DS 632: System Simulation. Summer I 2014

Chapter 2: Basics of Queuing Theory
Notes

Queue and Queuing

Queue or waiting line is a part of many day-to-day applications, e.g., bank, grocer, post office, cafeteria, ticket counter at a game, etc. Waiting is associated with inefficiencies or loss of productivity.

Queuing theory is the study of waiting. It uses queuing models to represent various types of Queuing Systems (systems that have some waiting). Queuing models enable find an appropriate balance between the cost of service and the amount of waiting.

Basic Queuing Process

A queuing process consists of customers arriving at a service facility, then waiting in a line (queue) if all servers are busy and eventually receiving service, and finally departing from the facility. A queuing system is a set of customers, a set of servers, and an order whereby customers arrive and are served.

More generally, a queuing system is a birth-and-death process with a population consisting of customers either waiting for service or currently in service. A birth occurs when a customer arrives at the service facility; a death occurs when a customer departs from the facility. The state of the system is the number of customers in the facility.

Queue Characteristics

A Queuing System is characterized by five components:

1. Arrival patterns
2. Service patterns
3. Number of servers
4. Capacity of the facility to hold customers
5. Order in which customers are served

Arrival Patterns

Arrival pattern of customers is usually specified by the *interarrival time,* the time between successive customer arrivals to the service facility. It may be deterministic (i.e., known exactly), or it may be a random variable whose probability distribution is presumed known. It may depend on the number of customers already in the system, or it may be state-dependent.

Customers are generated over time by a Poisson process (i.e., the number of customers generated until any specific time, is a Poisson distribution). Probability distribution of time between consecutive arrivals is an exponential distribution.

Customers wait before being served.

Service Patterns

A service mechanism is involved in implementing the process. A service mechanism has three components:

- Servers (or service channels)
- Service facilities contain servers
- Service time (or holding time)

Service pattern is specified by the *service time,* the time required by one server to serve one customer. Service time may be deterministic, or it may be a random variable whose probability distribution is known. It may depend on the number of customers already in the facility, or it may be state dependent.

Service time distribution is the exponential distribution. Other types are Erlang (gamma) distribution.

Customer may be serviced by one server or requires a sequence of servers.

System Capacity

System capacity is the maximum number of customers, both those in service and those in the queue(s), permitted in the service facility at the same time. A system that has no limit on the number of customers permitted inside the facility has *infinite capacity*; a system with a limit has *finite capacity*.

Queue Discipline

The queuing process follows a queue discipline. The queue discipline is the order in which customers are served. This can be on a first-in, first-out (FIFO) basis (i.e., service in order of arrival), a last-in, first-out (LIFO) basis (i.e., the customer who arrives last is the next served), a random basis, or a priority basis.

Queuing Systems are labeled as follows:

v/w/x/y/z

where,

v indicates the distribution of inter arrival times,
w indicates the distribution of service times,
x signifies the number of available servers,
y represents system's capacity, and
z designates the queue discipline.

The following table gives representation of some of the models:

Queue Characteristic	Symbol	Meaning
Interarrival time or Service time	D M E_k G	Deterministic Exponentially distributed (Markovian) Erlang-type-k (k = 1, 2,…,) distributed Any other distribution
Queue discipline	FIFO LIFO SIRO PRI GD	First in, first out Last in, first out Service in random order Priority ordering Any other special ordering

Thus, a M/D/2/5/LIFO system has
- Exponentially distributed interarrival times
- Deterministic service times
- Two servers are available for providing service
- Has a limit of five customers
- Last customer to arrive is the next customer to receive service

A D/D/1 system has
- Deterministic interarrival time
- Deterministic service time
- Has one server

Queuing: Terminology and Notations

State of system = number of customers in queuing system.
Queue length = number of customers in waiting for service to begin.
= state of system minus number of customers being served.
$N(t)$ = number of customers in queuing system at time t ($t \geq 0$).
$P_n(t)$ = probability of exactly n customers in queuing system at time t, given number at time 0.
s = number of servers (parallel service channels) in queuing system..
λ_n = mean arrival rate (expected number of arrivals per unit time) of new customers when n customers are in system.
μ_n = mean service rate for overall system (expected number of customers completing service per unit time) when n customers are in system.

Note: μ_n represents combined rate at which all busy servers (those serving customers) achieve service completions.

λ = when λ_n is a constant for all n.
μ = when the mean service rate per busy server is a constant for all $n \geq 1$.
$\mu = s\mu$ when $n \geq s$ (when all s servers are busy).
$\frac{1}{\lambda}$ = expected interarrival time
$\frac{1}{\mu}$ = expected server time
$\rho = \frac{\lambda}{s\mu}$ is the utilization factor for service facility

System States

Transient state is state of the system affected by initial state and by the time that has since elapsed.

A system has reached a *steady state* condition, where the probability distribution of the state of the system remains the same over time.

Steady State Solutions. If a system is in steady state, then:

P_n = probability of exactly n customers in queuing system.

L = expected number of customers in queuing system = $\sum_{n=0}^{\infty} nP_n$

L_q = expected queue length = $\sum_{n=s}^{\infty} (n-s)P_n$

ω = waiting time in system for each individual customer

$W = E(\omega)$

ω_q = waiting time in queue for each individual customer

$W_q = E(\omega_q)$

Relationship between L, W, L_q, and W_q

$L = \lambda W$

$L_q = \lambda W_q$

$W = W_q + \frac{1}{\mu}$

Role of Exponential Distribution in Queuing Theory

Suppose that a random variable (r. v.) T, represents either interarrival or service times. This r. v. is said to have an exponential distribution with parameter α, if its p.d.f. is

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

The cumulative probabilities are

$$P(T \leq t) = 1 - e^{-\alpha t}$$
$$P(T > t) = e^{-\alpha t} \quad (t \geq 0)$$

The expected value and variance of T are:

$$E(T) = \frac{1}{\alpha}$$
$$Var(T) = \frac{1}{\alpha^2}$$

M/M/1 Systems

An M/M/1 system is a queuing system having:

- Exponentially distributed interarrival times, with parameter λ,
- Exponentially distributed service times, with parameter μ,
- One server,

- No limit on system capacity, and
- Queue discipline of FIFO.

Where,

λ is the average customer arrival rate.
μ is the average service rate of customers.
$\frac{1}{\lambda}$ is the expected interarrival time.
$\frac{1}{\mu}$ is the expected time to serve a customer.

Since exponentially distributed interarrival times with mean $\frac{1}{\lambda}$ are equivalent over a time interval τ, to a Poisson distributed arrival pattern with mean $\lambda\tau$, M/M/1 systems are often referred to as single-server, infinite-capacity, queuing systems having Poisson input and exponential service times.

Markovian Model birth and death process

An M/M/1 is a Poisson birth-and-death process. In this system, inputs (arriving customers) and outputs (leaving customers) occur according to a birth-and-death process.

Birth → arrival of a new customer into the queuing system
Death → departure of a served customer

State of the system at time t ($t \geq 0$) → N(t) [number of customers in the queuing system at time t]

Birth-and-death process describe probabilistically how N(t) changes as t increases.

Assumption 1. Given N(t) = n, current probability of remaining time until the next birth is exponential with parameter λ_n (n = 0, 1, 2, …).

Assumption 2. Given N(t) = n, current probability of remaining time until the next death is exponential with parameter μ_n (n = 0, 1, 2, …).

Assumption 3. Random variable of assumption 1 and random variable of assumption 2 are mutually independent.

The next transition in the state of the process is either,

$n \rightarrow n+1$ (a single birth)

or

$n \rightarrow n-1$ (a single death)

The birth-and-death process is a special type of continuous time Markov Chain. The following principle holds for this process:

Rate In = Rate Out. For any state of the system n (n = 0, 1, 2, …) mean entering rate = mean leaving rate. This is called the **balance equation** for state n. After constructing the balance equations for all states in terms of the unknown P_n probabilities, we can solve this system of equations, (plus an equation stating that the probabilities must sum to 1) to find these probabilities.

Balance Equations Table

State	Rate In = Rate Out
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$
.	
.	

Using the above principles, yields following probabilities:

State:	
0	$P_1 = \frac{\lambda_0}{\mu_1} P_0$
1	$P_2 = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$
n	$P_{n+1} = \frac{\lambda_n}{\mu_{n+1}} P_n = \frac{\lambda_n \lambda_{n-1} \ldots \lambda_0}{\mu_n \mu_{n-1} \ldots \mu_1} P_0$

Let $C_n = \frac{\lambda_{n-1} \lambda_{n-2} \ldots \lambda_0}{\mu_n \mu_{n-1} \ldots \mu_1}$, for n = 1, 2, …,

define $C_n = 1$, for n = 0.

The steady state probabilities are:

$P_n = C_n P_0$, for n = 0, 1, 2, …

$$P_0 = \left(\sum_{n=0}^{\infty} C_n\right)^{-1}$$

Key Measures of Performance for Queuing Systems

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

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

where $\bar{\lambda}$ is the average arrival rate over the long run

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$$

Measures of Performance for M/M/1 Queuing System

$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n$, for n = 0, 1, 2, …

$P_n = \rho^n P_0$, for n = 0, 1, 2, …

where $P_0 = 1 - \rho$

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

$$L = \left(\frac{\rho}{1-\rho}\right) = \frac{\lambda}{\mu-\lambda}$$

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

$$W = \frac{1}{\mu-\lambda}$$

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)}$$