QUEUING THEORY

M.Venkatasami,
Ph.D. (Processing and Food Engineering)
Department of Food Process Engineering
AEC&RI, TNAU.

INTRODUCTION

Queues → Waiting lines

Formation of queue

- Production/operation system
- Number of customers exceeds the number of service facilities
- Service facilities do not work efficiently
- More time than prescribed to serve a customer

E.g., Bus stops, petrol pumps, restaurants, ticket booths, doctors' clinics, bank counters

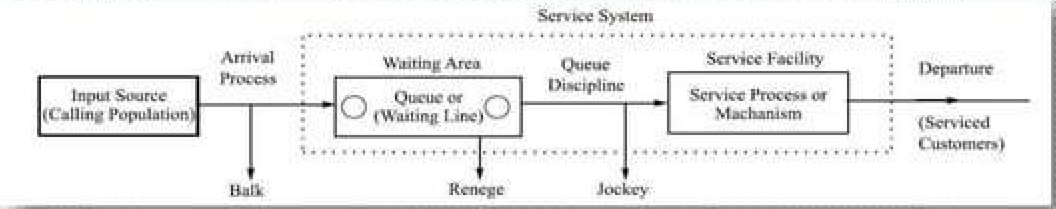
SITUATIONS

The arrival rate (or time) of customers ⇐ **Not possible to accurately predict** ⇒ Service rate (or time) of service facility or facilities.

- Used to determine the level of service (either the service rate or the number of service facilities)
- Balances the following two conflicting costs

1. Cost of offering the service	2. Cost incurred due to delay in offering service
• Service facilities and their operation	• Cost of customers waiting for service

THE STRUCTURE OF A QUEUING SYSTEM



The major components

- Calling population (or input source)
- Queuing process
- Queue discipline
- Service process (or mechanism)

CALLING POPULATION CHARACTERISTICS

- 1. Size of calling population
  - Homogeneous
    - Finite
    - Infinite
  - Subpopulations
    - Finite
    - Infinite

- 2. Behavior of the arrivals
  - Patient customer
  - Impatient customer
  - Balking customers
  - Reneging customers
  - Jockeying customers

- 3. Pattern of arrivals at the system
  - Static arrival
  - Dynamic arrival

❖ Batches
❖ Individually
❖ Scheduled time
❖ Unscheduled time

ARRIVAL TIME DISTRIBUTION

Probability distributions

- Poisson distribution
- Exponential distribution
- Erlang distribution

Poisson distribution

- Describes the arrival rate variability
- Number of random arrivals at a service facility in a fixed period of time.

Exponential distribution

- Describes the average time between arrivals (inter-arrival time) when arrival rate is Poisson

PROBABILITY DISTRIBUTION FUNCTION

- No of customers arrive = n
- Time interval = 0 to t
- The expected (or average) number of arrivals per time unit = λ
- The expected number of arrivals in a given time interval 0 to t = λt

Poisson probability distribution function

$$P(x=n)= e^{-\lambda t}\left((\lambda t)^n / n!\right) \qquad \text{for } n=0,1,2,\ldots$$

- The probability of no arrival in the given time interval 0 to t

$$P(x=0)= e^{-\lambda t}\left((\lambda t)^0/0!\right)= e^{-\lambda t} \qquad \text{for } n=0,1,2,\ldots$$

Cont.

- The time between successive arrivals = T (continuous random variable)
- A customer can arrive at any time

The probability of no arrival in the time interval 0 to t = The probability that T exceeds t.

$$P(T>t)=P(x=0)=e^{-\lambda t}$$

The cumulative probability → The time T between two successive arrivals is t or less

$$P(T \le t) = 1 - P(T > t) = 1 - e^{-\lambda t}\ ;\ t \ge 0$$

Cont.

- The expression for $P(T \leq t)$ ⟹ the cumulative probability distribution function of T.
- The distribution of the random variable T is referred to as the exponential distribution,
- whose probability density function can be written as follows:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{For } \lambda, t \geq 0 \\ 0 & \textit{otherwise} \end{cases}$$

Poisson distribution ⟹ Arrival of customers at a service system, $\mu = \sigma = \lambda$

Exponential distribution ⟹ The time between successive arrivals, $\mu = \sigma = 1/\lambda$

QUEUING PROCESS

- Refers to the number of queues – single, multiple or priority queues and their lengths

The type of queue	The layout of service mechanism
The length (or size) of a queue	Operational situations such as physical space, legal restrictions, and attitude of the customers

- Finite (or limited) source queue.
- Infinite (or unlimited) source queue
- Multiple queues - finite or infinite

QUEUE DISCIPLINE

- The order (or manner) in which customers from the queue are selected for service

Static Queue Disciplines

- First-come, firstserved (FCFS)
- Last-come, first-served (LCFS)

Dynamic Queue Disciplines

- Service in random Order (SIRO)
- Priority service
- Pre-emptive priority (or Emergency)
- Non-pre-emptive priority

SERVICE PROCESS (OR MECHANISM)

- The service mechanism (or process) is concerned with the manner in which customers are serviced and leave the service system

- The arrangement (or capacity) of service facilities
- The distribution of service times

THE ARRANGEMENT (OR CAPACITY) OF SERVICE FACILITIES

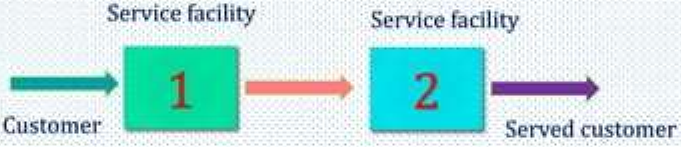
- Series arrangement
- Parallel arrangement
- Mixed arrangement

ARRANGEMENT OF SERVICE FACILITIES

SERIES ARRANGEMENT

Single Queue, Single Serve

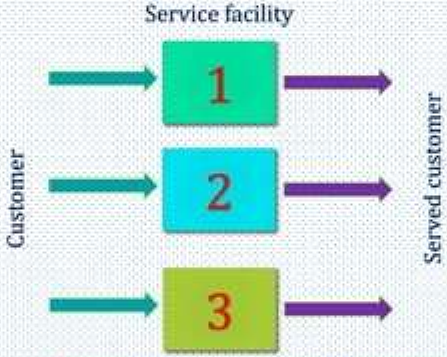
Single Queue, Multiple Servers



Cont.

PARALLEL ARRANGEMENT

Single Queue, Multiple Service



Multiple Queue, Multiple Servers

Cont.

MIXED ARRANGEMENT

Single Queue, Multiple Service

SERVICE TIME DISTRIBUTION

- The time taken by the server from the commencement of service to the completion of service for a customer is known as the service time.

AVERAGE SERVICE RATE

- The service rate measures the service capacity of the facility in terms of customers per unit of time
- μ is the average service rate
- The expected number of customers served during time interval 0 to t will be μt.

If service starts at zero time, the probability that service is not completed by time t is given by,

$$P(x = 0) = e^{-\mu t}$$

Cont.

- Service time = T (random variable)
- The probability of service completion within time t is given by:

$$P(T \leq t) = 1 - e^{-\mu t}, t \geq 0$$

AVERAGE LENGTH OF SERVICE TIME

- The fluctuating service time is described by the negative exponential probability distribution, denoted by $1/\mu$

Queue size	• Average number of customers waiting in the system for service
Queue length	• Average number of customers waiting in the system and being served

PERFORMANCE MEASURES OF A QUEUING SYSTEM

- In steady state systems, the operating characteristics do not vary with time

NOTATIONS

n	Number of customers in the system (waiting and in service)
Pn	Probability of n customers in the system
λ	Average customer arrival rate or average number of arrivals per unit of time in the queuing system
μ	Average service rate or average number of customers served per unit time at the place of service
Po	Probability of no customer in the system
s	Number of service channels (service facilities or servers)
N	Maximum number of customers allowed in the system

Cont.

Ls	Average number of customers in the system (waiting and in service)
Lq	Average number of customers in the queue (queue length)
Ws	Average waiting time in the system (waiting and in service)
Wq	Average waiting time in the queue
Pw	Probability that an arriving customer has to wait (system being busy), 1 – Po = (λ/μ)

$$\frac{\lambda}{\mu} = \rho = \frac{\text{Average service completion time } (1/\mu)}{\text{Average interarrival time } (1/\lambda)}$$

ρ : Percentage of time a server is busy serving customers, i.e., the system utilization

GENERAL RELATIONSHIPS

LITTLE'S FORMULA

$Ls = \lambda\, Ws$

$L_q = \lambda\, W_q$

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

$Ls = L_q + \frac{\lambda}{\mu}$

- Valid for all queueing models
- Developed by J. Little
- If the queue is finite, λ is replaced by λe

QUEUING MODEL

Traditional queuing theory is concerned with obtaining closed form solutions for,

Steady state probabilities $p_n = P(N=n)$

or

The performance measures Ls, L_q, Ws, and W_q for simple queuing systems

CLASSIFICATION OF QUEUING MODELS

- QT models are classified by using special (or standard) notations
- Described initially by D.G. Kendall in the form (*a/b/c*)
- A.M. Lee added the symbols *d* and *c* to the Kendall's notation.

Cont.

- The standard format used to describe queuing models is as follows:

$$\{(a/b/c) : (d/c)\}$$

- a = arrivals distribution
- b = service time distribution
- c = number of servers (service channels)
- d = capacity of the system (queue plus service)
- e = queue (or service) discipline

- In place of notation a and b, other descriptive notations are used for the arrival and service times distribution:

M = Markovian (or Exponential) interarrival time or service-time distribution
D = Deterministic (or constant) interarrival time or service time
GI = General probability distribution – normal or uniform for inter-arrival time

Cont.

In a queuing system,

M/M/1

M	• The number of arrivals is described by a Poisson probability distribution, λ
M	• The service time is described by an exponential distribution, μ
1	• A single server

$\frac{\lambda}{\mu} < 1, Infinite\ queue\ length\ models$

$\frac{\lambda}{\mu} > 1\ \ Finite\ queue\ length\ models$

Single Server → *Finite queue length*, *Infinite queue length*

Multiple server → *Finite queue length*, *Infinite queue length*

SINGLE-SERVER QUEUING MODELS

Model I: {(M/M/1): (∞/FCFS)} Exponential Service – Unlimited Queue

(A) Expected number of customers in the system

$$L_s = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}; \quad \rho = \frac{\lambda}{\mu}$$

(B) Expected number of customers waiting In the queue

$$L_q = \frac{\lambda}{\lambda-\mu} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)}; \; 1-P_0 = \frac{\lambda}{\mu}$$

(C) Expected waiting time for a customer in the queue:

$$W_q = \lambda\left(1-\frac{\lambda}{\mu}\right)\frac{1}{(\mu-\lambda)^2} = \frac{\lambda}{\mu(\mu-\lambda)} \text{ or } \frac{L_q}{\lambda}$$

(d) Expected waiting time for a customer in the system

$$W_s = W_q + \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu} = \frac{1}{\mu-\lambda} \text{ or } \frac{L_s}{\lambda}$$

Cont.

Model II: {(M/M/1) : (∞/SIRO)} $P_n = (1-\rho)\,\rho^n \,;\; n = 1, 2, \ldots$

- Identical to the model I with the only difference in queue discipline
- The derivation of P_n is independent of any specific queue discipline
- Other results will also remain unchanged as long as P_n remains unchanged

Model III: {(*M*/*M*/1) : (*N*/FCFS)} Exponential Service – Finite (or Limited) Queue

(A) Expected number of customers in the system

$$L_s = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(N+1)\,\rho^{N+1}}{1-\rho^{N+1}} \,; & \rho \neq 1\ (\lambda \neq \mu) \\[2ex] \dfrac{N}{2} \,; & \rho = 1\ (\lambda = \mu) \end{cases}$$

Cont.

Model III: {(M/M/1) : (N/FCFS)} Exponential Service – Finite (or Limited) Queue

- Expected number of customers waiting in the queue:

$$L_q = L_s - \frac{\lambda}{\mu} = L_s - \frac{\lambda(1-P_N)}{\mu}$$

- Expected waiting time of a customer in the system (waiting + service):

$$W_s = \frac{L_q}{\lambda(1-P_N)} + \frac{1}{\mu} = \frac{L_s}{\lambda(1-P_N)}$$

- Expected waiting time of a customer in the queue:

$$W_q = W_s - \frac{1}{\mu} \text{ or } \frac{L_q}{\lambda(1-P_N)}$$

MULTI-SERVER QUEUING MODELS

Model IV: {(M/M/s) : (∞/FCFS)} Exponential Service – Unlimited Queue

The expected number of customers waiting in the queue (length of line):

$$L_q = \left[\frac{1}{(s-1)!} \left(\frac{\lambda}{\mu} \right)^s \frac{\lambda\mu}{(s\mu - \lambda)^2} \right] P_0$$

Cont.

The expected number of customers in the system:

$$L_s = L_q + \frac{\lambda}{\mu}$$

The expected waiting time of a customer in the queue:

$$W_q = \left[\frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^s \frac{\mu}{(s\mu-\lambda)^2}\right]P_0 = \frac{L_q}{\lambda}$$

The expected waiting time that a customer spends in the system:

$$W_s = W_q + \frac{1}{\mu} = \frac{L_q}{\lambda} + \frac{1}{\mu}$$

Cont.

Model V: $\{(M/M/s) : (N/FCFS)\}$ Exponential Service – Limited (Finite) Queue

The expected number of customers in the queue

$$L_q = \frac{(s\rho)^s \rho}{s!(1-\rho)^2} \left[1-\rho^{N-s+1} - (1-\rho)(N-s+1)\rho^{N-s}\right] P_0$$

Cont.

The expected number of customers in the system:

$$L_s = L_q + \left(\frac{\lambda}{\mu}\right)(1 - P_N) = L_q + s - P_0 \sum_{n=0}^{s-1} \frac{(s-n)}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

The expected waiting time in the system:

$$W_s = \frac{L_s}{\lambda\,(1 - P_N)}$$

The expected waiting time in the queue:

$$W_q = W_s - \frac{1}{\mu} = \frac{L_q}{\lambda(1 - P_N)}$$

FINITE CALLING POPULATION QUEUING MODELS

- Model VI: {(M/M/1) : (M/GD)} Single Server – Finite Population (Source) of Arrivals
- Model VII: {(M/M/s) : (M/GD)} Multiserver – Finite Population (Source) of Arrivals

MULTI-PHASE SERVICE QUEUING MODEL

- Model VIII: {(M/Ek / 1) : (∞ / FCFS)} Erlang Service Time Distribution with k-Phases

SPECIAL PURPOSE QUEUING MODELS

- Model IX: Single Server, Non-Exponential Service Times Distribution – Unlimited Queue
- Model X: Single Server, Constant Service Times – Unlimited Queue

THANK YOU