Queuing Theory

Queuing Theory

The ***input process*** is usually called the **arrival process.** Arrivals are called **customers.**

Assumptions:

1. No more than one arrival can occur at a given instant. For a case like a restaurant, if more than one arrival can occur at a given instant, we say that **bulk arrivals** are allowed.

2. The arrival process is unaffected by the number of customers present in the system.

There are two common situations in which the arrival process may depend on the number of customers present. The first occurs when arrivals are drawn from a small population. Models in which arrivals are drawn from a small population are called **finite source models.** Another situation in which the arrival process depends on the number of customers present occurs when the rate at which customers arrive at the facility decreases when the facility becomes too crowded.

- **If the arrival process is unaffected by the number of customers present, we usually describe it by specifying a probability distribution that governs the time between successive arrivals.**

Examples of Queuing Systems

Situation	Input Process	Output Process
Bank	Customers arrive at bank	Tellers serve the customers
Pizza parlor	Requests for pizza delivery are received	Pizza parlor sends out truck to deliver pizzas
Hospital blood bank	Pints of blood arrive	Patients use up pints of blood
Naval shipyard	Ships at sea break down and are sent to shipyard for repairs	Ships are repaired and return to sea

The Output or Service Process

To describe the output process (often called the service process) of a queuing system, we usually specify a probability distribution—the **service time distribution**—which governs a customer's service time.

Assumptions:

The service time distribution is independent of the number of customers present. This implies, for example, that the server does not work faster when more customers are present.

- **servers in parallel** and **servers in series.** Servers are in parallel if all servers provide the same type of service and a customer need only pass through one server to complete service.

  Servers are in parallel if all servers provide the same type of service and a customer need only pass through one server to complete service. For example, the tellers in a bank are usually arranged in parallel; any customer need only be serviced by one teller, and any teller can perform the desired service. Servers are in series if a customer must pass through several servers before completing service. An assembly line is an example of a series queuing system.

- **Queue discipline**

It describes the method used to determine the order in which customers are served.

- ❑ **FCFS discipline** (first come, first served), in which customers are served in the order of their arrival.
- ❑ **LCFS discipline** (last come, first served), the most recent arrivals are the first to enter service.
- ❑ **SIRO discipline** (service in random order)

- **Method Used by Arrivals to Join Queue**

Another factor that has an important effect on the behavior of a queuing system is the method that customers use to determine which line to join

Modeling the Arrival Process

we assume that at most one arrival can occur at a given instant of time.

We define t_i to be the time at which the ith customer arrives.

For $i \geq 1$, we define $T_i = t_{i+1} - t_i$ to be the ith interarrival time.

In modeling the arrival process, we assume that the T_i's are independent, continuous random variables described by the random variable **A**.

We assume that **A** has a density function $a(t)$.
for small Δt, $P(t \leq \mathbf{A} \leq t + \Delta t)$ is approximately $\Delta t a(t)$. Of course, a negative interarrival time is impossible.

This allows us to write

$$P(\mathbf{A} \leq c) = \int_0^c a(t)dt \qquad \text{and} \qquad P(\mathbf{A} > c) = \int_c^\infty a(t)dt$$

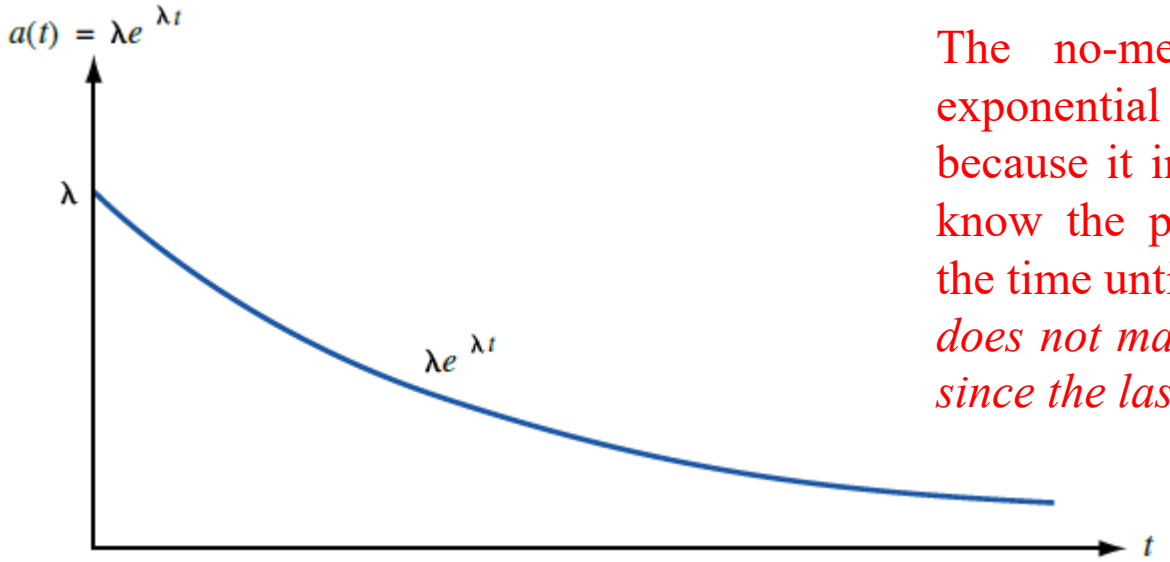
The exponential Distribution

We define λ to be the **arrival rate,** which will have units of arrivals per hour.

We define $\frac{1}{\lambda}$ to be the mean or average interarrival time. Without loss of generality, we assume that time is measured in units of hours. Then $\frac{1}{\lambda}$ will have units of hours per arrival.

$$\frac{1}{\lambda} = \int_0^\infty ta(t)dt$$

In most applications of queuing, an important question is how to choose **A** to reflect reality and still be computationally tractable. The most common choice for **A** is the **exponential distribution.** An exponential distribution with parameter λ has a density $a(t) = \lambda e^{-\lambda t}$.



Density Function for Exponential Distribution

The no-memory property of the exponential distribution is important, because it implies that if we want to know the probability distribution of the time until the next arrival, then *it does not matter how long it has been since the last arrival.*

The Poisson Distribution

If interarrival times are exponential, then the probability distribution of the number of arrivals occurring in any time interval of length t *can be defined by Poisson's distribution.*

A discrete random variable $\mathbf{N}$ has a Poisson distribution with parameter λ if, for $n = 0, 1, 2, \ldots,$

$$P(\mathbf{N} = n) = \frac{e^{-\lambda}\lambda^n}{n!} \qquad (n = 0, 1, 2, \ldots)$$

If $\mathbf{N}$ is a Poisson random variable, it can be shown that $E(\mathbf{N}) = \text{var } \mathbf{N} = \lambda$. If we define $\mathbf{N}_t$ to be the number of arrivals to occur during any time interval of length t,

$$P(\mathbf{N}_t = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \qquad (n = 0, 1, 2, \ldots)$$

Since $\mathbf{N}_t$ is Poisson with parameter λt, $E(\mathbf{N}_t) = \text{var } \mathbf{N}_t = \lambda t$. An average of λt arrivals occur during a time interval of length t, so λ may be thought of as the average number of arrivals per unit time, or the arrival rate.

Consider the following two assumptions:

1 Arrivals defined on nonoverlapping time intervals are independent (for example, the number of arrivals occurring between times 1 and 10 does not give us any information about the number of arrivals occurring between times 30 and 50).

2 For small Δt (and any value of t), the probability of one arrival occurring between times t and $t + \Delta t$ is $\lambda \Delta t + o(\Delta t)$, where $o(\Delta t)$ refers to any quantity satisfying

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

Also, the probability of no arrival during the interval between t and $t + \Delta t$ is $1 - \lambda \Delta t + o(\Delta t)$, and the probability of more than one arrival occurring between t and $t + \Delta t$ is $o(\Delta t)$.

If assumptions 1 and 2 hold, then $\mathbf{N}_t$ follows a Poisson distribution with parameter λt, and interarrival times are exponential with parameter λ; that is, $a(t) = \lambda e^{-\lambda t}$.

Problem:

The number of pizzas ordered per hour at Dominos's outlet follows a Poisson distribution, with an average of 30 pizzas per hour being ordered.

1. Find the probability that exactly 60 pizzas are ordered between 10 P.M. and 12 midnight.
2. Find the mean and standard deviation of the number of pizzas ordered between 9 P.M. and 1 A.M.
3. Find the probability that the time between two consecutive orders is between 1 and 3 minutes.

The Erlang Distribution

If interarrival times do not appear to be exponential, they are often modeled by an Erlang distribution. An Erlang distribution is a continuous random variable (call it **T**) whose density function $f(t)$ is specified by two parameters: a rate parameter R and a shape parameter k (k must be a positive integer). Given values of R and k, the Erlang density has the following probability density function:

$$f(t) = \frac{R(Rt)^{k-1}e^{-Rt}}{(k-1)!} \qquad (t \geq 0)$$

Using integration by parts, we can show that if **T** is an Erlang distribution with rate parameter R and shape parameter k, then

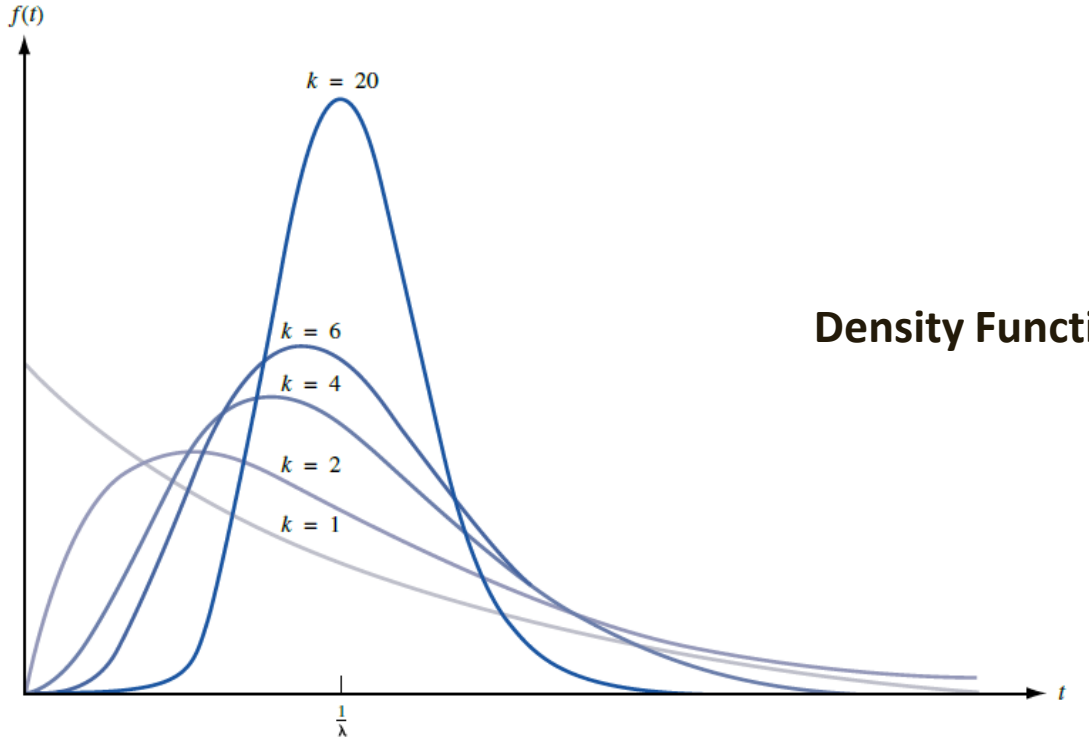
$$E(\mathbf{T}) = \frac{k}{R} \qquad \text{and} \qquad \text{var } \mathbf{T} = \frac{k}{R^2}$$

To see how varying the shape parameter changes the shape of the Erlang distribution, we consider for a given value of λ, a family of Erlang distributions with rate parameter $k\lambda$ and shape parameter k.

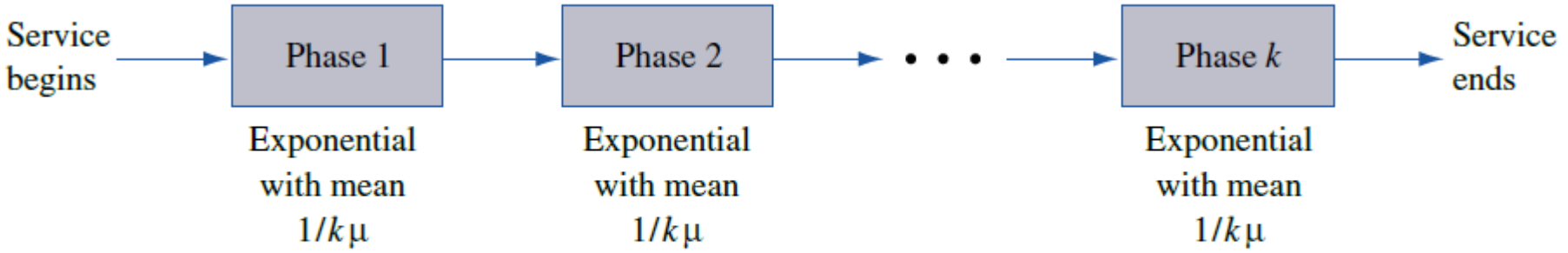
- for $k = 1$, the Erlang distribution is an exponential distribution with parameter R. As k increases, the Erlang distribution behaves more and more like a normal distribution. For extremely large values of k, the Erlang distribution approaches a random variable with zero variance (that is, a constant interarrival time). Thus, by varying k, we may approximate both skewed and symmetric distributions.

- It can be shown that an Erlang distribution with shape parameter k and rate parameter $k\lambda$ has the same distribution as the random variable $\mathbf{A}1+\mathbf{A}2+\ldots+\mathbf{A}k$, where each $\mathbf{A}i$ is an exponential random variable with parameter $k\lambda$, and the $\mathbf{A}i$'s are independent random variables.

- If we model interarrival times as an Erlang distribution with shape parameter k, we are really saying that the interarrival process is equivalent to a customer going through k phases (each of which has the no-memory property) before arriving. For this reason, the shape parameter is often referred to as the *number of phases* of the Erlang distribution.



Density Functions for Erlang Distributions



Representation of Erlang Service Time

Modeling the Service Process

We assume that the service times of different customers are independent random variables and that each customer's service time is governed by a random variable **S** having a density function $s(t)$. We let $1/\mu$ be the mean service time for a customer. Of course,

$$\frac{1}{\mu} = \int_0^\infty ts(t)dt$$

The variable $\frac{1}{\mu}$ will have units of hours per customer, so μ has units of customers per hour. For this reason, we call μ the service rate.

The Kendall–Lee Notation for Queuing Systems

Kendall (1951) devised the following notation. Each queuing system is described by six characteristics:

1/2/3/4/5/6

The first characteristic specifies the nature of the arrival process. The following standard abbreviations are used:

M = Interarrival times are independent, identically distributed (iid) random variables having an exponential distribution.

D = Interarrival times are iid and deterministic.

E_k = Interarrival times are iid Erlangs with shape parameter k.

GI = Interarrival times are iid and governed by some general distribution.

The second characteristic specifies the nature of the service times:

M = Service times are iid and exponentially distributed.

D = Service times are iid and deterministic.

E_k = Service times are iid Erlangs with shape parameter k.

G = Service times are iid and follow some general distribution.

The third characteristic is the number of parallel servers.

The fourth characteristic describes the queue discipline:

FCFS = First come, first served

LCFS = Last come, first served

SIRO = Service in random order

GD = General queue discipline

The fifth characteristic specifies the maximum allowable number of customers in the system (including customers who are waiting and customers who are in service).

The sixth characteristic gives the size of the population from which customers are drawn. Unless the number of potential customers is of the same order of magnitude as the number of servers, the population size is considered to be infinite. In many important models 4/5/6 is $GD/\infty/\infty$. If this is the case, then 4/5/6 is often omitted.

As an illustration of this notation, $M/E_2/8/\text{FCFS}/10/\infty$ might represent a health clinic with 8 doctors, exponential interarrival times, two-phase Erlang service times, an FCFS queue discipline, and a total capacity of 10 patients.

Birth Death Process

We define the number of people present in any queuing system at time t to be the **state** of the queuing system at time t. For $t = 0$, the state of the system will equal the number of people initially present in the system. Of great interest to us is the quantity $P_{ij}(t)$ which is defined as the probability that j people will be present in the queuing system at time t, given that at time 0, i people are present. Note that $P_{ij}(t)$ is analogous to the n-step transition probability $P_{ij}(n)$ (the probability that after n transitions, a Markov chain will be in state j, given that the chain began in state i)

for many queuing systems, $P_{ij}(t)$ will, for large t, approach a limit π_j, which is independent of the initial state i. We call π_j the **steady state,** or equilibrium probability, of state j.

Similarly, it turns out that for many queuing systems, $P_{ii}(t)$ will, for large t, approach a limit π_j, which is independent of the initial state i. We call π_i the **steady state,** or equilibrium probability, of state j.

A **birth–death process** is a continuous-time stochastic process for which the system's state at any time is a nonnegative integer. If a birth–death process is in state j at time t, then the motion of the process is governed by the following laws.

Laws of Motion for Birth–Death Processes

Law 1 With probability $\lambda_j \Delta t + o(\Delta t)$, a birth occurs between time t and time $t + \Delta t$. A birth increases the system state by 1, to $j + 1$. The variable λ_j is called the **birth rate** in state j. In most queuing systems, a birth is simply an arrival.

Law 2 With probability $\mu_j \Delta t + o(\Delta t)$, a death occurs between time t and time $t + \Delta t$. A death decreases the system state by 1, to $j - 1$. The variable μ_j is the **death rate** in state j. In most queuing systems, a death is a service completion. Note that $\mu_0 = 0$ must hold, or a negative state could occur.

Law 3 Births and deaths are independent of each other.

Laws 1–3 can be used to show that the probability that more than one event (birth or death) occurs between t and $t + \Delta t$ is $o(\Delta t)$. Note that any birth–death process is completely specified by knowledge of the birth rates λ_j and the death rates μ_j. Since a negative state cannot occur, any birth–death process must have $\mu_0 = 0$.

Relation of Exponential Distribution to Birth–Death Processes

Most queuing systems with exponential interarrival times and exponential service times may be modeled as birth–death processes. To illustrate why this is so, consider an $M/M/1/\text{FCFS}/\infty/\infty$ queuing system in which interarrival times are exponential with parameter λ and service times are exponentially distributed with parameter μ. If the state (number of people present) at time t is j, then the no-memory property of the exponential distribution implies that the probability of a birth during the time interval $[t, t + \Delta t]$ will not depend on how long the system has been in state j. This means that the probability of a birth occurring during $[t, t + \Delta t]$ will not depend on how long the system has been in state j and thus may be determined as if an arrival had just occurred at time t. Then the probability of a birth occurring during $[t, t + \Delta t]$ is

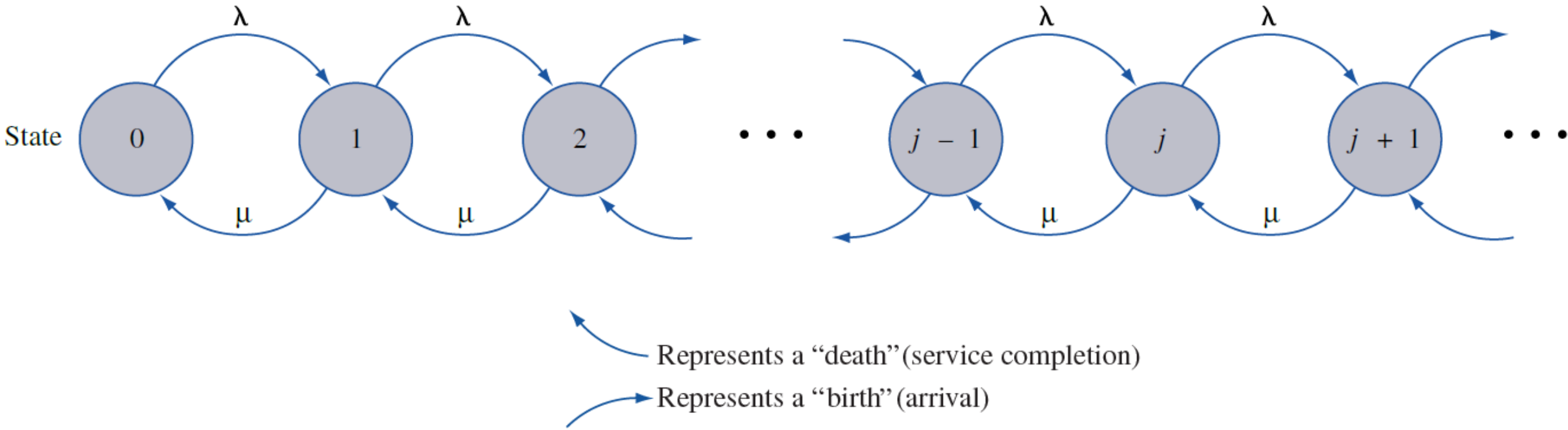
$$\int_0^{\Delta t} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda \Delta t}$$

By the Taylor series expansion given in Section 11.1,

$$e^{-\lambda \Delta t} = 1 - \lambda \Delta t + o(\Delta t)$$

This means that the probability of a birth occurring during $[t, t + \Delta t]$ is $\lambda \Delta t + o(\Delta t)$. From this we may conclude that the birth rate in state j is simply the arrival rate λ.

To determine the death rate at time t, note that if the state is zero at time t, then nobody is in service, so no service completion can occur between t and $t + \Delta t$. Thus, $\mu_0 = 0$.



Rate Diagram for *M*/*M*/1/FCFS/∞/∞ Queuing System

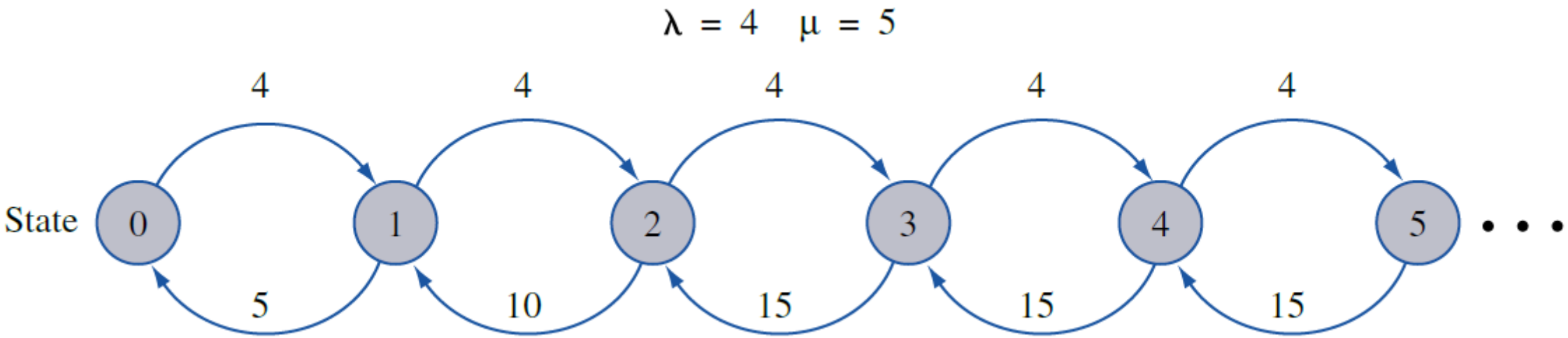
More complicated queuing systems with exponential interarrival times and exponential service times may often be modeled as birth–death processes by adding the service rates for occupied servers and adding the arrival rates for different arrival streams. For example, consider an $M/M/3/\text{FCFS}/\infty/\infty$ queuing system in which interarrival times are exponential with $\lambda = 4$ and service times are exponential with $\mu = 5$. To model this system as a birth–death process, we would use the following parameters (see Figure 10):

$$\lambda_j = 4 \qquad (j = 0, 1, 2, \ldots)$$

$$\mu_0 = 0, \quad \mu_1 = 5, \quad \mu_2 = 5 + 5 = 10, \quad \mu_j = 5 + 5 + 5 = 15 \qquad (j = 3, 4, 5, \ldots)$$

If either interarrival times or service times are nonexponential, then the birth–death process model is not appropriate.[†] Suppose, for example, that service times are not exponential and we are considering an $M/G/1/\text{FCFS}/\infty/\infty$ queuing system. Since the service times for an $M/G/1/\text{FCFS}/\infty/\infty$ system may be nonexponential, the probability that a death (service completion) occurs between t and $t + \Delta t$ will depend on the time since the last service completion. This violates law 2, so we cannot model an $M/G/1/\text{FCFS}/\infty/\infty$ system as a birth–death process.

[†]A modified birth–death model can be developed if service times and interarrival times are Erlang distributions.



Rate Diagram for *M*/*M*/3/FCFS/∞/∞ Queuing System

The $M/M/1/GD/\infty/\infty$ queuing system has exponential interarrival times (we assume that the arrival rate per unit time is l) and a single server with exponential service times (we assume that each customer's service time is exponential with rate m).

Often we are interested in the amount of time that a typical customer spends in a queuing system. We define W as the expected time a customer spends in the queuing system, including time in line plus time in service, and W_q as the expected time a customer spends waiting in line. Both W and W_q are computed under the assumption that the steady state has been reached. By using a powerful result known as **Little's queuing formula,** W and Wq may be easily computed from L and Lq. We first define (for any queuing system or any subset of a queuing system) the following quantities:

λ = average number of arrivals *entering* the system per unit time
L = average number of customers present in the queuing system
L_q = average number of customers waiting in line
L_s = average number of customers in service
W = average time a customer spends in the system
W_q = average time a customer spends in line
W_s = average time a customer spends in service

(In these definitions, all averages are steady-state averages).

For *any* queuing system in which a steady-state distribution exists, the following relations hold:

$$L=\lambda W$$
$$L_q = \lambda W_q$$
$$L_s = \lambda\ W_s$$

We define $\rho = \frac{\lambda}{\mu}$. For reasons that will become apparent later, we call ρ the **traffic intensity** of the queuing system.

if $\rho \geq 1$, no steady-state distribution exists. Since $\rho = \frac{\lambda}{\mu}$, we see that if $\lambda \geq \mu$ (that is, the arrival rate is at least as large as the service rate), then no steady-state distribution exists.

If $\rho > 1$, it is easy to see why no steady-state distribution can exist. Suppose $\lambda = 6$ customers per hour and $\mu = 4$ customers per hour. Even if the server were working all the time, she could only serve an average of 4 people per hour. Thus, the average number of customers in the system would grow by at least $6 - 4 = 2$ customers per hour. This means that after a long time, the number of customers present would "blow up," and no steady-state distribution could exist. If $\rho = 1$, the nonexistence of a steady state is not quite so obvious, but our analysis does indicate that no steady state exists.

$$L = \frac{\rho}{1 - \rho} \qquad L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

$$W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1 - \rho)} = \frac{1}{\mu - \lambda} \qquad W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)}$$

Notice that (as expected) as ρ approaches 1, both W and W_q become very large. For ρ near zero, W_q approaches zero, but for small ρ, W approaches $\frac{1}{\mu}$, the mean service time.

Suppose that all car owners fill up when their tanks are exactly half full. At the present time, an average of 7.5 customers per hour arrive at a single-pump gas station. It takes an average of 4 minutes to service a car. Assume that interarrival times and service times are both exponential.

1 For the present situation, compute L and W.

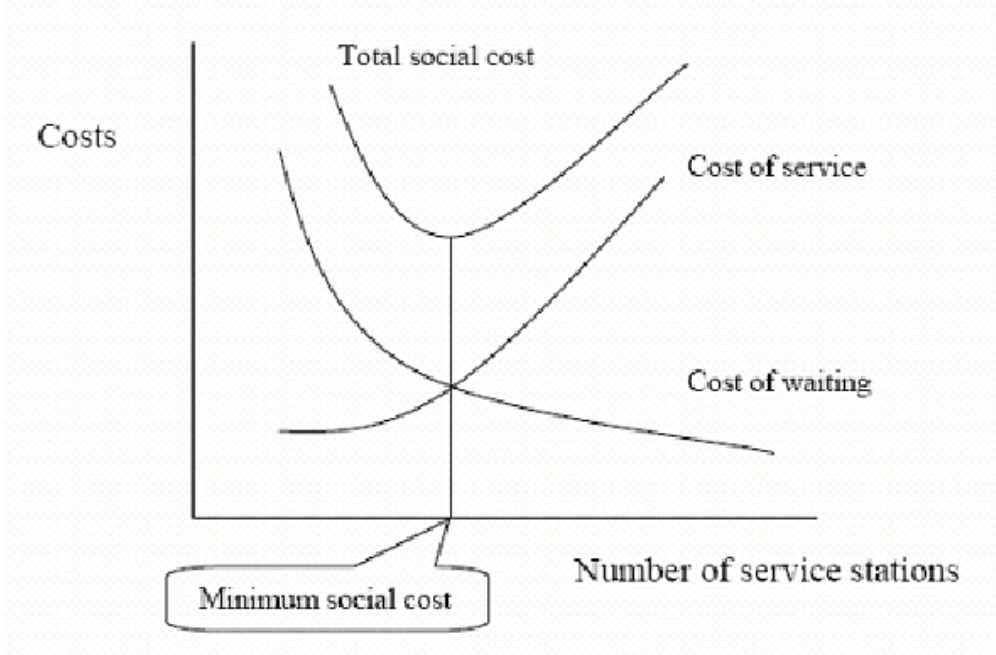
2 Suppose that a gas shortage occurs and panic buying takes place. To model this phenomenon, suppose that all car owners now purchase gas when their tanks are exactly three-quarters full. Since each car owner is now putting less gas into the tank during each visit to the station, we assume that the average service time has been reduced to $3\frac{1}{3}$ minutes. How has panic buying affected L and W?

Solution **1** We have an $M/M/1/GD/\infty/\infty$ system with $\lambda = 7.5$ cars per hour and $\mu = 15$ cars per hour. Thus, $\rho = \frac{7.5}{15} = .50$. From (26), $L = \frac{.50}{1-.50} = 1$, and from (28), $W = \frac{L}{\lambda} = \frac{1}{7.5} = 0.13$ hour. Hence, in this situation, everything is under control, and long lines appear to be unlikely.

2 We now have an $M/M/1/GD/\infty/\infty$ system with $\lambda = 2(7.5) = 15$ cars per hour. (This follows because each car owner will fill up twice as often.) Now $\mu = \frac{60}{3.333} = 18$ cars per hour, and $\rho = \frac{15}{18} = \frac{5}{6}$. Then

$$L = \frac{\frac{5}{6}}{\quad} = 5 \text{ cars} \qquad \text{and} \qquad W = \frac{L}{\lambda} = \frac{5}{15} = \frac{1}{3} \text{ hours} = 20 \text{ minutes}$$

Queuing costs



Service mechanics come to take spares at a shop at 6/hour on the average. Waiting for them costs Rs. 8/- per hour. A Shop attendant's wage is Rs. 5/- per hour. There is only one counter. The service rates are: 1 attendant: 8/hour, 2 attendants: 12/hour, 3 attendants: 16/hour. With usual M/M/1 assumptions, find how many attendants to choose.

No. of Atten-dants (N)	Arrival Rate per hour (λ)	Service Rate per hour (μ)	Utilization Factor ρ = λ/ μ	Waiting time (W) in hour W = ρ/((1 – ρ)*λ)	Service Cost (Rs.) SC = N*8hr*5	Waiting Cost (Rs.) WC = W*8hr*λ*8	Total Cost TC = SC+WC
1	6	8	3/4	1/2	1*8*5 = 40	(1/2)*8*6*8 = 192	232
2	6	12	1/2	1/6	2*8*5 = 80	(1/6)*8*6*8 = 64	144 Lowest
3	6	16	3/8	1/10	3*8*5 = 120	(1/10)*8*6*8 = 38.4	158.4

Machines fail at 4 per hour and the cost of non-productive machine is Rs. 9 per hour. A fast repairman charges Rs. 6 per hour and repairs at 7 per hour. A slow repairman charges Rs. 3 per hour but repairs at 5 per hour. With usual M/M/1 assumptions, find which repairman to hire.

Repair-man	Arrival Rate per hour (λ)	Service Rate per hour (μ)	Utilization Factor $\rho = \lambda/\mu$	Average No. of machines in repair (L) $L = \rho/(1-\rho)$	M/c Idle Cost (Rs.) IC = L*8hr *hourly cost	Repairman Cost (Rs.) RC = 8 hr* hourly cost	Total Cost TC = SC+WC
Fast	4	7	4/7	(4/7)/(3/7) = 4/3	(4/3)*8*9 = 96	8*6 = 48	144 Lower
Slow	4	5	4/5	(4/5)/(1/5) = 4	4*8*9 = 288	8*3 = 24	312

Take home question

Machinists who work at a tool-and-die plant must check out tools from a tool center. An average of ten machinists per hour arrive seeking tools. At present, the tool center is staffed by a clerk who is paid $6 per hour and who takes an average of 5 minutes to handle each request for tools. Since each machinist produces $10 worth of goods per hour, each hour that a machinist spends at the tool center costs the company $10. The company is deciding whether or not it is worthwhile to hire (at $4 per hour) a helper for the clerk. If the helper is hired, the clerk will take an average of only 4 minutes to process requests for tools. Assume that service and interarrival times are exponential. Should the helper be hired?