in the square block of cells `A2:C4` and the right-hand-side vector in a single column block of cells `E2:E4` as shown below.

	A	B	C	D	E
1	Coefficient Matrix				RHS
2	3	4	5		4
3	2	2	5		3
4	1	6	-2		1

The solution to the system, namely $A^{-1}\mathbf{b}$ is a 3×1 array; therefore, a column of three cells for storing the answer must be selected (highlighted). Choosing the cells `G2:G4` for the answer, select those three cells by placing the mouse in cell `G2` and dragging the mouse down three cells. While the three cells are highlighted, type the following (the typing will be appear in cell `G2` since that is where the selection started)

```
=MMULT(MINVERSE(A2:C4),E2:E4)
```

but do not hit the <enter> key. Note that the `MMULT()` function multiplies two arrays, and the `MINVERSE()` function produces the inverse of an array. In Excel, matrix functions always begin with the letter `M`. When finished typing, hold down the <ctrl> and <shift> keys and while holding these two key down, hit the <enter> key. The answer (0.75, 0.125, 0.25) should appear in the highlighted cells `G2:G4`.

Simulation of Waiting Times in a Single-Server Workstation. Consider a G/G/1 queueing system in which each job is numbered sequentially as it arrives. Let the service time of the n^{th} job be denoted by the random variable S_n, the delay time (time spent in the queue) by the random variable D_n, and the inter-arrival time between the n-1^{st} and n^{th} job by the random variable A_n. The delay time of the n^{th} job must equal the delay time of the previous job, plus the previous job's service time, minus the inter-arrival time; however, if inter-arrival time is larger than the previous job's delay time plus service time, then the queueing delay will be zero. In other words, the following must hold

$$D_n = \max\{0,\, D_{n-1} + S_{n-1} - A_n\}\,. \tag{3.22}$$

If we can generate observations of the random variables A_n and S_n for $n = 1, \cdots, n_{\max}$ we will have simulated the arrival and service times for $n_{\max}$ jobs and thus be able to simulate their delays using (3.22). In the Appendix of Chap. 2, the Excel function `RAND()` was used to generate random numbers which are defined as a sequence of numbers appearing to have a continuous uniform distribution between 0 and 1. General random variates can be obtained by the following property that is used to relate random numbers to any other random variable.

Property 3.7. *Let R be a random variable with a continuous uniform distribution between zero and one, and let F be an arbitrary CDF. If the inverse of the function F exists, denote it by F^{-1}; otherwise, let $F^{-1}(a) = \min\{t|F(t) \geq a\}$. Then the random variable X defined by*

$$X = F^{-1}(R),$$

has a distribution function given by F; that is,

$$P\{X \le a\} = F(a) \quad for\ -\infty < a < \infty.$$

To illustrate the use of this property, consider the Excel function

GAMMAINV(*probability, shape_parameter, scale_parameter*)

that yields the inverse of the gamma CDF evaluated at the specified probability with the given shape, α, and scale, β, parameters (review p. 19); thus,

```
=GAMMAINV(RAND(),4,3)
```

will generate gamma random variates with mean 12 and standard deviation 6 (because the mean is the shape times scale and the variance is shape times scale squared).

To begin a simulation of Example 3.7, type the following in the first three rows of an Excel spreadsheet.

	A	B	C
1	InterArrive	Service	Delay
2	0	`=GAMMAINV(RAND(),4,3)`	0
3	`=GAMMAINV(RAND(),0.25,60)`	`=GAMMAINV(RAND(),4,3)`	`=MAX(0,C2+B2-A3)`

Notice that the references in the `C3` cell are relative references and that two of the references are to the previous row, but the third reference (`A3`) is to the same row. Also, remember that the Erlang distribution is a gamma distribution whose shape parameter is an integer. Now copy the third row down for several thousands of rows and obtain an average of the values in the `C` column. This average is an estimate for the mean cycle time. However, because of the large variablity in the inter-arrival times, the simulation needs to be repeated several times to obtain a good estimate. Reporting the simulation results together with an estimate of its variability is briefly discussed in the next few paragraphs.

Confidence Intervals. Simulations are statistical experiments; therefore, results should never be reported without giving some idea of the accuracy or variability of the statistical information. Assume there is a data set $\{x_1, \cdots, x_n\}$ containing n data points from independent and identically distributed observations. Our goal is to estimate the underlying true (but unknown) mean of the distribution that produced the data. For any data set, the sample mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.23}$$

and the sample variance is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right). \tag{3.24}$$

Since an estimate for the true mean is desired, the temptation may be to report the sample mean only; however, a single value will provide an estimate but it gives no information on the variability of the estimate. To include information about variability, a *confidence interval* is often used. For example, a 95% confidence interval for the mean implies that if the same experiment were repeated 100 times, approximately 95 of those confidence intervals would contain the true mean; that is, we expect to be correct approximately 19 out of 20 times.

Under the assumption of normally distributed data and unknown variance, the $1-\alpha$ confidence interval for the mean is given by

$$(\bar{x}_n - t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}},\ \bar{x}_n + t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}) \tag{3.25}$$

where $t_{n-1,\alpha/2}$ is a critical value based on the Student-t distribution. Statistical tests are usually better as the degrees-of-freedom increases. (As a rule of thumb, a statistical test loses a degree-of-freedom whenever a parameter must be estimated by the data set; thus, the t-test has only $n-1$ degrees-of-freedom instead of n because we use the data to estimate the variance.)

If using Excel, the function `=TINV(0.05, 24)` would yield the critical value for a 95% t-statistic for a sample of 25 data points. Notice that Excel automatically splits the error into a right-hand error and a left-hand error; thus, if it were desired to obtain the critical value for a 90% confidence interval of a sample of 100 points, the function `=TINV(0.10, 99)` would be used. (As an historical note: when statistical tables were primarily used to obtain the critical value for the statistics, the rule of thumb was to use the z-statistic for large sample sizes; however, with Excel, there is no reason to switch to the z-statistic since Excel does not have a problem with large sample sizes.)

When applying confidence intervals to simulations, care must be taken not to violate the independence assumption. Because sequential output from a simulation are usually correlated, it is best to form a random sample by performing several replicates of the same simulation, where each replicate starts with a different random number seed. The random sample for the confidence interval then comes from the summary statistics of each replicate.

Problems

3.1. Consider a facility open 24 hours per day with a single machine that is used to service only one type of job. The company policy is to limit the number of jobs within the facility at any one time to 4. The mean arrival rate of jobs is 120 jobs per day, and the mean processing time for a job is 15 minutes. Both the processing and inter-arrival times are assumed to be exponentially distributed. Answer the following questions regarding the long-run behavior of the facility.
(a) What is the average number of jobs that arrive to the facility (but not necessarily get in) per hour?

(b) What is the probability that there are no jobs at the facility?
(c) What is the average number of jobs within the facility?
(d) What is the average number of jobs lost per day due to the limited capacity of the facility?
(e) What is the average throughput rate per hour?
(f) What is the average amount of time, in minutes, that a job spends within the facility?

3.2. Consider a single server system with a limit of 3 jobs (an $M/M/1/3$ system). Let λ be the mean arrival rate and μ be the mean service rate.
(a) Use the singleton subset partition method to derive a system of balance equations (note the last equation is the probability norming equation):

$$\begin{aligned} \lambda p_0 - \mu p_1 &= 0 \\ \lambda p_0 + \mu p_2 - (\lambda + \mu) p_1 &= 0 \\ \lambda p_1 + \mu p_3 - (\lambda + \mu) p_2 &= 0 \\ \lambda p_2 - \mu p_3 &= 0 \\ p_0 + p_1 + p_2 + p_3 &= 1. \end{aligned}$$

(b) Use the subset partition between successive nodes to derive a system of balance equations.
(c) Solve for each p_i in terms of p_0 for each set of balance equations (a and b) to establish that they yield the same solution.

3.3. Consider a two-server system with exponentially distributed inter-arrival and service times. Let λ be the mean arrival rate and μ be the mean service rate of each server. The system has a limit of 3 jobs at any time. The servers work on jobs independently (only one server is working when there is only one job in the system).
(a) Develop the labeled directed arc network for this system.
(b) Write a system of equations, balance and norming equations, for this system.
(c) Solve this system for the general form of the steady-state probabilities.
(d) Write the equation for server utilization in terms of the steady-state probabilities.
(e) What is the mean number of jobs lost per unit time due to the limited system capacity?
(f) What is the system throughput rate? Note that throughput means completed jobs.

3.4. Consider a single-server system with two types of jobs. The system has a limited capacity of three total jobs in the system at any time. The job classes have different mean arrival and service rates, but all are assumed to be exponentially distributed. Let λ_1 be the mean arrival rate and μ_1 be the mean service rate of job type 1, and let λ_2 be the mean arrival rate and μ_2 be the mean service rate of job type 2. Job class 1 are high priority items and, as such, they have preemptive priority over jobs of type 2 on the server. Space within the system limit of three jobs is on a first-come first-service basis; thus, once a low-priority job is in the system, it cannot be replaced by a high-priority job. Although all low-priority jobs must wait until all high-priority jobs have been processed, even if they arrive when a low-priority job

is being serviced. Develop the labeled directed arc network for this system. Hint: there are ten different states and the number of each job type must be accounted for separately.

3.5. Consider the $M/M/1/3$ system of Problem 3.2 with an effective arrival rate given by the equation

$$\lambda_e = \lambda(1 - p_3).$$

Compute the effective arrival rate as a function of μ for the following situations:

λ	$\lambda = \mu$	$\lambda = 2\mu$	$\lambda = 3\mu$	$\lambda = 4\mu$
λ_e	?	?	?	?

3.6. Consider solving the set of steady-state equations for a system with a limit on the number of jobs allowed (example $M/M/1/3$). Suppose there are $n_{\max}$ steady-state equations derived from the flow-in equals flow-out approach. Show that if only these equations (omitting the norming equation) are used and if they are linearly independent, then the solution for p_n cannot satisfy the conditions for a pmf. This result leads to the conclusion that this set of equations must be dependent and, therefore, the norming equation must be used in place of one of the other equations.

3.7. Jobs arrive at a single machine for processing. Jobs arrive in groups of two (always) with an exponentially distributed time between groups with mean rate λ. The single server works on individual jobs. The service time is exponentially distributed with a mean rate μ. Let p_n be the probability that there are n jobs in the system in steady-state. Note that there is no limit to the number of jobs allowed into this system. Draw the state diagram with labeled arcs and write the steady-state equations for states 0, 1, 2, 3, 4, and 5. What is the relationship between λ and μ that guarantees that a steady-state exists?

3.8. Redo Problem 3.7 under the assumption that the group size is one with probability 1/2 and two with probability 1/2.

3.9. Consider a factory with a two-identical servers where jobs can be run on either of the two servers. All jobs have the mean-arrival rate of λ and the same mean-service rate μ, and both distributions are assumed to be exponential. Assume that there is no limit on the number of jobs allowed in the system. Thus, the system is an $M/M/2/\infty$ queue.
(a) Develop the steady-state diagram connecting the states of the system.
(b) Develop the system of equations that the steady-state probabilities must satisfy.
(c) Develop the general probability relationship for p_n in terms of p_0.
(d) Develop a formula for p_0. Hint: the appropriate service rate when both servers are busy is 2μ.

3.10. For the $M/M/1/\infty$ model, show that the expected output rate of jobs is equal to the mean input rate λ.

3.11. For the $M/M/1/\infty$ model derive, from the p_n's, an expression for the queue work-in-process WIP_q.

3.12. Using Little's Law, obtain the cycle time in the queue, CT_q, from the result of Problem 3.11.

3.13. The cycle time in the system is logically the cycle time in the queue plus the expected service time

$$CT_s = CT_q + E[T_s].$$

For the $M/M/1/\infty$ model derive an expression for CT_q using the CT_s result of Eq. (3.12).

3.14. Consider an $M/M/1/\infty$ system with a mean arrival rate of $\lambda = 5$ jobs per hour. Compute the system performance measures (WIP_s, CT_s, th_s, u) for several different service rates $\mu \in \{5.5, 6, 7, 8, 9, 10\}$. Graph the WIP_s and CT_s as a function of the system utilization factor u.

3.15. Determine the impact of an arrival rate of 5 per day in Example 3.4 ($\lambda = 5, \mu = 3, \gamma = 2$ in Eq. 3.13) as it reflects on the system parameters.
(a) Write the system of equations for the steady-state probabilities.
(b) Obtain the system performance measures: CT_s, CT_q, WIP_s, WIP_q, utilization u, mean service time $E[T_s]$, and throughput λ_e.

3.16. For a system with non-identical service rates (see Sect. 3.5) and a limit of N jobs in the system (Eq. 3.13), obtain an expression for the mean service time per job, $E[T_s]$, as a function of the mean throughput rate λ_e, the steady-state probabilities p_n and the mean-service rates μ and γ.

3.17. Solve Problem 3.16 for the probabilities given the parameters: $n_{\max} = 4$, $\lambda = 3$, $\mu = 3$, and $\gamma = 2$.

3.18. Consider a two-server system with non-identical machines, exponentially distributed inter-arrival and service times, and a limit of four jobs. The mean inter-arrival rate is λ. The mean service rates are $\gamma < \mu$. Jobs cannot be split across machines. When there is not a queue of waiting jobs and the faster machine completes processing first, the job on the slower machine is immediately moved to the faster machine to complete processing.
(a) Develop the steady-state diagram of the number of jobs in the system and the flow rates between states.
(b) Develop the system of equations describing the steady-state probabilities of being in each state.
(c) Solve this system of equations.

3.19. For Problem 3.18, obtain the system parameters: CT_s, CT_q, WIP_s, WIP_q, u, mean service time $E[T_s]$, the expected number of busy servers (EBS), and throughput th_s.

3.20. A workstation has two different machines for performing two distinct processing tasks. The workstation has one operator that performs all work done in the

workstation on all jobs. That is, the operator stays with a job and moves it from machine to machine to accomplish the necessary processing. Jobs arrive to the workstation at a mean rate λ (exponentially distributed inter-arrival times). Each job is first processed by the operator on Machine 1 which takes an exponentially distributed length of time with mean rate μ. Then the job and operator go to Machine 2 for further processing. The processing time on the second machine is also exponentially distributed but with a mean rate γ. The operator works on one job at a time and completes it before starting on a new job. The company limits the jobs in this workstation to 3.
(a) Define an appropriate state space representation for this model.
(b) Using your state space, develop a state diagram to model this situation.
(c) Write the utilization equation for machine one, using the state probabilities.
(d) Write the operator utilization equation, using the state probabilities.
(e) Write the workstation work-in-process equation, using the state probabilities.
(f) Write the throughput equation, using the state probabilities.

3.21. A company has a special purpose processing area that makes parts used throughout the company. A variety of different parts are made on a single machine and transported to various locations within the company for storage until they are needed in that area. The company has a very experienced employee who does the analysis of the parts currently available throughout the company and then decides what part type is to be made next at this machine. The part-needs analysis and release for processing is performed by this employee in two steps. The needs-analysis step takes 1/2 hour on average, but with the variety of parts to be analyzed, this time is exponentially distributed. Historical data indicates that 7 of every 9 parts analyses results in a standard part-type release and, since the part processing information is already on file, the part order is then released to the machine immediately.

Two of every nine analyses, however, results in the need for a special-purpose part for which the processing data are not available. Thus, this employee then develops a complete processing plan for the part. This processing plan development time averages an additional 2.5 hours. Due to the variety of the special purpose parts, it has been observed that this extra preparation time also is exponentially distributed. The order development employee is additionally charged with keeping the flow of jobs within the machine area reasonably smooth and timely. Towards this objective, the employee has developed the following release strategy. If there are 3 part orders already in the machining area, the employee holds the current completed order at her desk until a part has been completed and shipped. Then the "ready" order is given to the machine area personnel. If there is a completed (but blocked) order on the analyses employee's desk, no new order analysis is started until the blocked order has been cleared and been released to the machining area.

The machining area has only one machine and the average time for processing an order is 70 minutes. Due to the variety of part types, this processing time is exponentially distributed.

Develop a model of the special parts processing workstation (order analyses through processing). This encompasses the analyses employee and the machine (there is an operator for the machine and it is not necessary to keep track of this

operator). First draw a diagram of every possible configuration that this workstation can encounter. From this set of configurations, develop a state-space representation for these configurations. Then draw a rate-connected state diagram relating all of these configurations. Develop the steady-state equations for the rate-state diagram. Solve these equations for the steady-state probabilities. And finally, develop the workstation performance measures for this problem (machine utilization, order-development employee utilization, and throughput).

3.22. Consider an $E_2/M/1/3$ model with the arrival rate of 3 jobs per hour and a service rate of 4 jobs per hour. Compute the steady state probabilities and the system performance measures of utilization, CT_s, WIP_s, and throughput. Note that this system has a capacity of 3 jobs.

3.23. Consider an $E_2/M/1/4$ model with the arrival rate of 3 jobs per hour and a service rate of 4 jobs per hour. Compute the steady state probabilities and the system performance measures of utilization, CT_s, WIP_s, and throughput. Note that this system has a capacity of 4 jobs.

3.24. Solve Problem 3.21 using a spreadsheet such as Excel.

3.25. Find the parameters of a GE_2 approximation for a random variable X with specified mean and squared coefficient of variation:

Case	$E[X]$	$C^2[X]$	λ_1	α	λ_2
i	1	5/4			
ii	4/3	3/2			
iii	5	2			
iv	5/8	5/2			

3.26. Develop a model of an $M/GE_2/1/3$ system and compute the system performance measures given the mean arrival rate is 0.2/hr and the service distribution has parameters $E[S] = 5$ hr and $C^2[S] = 2$.

3.27. Develop a model of an $M/GE_2/1/3$ system and compute the system performance measures given the mean arrival rate is 3/hr and the service distribution has parameters $\mu = 3$/hr, $\alpha = 0.5$, and $\gamma = 4$/hr.

3.28. Solve Problems 3.25 and 3.26 using a spreadsheet such as Excel.

3.29. Develop the node-arc diagram for an $M/GE_2/2/3$ system (identical machines).

3.30. Using the approximation of Eq. 3.19, compute the cycle time in an $M/G/1$ system for three systems with the same arrival rates of $\lambda = 4$ and service times $E[T_s] = 0.2$, but different squared coefficients of variation ($C^2[T_s] = 1/2, 1, 2$).

3.31. Using the data from Problem 3.30, except for λ, develop a graph of the system WIP_s over the utilization from 0.1 to 0.95 in steps of 0.05. Insert three curves into the graph, based on the squared coefficients of variation ($C^2[T_s] = 0.5, 1, 2$).

3.32. Using the approximation of Property 3.6 and Eq. (3.21), compute the cycle time in the system for three systems with the same mean arrival rates of $\lambda = 4$ and mean service times of $E[T_s] = 0.4$, but different squared coefficients of variation ($C^2[T_s] = 1/2, 1, 2$). Note here that one machine is not adequate since $u > 1$, so assume that there are two-identical machines available, i.e., use an $M/G/2$ system.

3.33. Consider a single-server system with two types of jobs. The system has a limited capacity of three total jobs in the system at any time. The job classes have different mean arrival and service rates, but all are assumed to be exponentially distributed. Let λ_1 be the mean arrival rate and μ_1 be the mean service rate of Job Type 1, and let λ_2 be the mean arrival rate and μ_2 be the mean service rate of Job Type 2. Jobs are served on a first-come first-serve basis (denoted as FCFS or FIFO).
(a) Develop the labeled directed arc network for this system. Hint: there are fifteen different states and the sequence of job types in the queue must be maintained.
(b) Write the equations linking the steady-state probabilities.
(c) Write a formula for computing (in terms of the p_i's) the total WIP_s, WIP_s by product type, throughput, throughput by product type, the system CT_s, CT_s by product type.

3.34. Consider a single-server system with two types of jobs. The system has a limited capacity of three total jobs in the system at any time. The job classes have different mean arrival and service rates, but all are assumed to be exponentially distributed. Let λ_1 be the mean arrival rate and μ_1 be the mean service rate of Job Type 1, and let λ_2 be the mean arrival rate and μ_2 be the mean service rate of Job Type 2. Jobs are served on a non-preemptive priority basis with job type 1 given preference; that is, once a job starts it can not be displaced from the machine.
(a) Develop the labeled directed arc network for this system. Hint: there are thirteen different states and the sequence of job types in the queue will always be Type 1's in front of Type 2's.
(b) Write the equations linking the steady-state probabilities.
(c) Write a formula for computing (in terms of the p_i's) the total WIP_s, WIP_s by product type, throughput, throughput by product type, the system CT_s, and CT_s by product type.

3.35. Team Computer Project. Consider a situation (factory) where there is a limit of 4 jobs allowed at any time; arrivals to a full system are lost. Assume that all inter-arrival and processing times are exponentially distributed with mean rates specified. Job processing has two steps (Step 1 uses Machine 1 and Step 2 uses Machine 2). That is, there are two independent processing steps that must be done in the sequence: Machine 1 then Machine 2. The system is automated with-respect-to job movement between the queue and machines and between machines and then from the last machine to shipping (not part of this problem). There currently is no space for a job to wait for processing at Machine 2 after it has completed processing at Machine 1. Therefore, the completed job is left on Machine 1 until Machine 2 becomes available.

Management would like to improve the factory throughput and they are want to know what throughput improvement could be gained if they would invest in a

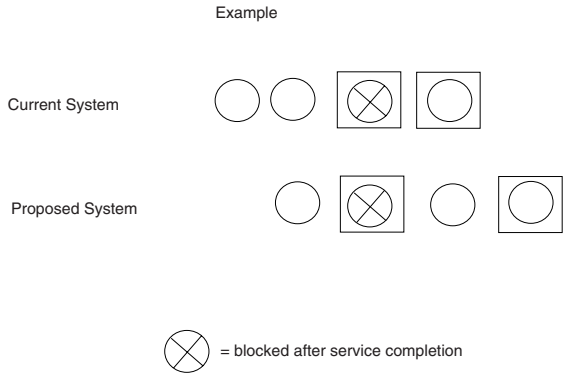


Fig. 3.6 Two configurations for Problem 3.35

conveyor between the machines. Develop a model and obtain the throughput for this system under the following two parameter sets: $\lambda = 6$, $\mu_1 = 8$, $\mu_2 = 7$ and $\lambda = 9$, $\mu_1 = 6$, $\mu_2 = 6$. Contrast the system throughput with and without a single buffer (job holding station) between the two machines for both configurations (see Fig. 3.6).

Develop a computer code to solve these two problems and evaluate the system throughput. Make it general in that the rate parameters are input or specified values within the spreadsheet that can be changed (such as merely changing parameter values between the data sets).

3.36. Model an $E_2/M/1/3$ system with a dependent arrival process in that once the system is full, the arrival process is shutoff until space is available in the system.

References

1. Allen, A.O. (1978). *Probability, Statistics, and Queueing Theory: With Computer Science Applications*, Academic Press, New York.
2. Altiok, T. (1996). *Performance Analysis of Manufacturing Systems*, Springer-Verlag, New York.
3. Hall, R.W. (1991). *Queueing Methods: For Services and Manufacturing*, Prentice-Hall, Englewood Cliffs, N. J.
4. Gross, D., and Harris, C.M. (1998). *Fundamentals of Queueing Theory*, Third Edition, John Wiley & Sons, New York.
5. Hopp, W.J. and Spearman M.L. (1996). *Factory Physics: Foundations of Manufacturing Management*, Irwin, Chicago.
6. Kendall, D.G. (1953). Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains. *Annals of Mathematical Statistics*, **24**:338–354.
7. Kingman, J.F.C. (1962). On queues in heavy traffic. *J. Royal Statist. Soc. Ser. B*, **32**:102–110.
8. Kraemer, W. and Langenbach-Belz, M. (1976). Approximate Formulae for the Delay in the Queueing System $GI/G/1$. *Congressbook, Eighth Int. Teletraffic Cong.*, Melbourne.

9. Sakasegawa, H. (1977). An Approximation Formula $L_q = \alpha\beta^{\rho}(1-\rho)$. *Annuals of the Institute of Statistical Mathematics*, **29**:67–75.
10. Whitt, W. (1983). The Queueing Network Analyzer. *The Bell System Technical Journal*, **62**:2779–2814.

http://www.springer.com/978-3-642-16617-4

Manufacturing Systems Modeling and Analysis
Curry, G.L.; Feldman, R.M.
2011, XVI, 338 p., Hardcover
ISBN: 978-3-642-16617-4