Queuing Theory

RK Jana

1

Preliminaries

A flow of customers coming towards the service facility forms a queue on account of lack of capacity to serve them all at a time.

Some Examples:

- Persons waiting at doctor's clinic
- Persons waiting at railway booking office
- Machines waiting to be repaired
- Ships waiting in the harbor to be unloaded
- Airplanes take off, landing

Customers may be: persons, machines, vehicles, parts etc

2

Applications of Queuing Theory

- Telecommunications
- Traffic control
- Determining the sequence of computer operations
- Predicting computer performance
- Health services (eg. control of hospital bed assignments)
- Airport traffic, airline ticket sales
- Layout of manufacturing systems

3

Basic Concepts

Input Process ***Queue*** ***Service Mechanism***

Queuing system: Customers ***arriving*** for service, ***waiting*** for ***service*** if it is not immediate, leaving the system after being served.

The theoretical study of waiting lines, expressed in mathematical terms.

4

Input Process

- Source (population size)
  - Finite
  - Infinite
- Arrival behavior
  - One by one
  - Batch
    - Fixed size
    - Variable size
  - Change of arrival pattern
    - Stationary
    - Non-stationary/Transient (time dependent)

5

Queue Discipline

- Queue size
  - Finite
  - Infinite
- Queue discipline
  - First come, first served (FCFS)
  - Last come, first served (LCFS)
  - Random selection for service (RSS)
  - Priority queue
    - Preemptive
    - non-preemptive

6

Service Mechanism

- Bulk service (computer with parallel processing, bus service)
- State dependent (service depends on number of waiting customers. Example: internet access)
- State independent
- Stationary/non-stationary

7

Number of Service Channels

- Single channel
- Parallel channels (provides identical service)
- Series (customers go through a number of services, public offices, manufacturing process)

8

Capacity of the System

- Finite source queue (finite capacity of waiting room, restriction on queue length)
- Infinite (no restriction on queue length)

9

QT in Performance Measurement

- Analyze waiting time distribution
- To know average waiting time of customer
- To know queue length distribution
- Calculate current work backlog
- Measurement of the idle time of server
- Measurement of the busy time of server
- System utilization

10

Questions In Queuing Systems

1. The number of people in the system (those being served and waiting in line).
2. The number of people in the queue (waiting for service).
3. The waiting time in the system (the interval between when an individual enters the system and when he or she leaves the system).
4. The waiting time in the queue (the time between entering the system and the beginning of service).

11

Notations

n:	# customers in the system (in queue & in service)
$P_n(t)$:	Transient state probability of exactly n customers in the system at time t (it is assumed that the system starts its operation at time zero)
P_n:	Steady state probability of exactly n customers in the system
E(s):	Expected number of customers in the system
E(q):	Expected number of customers in the queue = E(s) – Number of customers being served

12

Notations: Continued…

λ_n : Expected number of arrivals per unit time (mean arrival rate) of customers when *n* customers are present in the system

μ_n : Expected number of customers served per unit time (mean service rate) when *n* customers are present in the system

λ: Mean arrival rate when λ_n is constant for all n

μ: Mean service rate when μ_n is constant for all $n \geq 1$

E(w_1): Expected waiting time per customer in the system

E(w_2): Expected waiting time per customer in the queue

13

The Poisson Process

- Axiom 1: The number of arrivals in non-overlapping intervals are statistically independent
- Axiom 2: The probability of more than one arrival between time t and time $(t + \Delta t)$ is $O(\Delta t)$ i.e., the probability of more two or more arrivals during the small time is negligible.
- Axiom 3: The probability that an arrival occurs between time t and time $(t + \Delta t)$ is $\{\lambda.\Delta t + O(\Delta t)\}$

14

The Arrival Theorem

If the arrivals are completely random, then the probability distribution of the number of arrivals in a fixed time interval follows Poisson distribution.

$$P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!},\ n = 0,1,2,\cdots$$

15

The Role of Exponential Distribution

Most analytic results for queuing situations involve the exponential distribution either as the distribution of inter-arrival times or service times or both.

The following three properties help to identify the set of circumstances in which it is reasonable to assume that an exponential distribution will occur.

- **Lack of memory**
- **Small service times**
- **Relation to the Poisson distribution**

16

Lack of Memory

In an arrival process, this property implies that the probability that an arrival will occur in the next few minutes is not influenced by when the last arrival occurred.

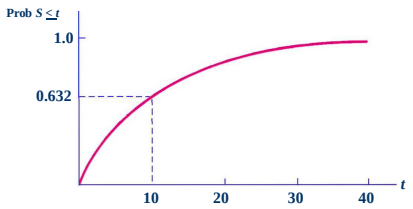
(a) There are many individuals who could potentially arrive at the system

(b) Each person decides to arrive independently of the other individuals

(c) Each individual selects his or her time of arrival completely at random

17

Small Service Time



This graph shows the probability that the service time *S* is less than or equal to *t* if the mean service time is 10.

18

Continued…

The graph showed that more than 63% of the service times were smaller than the average service time (10).

Compare this to the normal distribution where only 50% of the service times are smaller than the average.

The practical implication is that an exponential distribution can best be used to model the distribution of service times in a system in which a large proportion of "jobs" take a very short time and only a few "jobs" run for a long time.

19

Relation with Poisson Distribution

If the time between arrivals has an exponential distribution with parameter, then in a specified period of time the number of arrivals will have a Poisson distribution.

20

Distribution of Inter-arrival Time

Let *T* be the time between two consecutive arrivals. If the arrivals on *n*-customers in time *t* follows Poisson distribution the *T* follows exponential distribution.

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & 0 \le t < \infty \\ 0 \quad , elsewhere \end{cases}$$

21

The Mean Arrival Time

$$E(T) = \int_{t=0}^{\infty} t f(t)\, dt$$
$$= \int_{t=0}^{\infty} \lambda t e^{-\lambda t}\, dt$$
$$= \frac{1}{\lambda}$$

22

Distribution of Service Time

If T be the inter-arrival time, then the probability of n-complete service in time T is given by:

$$\phi_n(t) = P(n\text{-service in time } T)$$
$$= \frac{e^{-\mu t}(\mu t)^n}{n!}, \; n = 0, 1, 2, \cdots$$

23

The Traffic Intensity

For Poisson arrival and departure with one server, the traffic intensity (ρ) is given by:

$$\rho = \frac{\text{mean arrival rate}}{\text{mean service rate}} = \frac{\lambda}{\mu}$$

The unit of ρ is Erlang

24

Queuing Models

The general model can be completely represented by Kendall's notation as follows:

(a / b / c) : (d / e / f)

a ≡ arrival distribution
b ≡ service distribution
c ≡ # service channels
d ≡ capacity of the system
e ≡ service discipline
f ≡ size of calling source

Standard Notations:

M ≡ Poisson arrival or departure distribution
E_k ≡ Earlangian or Gamma arrival or departure distribution
GI ≡ General Independent arrival distribution

25

(M / M / 1) : (∞ / FCFS / ∞)

Single channel infinite population model

In a steady state condition ρ < 1, it can be shown that

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho,\ \rho < 1$$

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) = \rho^n (1-\rho),\ \rho < 1, n \geq 0$$

26

Characteristics of the Model

1. Expected number of customers in the system *E(n)*:

$$E(n) = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\rho \sum_{n=0}^{\infty} n\rho^{n-1}$$

$$= (1-\rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n) = (1-\rho)\rho \frac{d}{d\rho} \sum_{n=0}^{\infty} \rho^n$$

$$= (1-\rho)\rho \frac{d}{d\rho}\left(\frac{1}{(1-\rho)}\right) = \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

27

Continued…

2. Expected queue length *E(q)*: Since there is one server, one customer is in service & *(n-1)* customers are in the queue.

$$E(q) = \sum_{n=1}^{\infty} (n-1)P_n = \sum_{n=1}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n$$

$$= \sum_{n=0}^{\infty} nP_n - \left(\sum_{n=0}^{\infty} P_n - P_0\right) = \frac{\rho}{1-\rho} - \{1-(1-\rho)\}$$

$$= \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)},\ \text{since} \sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} (1-\rho)\rho^n = 1$$

28

Continued…

3. Probability of queue size greater than some finite number *N*:

$$P(\text{Queue size } \geq N) = \sum_{n=N}^{\infty} P_n = \sum_{n=N}^{\infty} (1-\rho)\rho^n$$

$$= (1-\rho)\rho^N \sum_{n=N}^{\infty} \rho^{n-N} = (1-\rho)\rho^N \sum_{r=0}^{\infty} \rho^r,\ \text{where } r = n - N$$

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

29

Continued…

4. Expected waiting time per customer in the system $E(w_1)$:

$$E(w_1) = \frac{\text{Expected \# customers in the system}}{\text{Arrival rate}} = \frac{E(n)}{\lambda} = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1-\rho)}$$

5. Expected waiting time per customer in the queue $E(w_2)$:

$$E(w_2) = E(w_1) - \text{service time of one customer} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$$

30

Example: 1

In a railway marshalling yard, goods trains arrive at a rate of 30 trains per day. Assuming that the inter-arrival time follows an exponential distribution and the service time (time taken to hump a train) distribution as Poisson with an average of 36 minutes, calculate:

(i) The average number of trains in the system.

(ii) The probability that the queue size exceeds 10.

(iii) Expected waiting time in the queue.

(iv) Average number of trains in the queue.

31

Solution: 1

$\lambda = 30/(24\times60) = 1/48$ trains/min

$\mu = 1/36$ trains/min

$\rho = \lambda/\mu = 36/48 = 0.75$

(i) Average number of trains in the system

$= \rho / (1 - \rho) = 0.75/(1 - 0.75) = 3$ trains

(ii) $P(\text{Queue size} \geq 10) = \rho^{10} = (0.75)^{10} = 0.06$

(iii) Expected waiting time in the queue: $\frac{\rho}{\mu(1-\rho)} = 108$ min = 1 hr 48 min

(iv) Average number of trains in the queue: $\frac{\rho^2}{1-\rho} = 2.25$ i.e., nearly 2 trains

32

Example: 2

In a single serve system, the arrival rate $\lambda = 5$ per hour and the service rate $\mu = 8$ per hour. Assuming the conditions of for the single channel queuing model, find out:

(i) The probability that the system is idle.

(ii) The probability that the queue size is at least 2.

(iii) Expected time that a customer is in the queue.

(iv) The probability that a customer is being served and nobody is waiting.

(v) Expected time a customer spends in the system.

(vi) Average number of customers in the queue.

33

Solution: 2

$\lambda = 5$ per hr, $\mu = 8$ per hr

$\rho = \lambda/\mu = 5/8 = 0.625$

(i) Probability that the system is idle

$= P$ (No customer in the system) $= 1 - \lambda/\mu$

(ii) P (At least 2 customers in the system) $= P(n \geq 2) = (\lambda/\mu)^2$

(iii) Expected time a customer is in the queue: $\lambda/\mu(\mu - \lambda)$

(iv) The probability that a customer is being served and nobody is waiting $= P_1 = (1 - \lambda/\mu)\lambda/\mu = (1 - \rho)\rho$

(v) Expected time a customer spends in the system $= 1/(\mu - \lambda)$

(vi) Average number of customers in the queue $= \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho}$

34

(M / M / 1) : (N / FCFS / ∞)

Single channel finite population model

Maximum number of customers allowed in the system: N

Maximum queue length: (N – 1)

N customers are in the system: No new arrival is permissible

In a steady state condition $\rho < 1$, it can be shown that

$$P_0 = \begin{cases} \frac{1-\rho}{1-\rho^{N+1}}, & \rho \neq 1 \\ \frac{1}{N+1}, & \rho = 1 \end{cases} \qquad P_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{N+1}}, & \rho \neq 1 \\ \frac{1}{N+1}, & \rho = 1 \end{cases}$$

35

Continued…

In this model, $\rho = \lambda/\mu$ may be $\geq$ or ≤ 1 for steady state condition. Because, the number of customers allowed in the system is controlled by the queue length and not by the relative rates of arrival (λ) or departure (μ).

36

Characteristics of the Model

1. Expected number of customers in the system $E(n)$:

$$E(n)=\begin{cases}\rho\left[\dfrac{1-(N+1)\rho^{N}+N\rho^{N+1}}{(1-\rho)(1-\rho^{N+1})}\right], & \rho\neq 1\\ \dfrac{N}{2}, & \rho=1\end{cases}$$

Characteristics of the Model

2. Expected queue length $E(q)$:

$$E(q)=\begin{cases}\rho\left[\dfrac{1-(N+1)\rho^{N}+N\rho^{N+1}}{(1-\rho)(1-\rho^{N+1})}\right]+\dfrac{(1-\rho)}{(1-\rho^{N+1})}, & \rho\neq 1\\ \dfrac{N}{2}+\dfrac{1}{N+1}-1, & \rho=1\end{cases}$$

Continued…

3. Expected waiting time per customer in the system $E(w_1)$:

$$E(w_1)=\begin{cases}\dfrac{1}{\mu}\left[\dfrac{1}{(1-\rho)}+\dfrac{N\rho^{N}}{(1-\rho^{N})}\right], & \rho\neq 1\\ \dfrac{N+1}{2\lambda}, & \rho=1\end{cases}$$

4. Expected waiting time per customer in the queue $E(w_2)$:

$$E(w_2)=\begin{cases}\dfrac{1}{\mu}\left[\dfrac{1}{(1-\rho)}+\dfrac{N\rho^{N}}{(1-\rho^{N})}-1\right], & \rho\neq 1\\ \dfrac{N+1}{2\lambda}-\dfrac{1}{\mu}, & \rho=1\end{cases}$$

Example: 3

Assume that the goods trains are coming in a yard at the rate of 30 trains per day and suppose that the inter-arrival time follows an exponential distribution. The service time for each train is assumed to be exponential with an average of 36 minutes. If the yard can admit 9 trains at a time, then calculate:

(i) The probability that the yard is empty.
(ii) The expected number of trains in the yard.
(iii) The expected number of trains in the queue.
(iv) Expected waiting time of a train in the yard.
(v) Expected waiting time of a train in the queue.

Solution: 3

$\lambda = 30/(24\times 60) = 1/48$ trains/min
$\mu = 1/36$ trains/min
$\rho = \lambda/\mu = 36/48 = 0.75 \neq 1$

(i) Probability that the yard is empty

$$P_0=\frac{1-\rho}{1-\rho^{N}} \text{ for } \rho\neq 1$$
$$=\frac{1-0.75}{1-(0.75)^{9}}=0.28$$

Solution 3: Continued…

(ii) The expected number of trains in the yard:

$$E(n)=\rho\left[\frac{1-(N+1)\rho^{N}+N\rho^{N+1}}{(1-\rho)(1-\rho^{N+1})}\right],\ \rho\neq 1$$
$$=0.75\left[\frac{1-10(0.75)^{9}+9(0.75)^{10}}{(1-0.75)\{1-(0.75)^{10}\}}\right]$$

Solution 3: Continued…

(iii) Expected number of trains in the queue:

$$E(q)=\rho\left[\frac{1-(N+1)\rho^{N}+N\rho^{N+1}}{(1-\rho)(1-\rho^{N+1})}\right]+\frac{(1-\rho)}{(1-\rho^{N+1})},\ \rho\neq 1$$

$$=0.75\left[\frac{1-10(0.75)^{9}+9(0.75)^{10}}{(1-0.75)\{1-(0.75)^{10}\}}\right]+\frac{(1-0.75)}{\{1-(0.75)^{10}\}}$$

Solution 3: Continued…

(iv) Expected waiting time of a train in the yard:

$$E(w_1)=\frac{1}{\mu}\left[\frac{1}{(1-\rho)}+\frac{N\rho^{N}}{(1-\rho^{N})}\right],\ \rho\neq 1$$

$$=36\left[\frac{1}{(1-0.75)}+\frac{9(0.75)^{9}}{\{1-(0.75)^{9}\}}\right]$$

(v) Expected waiting time of a train in the queue:

$$E(w_2)=\frac{1}{\mu}\left[\frac{1}{(1-\rho)}+\frac{N\rho^{N}}{(1-\rho^{N})}-1\right],\ \rho\neq 1$$

$$=36\left[\frac{1}{(1-0.75)}+\frac{9(0.75)^{9}}{\{1-(0.75)^{9}\}}-1\right]$$