Queueing Theory

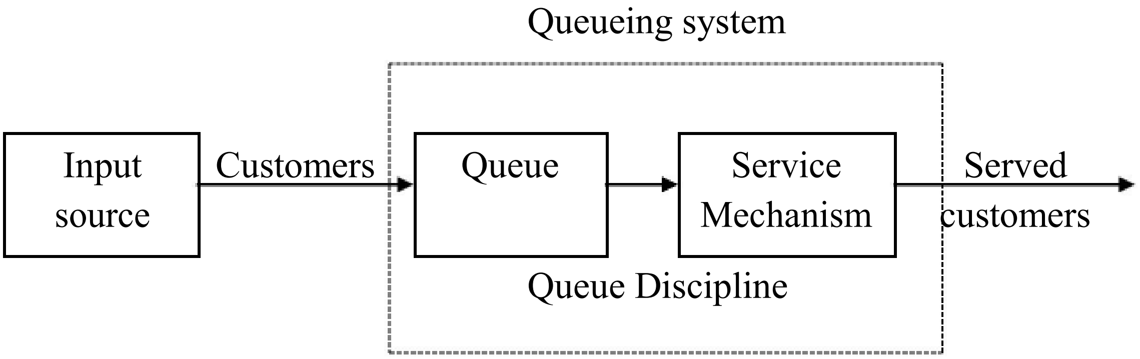
Introduction

- ✓ Queues (waiting line) are a part of everyday life.
- ✓ Providing too much service involves excessive costs. And not providing enough service capacity causes the waiting line to become excessively long.
- ✓ The ultimate goal is to achieve an economic balance between the cost of service and the cost associated with the waiting for that service.
- ✓ Queueing theory is the study of waiting in all these various guises.

Prototype Example—Doctor Requirement in a Emergence Room

- ✓ Consider assigning an extra doctor to the emergency room, which has one doctor already.
- ✓ How much can we reduce the average waiting time for patients if the extra doctor is hired?

Basic Structure of Queueing Models

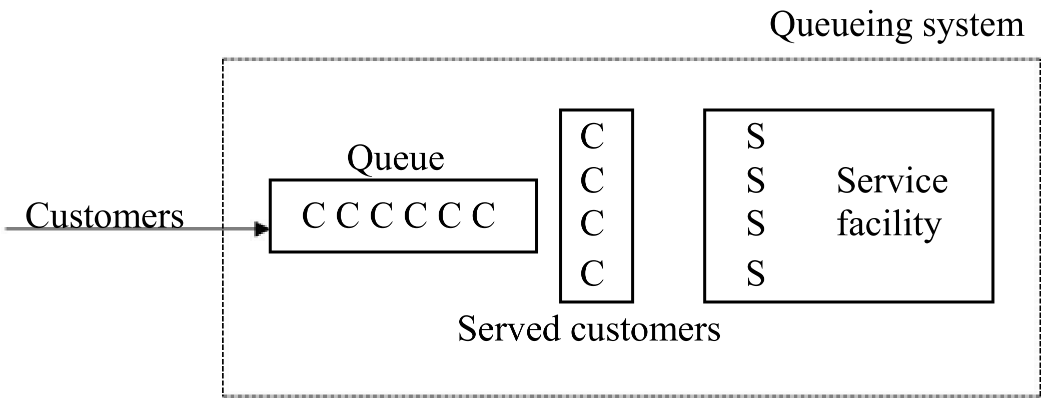


- ✓ Input Source (Calling Population)
  - ➢ One characteristic of the input source is its size. The size is the total number of customers. The size may be infinite (default one) or finite.
  - ➢ When will each one arrive? Associate with a distribution—usually, Poisson distribution (the number of customers generated until any specific time) or Exponential distribution (**interarrival time**).
  - ➢ A customer may be balking, who refuses to enter the system and is lost if the queue is too long.

- ✓ **Queue**
  - ➢ The queue is where customers wait before being served.
  - ➢ A queue is characterized by the maximum permissible number of customers that it can contain. Queue may be infinite (default one) or finite.
- ✓ **Queue Discipline**
  - ➢ Refers to the order in which members of the queue are selected for service.
  - ➢ First-come-first-serve is normally used.
- ✓ **Service Mechanism**
  - ➢ Consists of one or more service facilities, each of which contains one or more parallel service channels, called servers.
  - ➢ At a given facility, the customer enters one of the parallel service channels and is served by that server.
  - ➢ Most elementary models assume one service facility with either one or a finite number of servers.
  - ➢ Service time is usually defined by a probability distribution.

An Elementary Queueing Process

- ✓ A single waiting line forms in the front of a single service facility, within which are stationed one or more servers. Each customer is serviced by one of the servers, perhaps after some waiting in the queue.



- ✓ **The prototype example is of this type.**
- ✓ **We usually label a queueing model as ----/----/----**
  - ➢ The first spot is for distribution of interarrival times. The second spot is for distribution of service times. The third one is for number of servers.
  - ➢ *M* = exponential distribution (Markovian), which is the most widely used.
  - ➢ *D* = degenerate distribution (constant time).
  - ➢ E_k = Erlang distribution.
  - ➢ *G* = general distribution (any arbitrary distribution allowed)

Terminology and Notation

- ✓ State of system = number of customers in queueing system.
- ✓ Queue length = number of customers waiting for service to begin = state of system minus number of customers being served.
- ✓ $N(t)$ = number of customers in queueing system at time t.
- ✓ $P_n(t)$ = probability of exactly n customers in queueing system at time t.
- ✓ s = number of servers (parallel service channels) in queueing system.
- ✓ λ_n = mean arrival rate (expected number of arrival per unit time) of new customers when n customers are in system.
  - ➢ When λ_n is a constant for all n, this constant is denoted by λ .
  - ➢ $1/\lambda$ is the expected interarrival time.
- ✓ μ_n = mean service rate for overall system (expected number of customer completing service per unit time) when n customers are in system.
  - ➢ μ_n represents combined rate at which all busy servers achieve service completions.
  - ➢ When the mean service rate *per busy server* is a constant for all $n \geq 1$, this constant is denoted by μ.
  - ➢ $\mu_n = s\mu$ when $n \geq s$ (all servers are busy).
  - ➢ $1/\mu$ is the expected service time.
- ✓ $\rho = \lambda/(s\mu)$ is the **utilization factor** for the service facility, i.e., the expected fraction of time the individual servers are busy.
- ✓ **Transient condition**—when a queueing system has recently begun, the state of the system will be greatly affected by the initial state and by the time that has since elapsed.
- ✓ **Steady-state condition**—after sufficient time has elapsed, the state of the system becomes essentially independent of the initial state and the elapsed time.
  - ➢ Queueing theory has tended to focus largely on the steady-state condition.
- ✓ More notations are defined in a steady-state condition.
- ✓ P_n = probability of exact n customers in queueing system.
- ✓ L = expected number of customers in queueing system = $\sum^{\infty} nP_n$.

$n=0$

- ✓ L_q = expected queue length (excludes customers being served) = $\sum_{n=s}^{\infty}(n-s)P_n$.
- ✓ $\mathcal{W}$ = waiting time in system (includes service time) for each customer.
- ✓ $W = E(\mathcal{W})$.
- ✓ $\mathcal{W}_q$ = waiting time in queue (exclude service time) for each customer.
- ✓ $W_q = E(\mathcal{W}_q)$.

Relationships between *L, W, L_q*, and *W_q*

- ✓ Assume that λ_n is a constant for all *n*.
- ✓ In a steady-state queueing process, $L = \lambda W$ (Little's formula) and $L_q = \lambda W_q$.
- ✓ If the λ_n are not equal, then λ can be replaced in these equation by λ, the average arrival rate over the long time.
- ✓ Assume that the mean service time ($1/\mu$) is a constant. Thus, $W = W_q + \frac{1}{\mu}$.
- ✓ These four fundamental quantities (*L, W, L_q*, and *W_q)* could be immediately determined as soon as one is found analytically.

Examples—commercial service system, transportation service system, internal service system, and social service system.

The Role of the Exponential Distribution

- ✓ The mostly commonly used distribution for interarrival and service time is exponential distribution.
- ✓ A random variable (interarrival or service times), *T*, is said to have an **exponential distribution** with parameter *α* if its probability density function is

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & for \quad t \geq 0 \\ 0 & for \quad t < 0 \end{cases}$$

- ✓ The cumulative probabilities are $P\{T \leq t\} = 1 - e^{-\alpha}$, $P\{T > t\} = e^{-\alpha t}$ for $t \geq 0$.
- ✓ $E(T) = \frac{1}{\alpha}$, $\text{var}(T) = \frac{1}{\alpha^2}$.
- ✓ Property 1: $f_T(t)$ is a strictly decreasing function of *t*.
  - ➢ $P\{0 \leq T \leq t\} > P\{t \leq T \leq t + t\}$ for any strictly positive of *t* and
  - ➢ The value *T* takes on is more likely to be "small" [less than half of

$E(T)$)] than "near" its expected value. Is this property real?

$$P\left\{0 \le T \le \frac{1}{2}\frac{1}{\alpha}\right\} = 0.393\ ,\ P\left\{\frac{1}{2}\frac{1}{\alpha} \le T \le \frac{3}{2}\frac{1}{\alpha}\right\} = 0.383\ .$$

- ➢ Not real when the service required is essentially identical for each customer, with the server always performing the same sequence of service operations.
- ➢ It is suitable for the situations where the specific tasks required of the server differ among customers (hospital or banking cases).

✓ Property 2: Lack of memory: $P\{T > t + \Delta t \mid T > \Delta t\} = P\{T > t\}$ for any positive of t and Δt.

- ➢ The probability distribution of the remaining time until the event occurs always is the same, regardless of how much time already has passed.
- ➢ The process "forgets" its history.
- ➢ The phenomenon occurs with the exponential distribution.
- ➢ For interarrival time, the time until next arrival is completely uninfluenced by when the last arrival occurred.

✓ Property 3: The minimum of several independent exponential random variables has an exponential distribution.

- ➢ Let $T_1, T_2, \ldots, T_n$ be independent exponential random variables with parameters α_1 , α_2 , …, α_n . Also, let U be the random variable that takes on the value equal to the minimum of the values of $T_1, T_2, \ldots, T_n$.
- ➢ If T_i represents the time until a particular event occurs, then U represents the time until the first of the n different events occurs.
- ➢ U indeed has an exponential distribution with parameter $\alpha = \sum_{i=1}^{n} \alpha_i$.
- ➢ If there are n different types of customers (interarrival time is exponential with parameter α_i), the interarrival time for the queueing system as a whole, has an exponential distribution with parameter $\alpha = \sum_{i=1}^{n} \alpha_i$.
- ➢ Suppose all n servers have the same exponential service-time distribution with parameter μ . The time until the next service completion from any server has an exponential distribution with parameter $\alpha = n\mu$.

✓ **Property 4: Relationship to the Poisson distribution.**

- ➢ Suppose that the time between consecutive occurrences of some particular kind of event has an exponential distribution with parameterα .

- Then, the number of occurrence by time t ($X(t)$) has a Poisson distribution with parameter αt .

$$P\{X(t) = n\} = \frac{(\alpha t)^n e^{-\alpha t}}{n!}$$

- With $n = 0$, $P\{X(t) = 0\} = e^{-\alpha t}$, which is just the probability from the exponential distribution that the first event occurs after time t.
- The mean of this Poisson distribution is $E\{X(t)\} = \alpha t$, so that the expected number of events per unit time is α . α is said to be the mean rate at which the events occurs.
- When the events are counted on a continuing basis, the counting process $\{X(t); t \geq 0\}$is said to be a **Poisson process** with parameter α .
- Define $X(t)$ as the number of service completions achieved by a continuously busy server in elapsed time t, where $\alpha = \mu$. For multiple-server queueing models, $X(t)$ can also be defined as the number of service completions achieved by n servers in elapsed time t, where $\alpha = n\mu$.
- Suppose the interarrival times have an exponential distribution with parameter λ . In this case, $X(t)$ is the number of arrivals in elapsed time t, where $\alpha = \lambda$ is the mean arrival rate. Therefore, arrivals occur according to a **Poisson input process** with parameter λ .

✓ Property 5: For all positive values of t, $P\{T \leq t + \Delta t \mid T > t\} \approx \alpha\Delta t$, for small Δt.

- The series expansion of e^x for any exponent x is $e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}$.
- $P\{T \leq t + \Delta t \mid T > t\} = P\{T \leq \Delta t\} = 1 - e^{-\alpha\Delta t} = 1 - 1 + \alpha\Delta t - \sum_{n=2}^{\infty} \frac{(-\alpha\ t)^n}{n!} \approx \alpha\Delta t.$

✓ Property 6: Unaffected by aggregation or disaggregation.

- If there are n different types of customers (each is a Poisson input process with parameter λ_i, the aggregated input is a Poisson with $\lambda = \lambda_1 + \lambda_2 \ldots + \lambda n$.
- Assuming that each arriving customer has a fixed probability p_i of being of type i, with $\lambda_i = p_i \lambda$ and $\sum_{i=1}^{n} p_i = 1$, the property says that the input process for customers of type i also must be Poisson with parameter λ_i.

The Birth-and-Death Process

✓ **birth** = arrival; **death** = departure; state, $N(t)$, is the number of customers in the queueing system at time t.

✓ The birth-and-death process describes probabilistically how $N(t)$ changes as t increases.

✓ Assumption 1: Given $N(t) = n$, the current probability distribution of the remaining time until the next birth (arrival) is exponential with parameter λ_n.

✓ Assumption 2: Given $N(t) = n$, the current probability distribution of the remaining time until the next death (service completion) is exponential with parameter μ_n .

✓ Assumption 3: The random variable of assumption 1 and the random variable of assumption 2 are mutually independent. The next transition in the state of the process is either $n \rightarrow n+1$ or $n \rightarrow n-1$ depending on whether the former or latter random variable is smaller.

✓ That is, the birth-and-death process can be illustrated by the rate diagram.

✓ The following analysis only focuses on the steady state condition.

✓ $E_n(t)$ = number of times that process enters state n by time t.

✓ $L_n(t)$ = number of times that process leaves state n by time t.

- ✓ $|E_n(t)) - {}_n(t)| \le 1$. Dividing both sides by t and letting $t \to \infty$.

$$\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| \le \frac{1}{t}\text{, so } \lim_{t\to\infty}\left|\frac{E_n(t)}{t} - \frac{L_n(t)}{t}\right| = 0 \quad .$$

- ✓ $\lim_{n\to\infty} \frac{E_n(t)}{t}$ = mean rate at which process enters state n.

  $\lim_{n\to\infty} \frac{L_n(t)}{t}$ = mean rate at which process leaves state n.

- ✓ **Rate In = Rate Out Principle**: for any state of the system n, mean entering rate = mean leaving rate. The equation expressing this principle is called the **balance equation** for state n.
  - ➢ P_i is the steady-state probability of being in state i.
  - ➢ Consider state 0: the mean entering rate of state 0 is $\mu_1 P_1$; the mean leaving rate of state is $\lambda_0 P_0$. The balance equation for state 0 is $\mu_1 P_1 = \lambda_0 P_0$
  - ➢ For every other state there are two possible transitions both into and out of the state. Therefore, each side of the balance equations for these states represents the sum of the mean rates for the two transitions involved.
  - ➢ We can write the balance equations for the other states.
  - ➢ Notice that there is always one “extra” variable in these equations.

    Solve P_1, P_2… in term of P_0.

- ✓ Thus, the steady-state probabilities are

  $P_n = C_n P_0$, where $C_n = \dfrac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n\,\mu_{n-1}\ldots\mu_1}$ for $n = 1, 2, \ldots$; when $n = 0$, $C_n = 1$.

- ✓ Then, use the sum of all probabilities equal 1 to solve for P_0.

$$\sum_{n=0}^{\infty} P_n = 1 \text{ implies that } \left(\sum_{n=0}^{\infty} C_n\right) P_0 = 1\text{, so that } P_0 = \Big(\sum_{n=0}^{\infty} C_n\Big)^{-1} .$$

- ✓ The key measures of performance for the queueing system (L, L_q, W, and W_q) can be obtained immediately after calculating the P_n.

$$L = \sum_{n=0}^{\infty} nP_n \,, \quad L_q = \sum_{n=s}^{\infty}(n-s)P_n \,,$$

$$W = \frac{L}{\bar{\lambda}} \,, \quad W_q = \frac{L_q}{\bar{\lambda}} \,,$$

Where $\bar{\lambda}$ is the average arrival rate over the long run and $\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$.

- ✓ The above calculations are based on the steady-state condition. Steady-state condition holds if $\lambda_n = 0$ or $\rho = \lambda/(s\mu) < 1$.

The *M*/*M*/*s* model

- ✓ All interarrival and service times are independently and identically distributed according to an exponential distribution. The number of servers is s.
- ✓ This model is a special case of the birth-and-death process where the queueing system's mean arrival rate and mean service rate per busy server are constant.
- ✓ When the system has just a single server ($s = 1$), the parameters for the birth-and-death process are $\lambda_n = \lambda$, and $\mu_n = \mu$. Rate diagram is as follow.
- ✓ When the system has multiple server ($s > 1$),
  - ➢ μ_n represents the mean service rate for the overall queueing system when there are n customers in the system.
  - ➢ The service rate per busy server is μ, the overall mean service rate for n busy servers must be $n\mu$.
  - ➢ Therefore, $\mu_n = n\mu$ when $n \leq s$, and $\mu_n = s\mu$ when $n \geq s$.
- ✓ When the maximum mean service rate $s\mu$ exceeds the mean arrival rate λ, that is, when $\rho = \frac{\lambda}{s\mu} < 1$, a queueing system will eventually reach a steady-state condition

We can use the results derived in the birth-and-death model.

Results for the Single-Server Case (*M*/*M*/1)

- ✓ The C_n factors reduce to $C_n = \frac{\lambda}{s\mu} = \rho^n$, *for n = 0, 1, 2,...*

- ✓ Therefore, $P_n = \rho^n P_0$, for n = *0, 1, 2,...,* *where* $Po = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1} = \left(\frac{1}{1-\rho}\right)^{-1} = 1-\rho$.

- ✓ Thus, $P_n = (1-\rho)\,\rho^n$, for n = *0, 1, 2,...*

- ✓ Consequently, $L = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \frac{\lambda}{\mu-\lambda}$

- ✓ $L_q = \sum_{n=1}^{\infty}(n-1)P_n = \frac{\lambda^2}{\mu(\mu-\lambda)}$

- ✓ When $\lambda \geq \mu$ the mean arrival rate exceeds the mean service rate, the preceding solution "blows up".

- ✓ Consider the case when $\lambda < \mu$ and the queue discipline is first-come-first-served. We can derive the probability distribution of the waiting time in the system W for a random arrival.

- ✓ If this arrival finds n customers already in the system, then the arrival will have to wait through $n+1$ exponential service times, including his/her own.

- ✓ Let $T_1, T_2, \ldots$ be independent service-time random variables having an exponential distribution with parameter μ, and let $S_{n+1}=T_1+T_2+\ldots+T_{n+1}$.

- ✓ $P\{W > t\} = \sum_{n=0}^{\infty} P_n \quad P\{S_{n+1} > t\} = e^{-\mu(1-\rho)t}$. That is, W has an exponential distribution

with parameter $\mu(1-\rho)$.

Therefore, $W = E(W) = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}$.

- ✓ Sometimes, we are concern about W_q , the waiting time in the queue.
- ✓ If this arrival finds no customers already in the system, there is no waiting time in queue. $P\{W_q = 0\} = P_0 = 1-\rho$.
- ✓ If this arrival finds $n > 0$ customers already in the system, then the arrival has to wait through n exponential service times until his/her own service begins.

- ✓ $P\{W_q > t\} = \sum_{n=1}^{\infty} P_n P\{S_n > t\} = \rho e^{-\mu(1-\rho)t}$.
- ✓ W_q does not quite have an exponential distribution, because $P\{W_q = 0\} > 0$.
- ✓ The conditional distribution of W_q, given that $W_q > 0$, does have an exponential distribution with parameter $\mu(1-\rho)$, because $P\{W_q > t \mid W_q > 0\} = \frac{P\{W_q > t\}}{P\{W_q > 0\}} = e^{-\mu(1-\rho)t}$.
- ✓ By deriving the mean of the (unconditional) distribution of W_q (or applying either $L_q = W_q$ or $W_q = W - \frac{1}{\mu}$), $W_q = E(W_q) = \frac{\lambda}{\mu(\mu-\lambda)}$.
- ✓ $P\{W_q > t\} = (1 - P\{W_q = 0\})e^{-s\mu(1-\rho)t}$, where $P\{W_q = 0\}) = \sum_{n=0}^{s-1} P_0$

Results for the Multiple-Server Case (*M/M/s*) — *s* > 1

- ✓ For $n = 1, 2, \ldots, s$, $C_n = \frac{(\lambda/\mu)^n}{n!}$. For $n = s, s+1, \ldots, C_n = \frac{(\lambda/\mu)^n}{s!s^{n-1}}$
- ✓ $P_0 =$

  $$1/\left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1 \quad \lambda/(s\mu)}\right].$$

- ✓ For $0 \le n \le s$, $P_n = \frac{(\lambda/\mu)^n}{n!} P_0$. For $n \ge s$, $P_n = \frac{(\lambda/\mu)^n}{s!s^{n-s}} P_0$.
- ✓ $L_q = P_0 = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2}$

  $W_q = L_q / \lambda$, $W = W_q + \frac{1}{\mu}$, $L = \lambda(W_q + \frac{1}{\mu}) = L_q + \frac{\lambda}{\mu}$

$$P\{W > t\} = e^{-\mu t}[1 + \frac{P_0(\lambda/\mu)^s}{s!\,(1-\rho)} + \left(\frac{1 - e^{-\mu t(s-1-\frac{\lambda}{\mu})}}{s - 1 - \lambda/\mu}\right)]$$

$P\{W_q > t\} = (1 - P\{W_q = 0\})e^{-s\mu(1-\rho)t}$, where $P\{W_q = 0\} = \sum_{n=0}^{s-1} P_0$