Queuing Theory & its Applications

Prof. (Dr.) K.R. Chowdhary
Email: kr.chowdhary@jietjodhpur.ac.in

Campus Director,
JIET College of Engineering, Jodhpur

Thursday 6th April, 2017

Queuing Theory Notation

Queuing characteristics:

- arrival process
- Service time distribution
- Number of servers
- System capacity
- Population size
- Service discipline

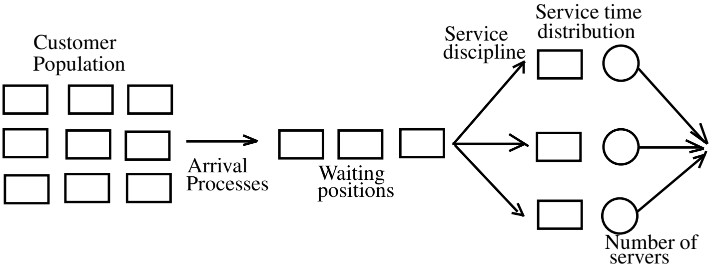

Figure 1: A Queuing system

Arrival Processes

Suppose jobs arrive at times $t_1, t_2, ..., t_j$

- Random variables $\tau_j = t_j - t_{j-1}$ are *inter-arrival times*
- There are many possible assumptions for the distribution of the τ_j. Typical assumptions for the τ_j :
  - Independent
  - Identically distributed
- Many other possible assumptions:
  - Bulk arrivals
  - Balking
  - Correlated arrivals

For Poisson arrival, the inter-arrival times are:

- IID (independent and identically distributed)
- exponentially distributed (i.e., $F(x) = 1 - e^{-x/a}$)

Other Queue Features

Service time:

- Interval spent actually receiving service (exclusive of waiting time)
- Like with arrival processes, there are many possible assumptions:
  - IID random variables
  - exponential service time distribution

Number of servers:

- Servers may or may not be identical
- Service discipline determines allocation of customers to servers

System capacity:

- Maximum no. of customers in system
- May be finite or infinite

Population size:

- Total number of potential customers
- May be finite or infinite

Service discipline:

- The order in which waiting customers are serviced
- Many possibilities, including
  - First-come-first-serve (FCFS), the most common
  - Last-come-first-serve (LCFS)
  - Last-come-first-serve preempt resume (LCFS-PR)
  - Round robin (RR) with finite quantum size
  - Processor sharing (PS) — RR with infinitesimal quantum size
  - Infinite server (IS)

Queuing Discipline Specification

Queuing follows Kendall's notation: Six queue attributes

- A: inter-arrival time distribution
- S: service time distribution
- m: number of servers
- B: number of buffers (system capacity)
- K: population size
- SD: service discipline

Inter-arrival and service time specifiers

- M exponential
- E_k Erlang with parameter k
- H_k hyperexponential with parameter k
- D deterministic
- G general (any distribution)

Omitted specifiers assume certain defaults:

- infinite buffer capacity
- infinite population size
- FCFS service discipline

Queuing Discipline Specification: Example

M/D/5/40/200/FCFS:

- Exponentially distributed interarrival times
- Deterministic service times
- Five servers
- Forty buffers (35 for waiting)
- Total population of 200 customers
- First-come-first-serve service discipline

M/M/1:

- Exponentially distributed interarrival times
- Exponentially distributed service times
- One server
- Infinite number of buffers
- Infinite population size
- First-come-first-serve service discipline

Example: typical bank

1. 5 tellers
2. customers form a single line and are serviced FCFS
3. excluding a run on the bank, waiting room is infinite
4. the population is infinite
5. bulk arrivals are possible if many people arrive together

- Service time and inter-arrival time distributions?
  - measure them with a watch at the bank
  - Or, make mathematically simplifying assumptions
  - Latter is most common and exponential distribution is typical
- Combining these facts and assumptions
  - M/M/1 queue
  - As we shall see, the mean queue length (including one in service) for an M/M/1 queue is

$$\frac{\lambda}{\mu - \lambda}$$

  - λ = mean inter-arrival time, and μ= mean service time

Notation and Basic "Facts"

- τ is job interarrival time
- $\lambda = 1/E[\tau]$ mean job arrival rate
- s is service time per customer (job)
- m is number of servers
- $\mu = 1/E[s]$ is mean service rate per server
- n_q is number of jobs waiting to receive service
- n_s is number of jobs in service
- $n = n_q + n_s$ is number of jobs in the system
- r is response time (service time plus queueing delay)
- w is waiting time (queueing delay only)

System must be "stable" to have an steady state solution:

- Number of jobs in the system is finite
- Requires the relation $\lambda < m\mu$ hold unless
    - the population is finite (queue length is bounded)
    - the buffer capacity is finite (arrivals are lost when queue is full)
    - (in these cases, system is always stable)

M / M /1 Queue Analysis

- $M/M/1$ is special case of a birth-death process
- $\lambda_i = \lambda_j$ for all i, j
- $\mu_i = \mu_j$ for all i, j

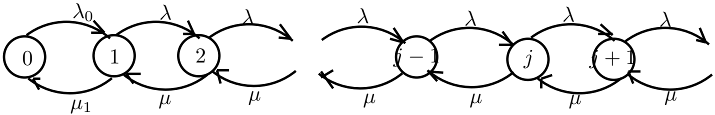

Figure 2: $M/M/1$ Queue.

- proba. of in state n, $p_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} p_0$
- p_0 is prob. of being in state 0, $p_1 = \frac{\lambda_0}{\mu_1} p_0$

M / M /1 Queue Analysis

- $p_n = (\frac{\lambda}{\mu})^n p_0,\ n = 1, 2, ..., \infty$
- $\rho = \frac{\lambda}{\mu}$, is called "traffic intensity"
- Mean queue length $E[n]$ or $\bar{n}$ is

$$\begin{aligned}\bar{n} &= \sum_{n=1}^{\infty} n p_n \\ &= \sum_{n=1}^{\infty} n(1-\rho)\rho^n \\ &= \frac{\rho}{1-\rho}\end{aligned}$$

- probability of n or more jobs in system: ρ^n

M / M /1 Queue Example

- Let there is a queue with $\mu = 0.5, \lambda = 0.3$
- then, we can calculate: utilization $U = \rho = \frac{\lambda}{\mu} = \frac{0.3}{0.5} = 0.6$
- mean number of jobs in the system $(\bar{n}) = \frac{\rho}{1-\rho} = 1.5$
- mean response time $\bar{r} = \frac{1}{\mu-\lambda} = \frac{1}{0.2} = 5.0$

M / M /m Queue

- *m* servers
- model for multiple tellers in a bank
- shared memory multiprocessors
- packet routing in Internet
- search engines to respond query
- packet & message communication in wireless mobile net
- many similar application.

Assumptions: All servers have same service rate μ, single queue for access to all servers, arrival rate λ