Queuing (Waiting Line) Theory

Queue or Waiting Line: An Example at a check-out counter

Queueing Examples In Real Life

Why do we need to study queueing and queueing theory?

- From a customer's prospective:
  - Line is too long
  - Perceived time to be served is too long
  - Someone cut line in front of you
- From a business prospective:
  - Multiple tasking, Productivity, Fairness
  - Put queueing theory in use
  - Combine intuition, common sense & queueing theory

A 30 minutes observation at a 911 call center (starts at 8:00am at time 0):

id	at Time	Server busy min	Server idle min	no in Queue
	0		2	
1	2	3		0
1	5		3	
2	8	10		0
3	11			1
4	12			2
5	17			3
2	18			
3	18	8		2
3	26			
4	26	2		1
4	28			
5	28	1		0
	29			
	30		1	
6	30	5		0

Characteristics of Queuing

- Arrival process: (λ) mean arrival rate per time unit (hour) versus mean inter-arrival time ($1/\lambda$). If 160 customers arrive for service at a bar in an eight hour day,
  - λ Arrival rate per hour?

- $\frac{1}{\lambda}$ inter-arrival time (hour)?

Please note: the mean arrival rate λ and the inter-arrival time 1/λ should initially have the same time units, an hour, for example.

- λ Arrival rate per 15 minutes?

- $\frac{1}{\lambda}$ inter-arrival time?

- Service process: (μ) mean service rate per time unit (hour) versus mean service time (1/μ). If the bar can serve 240 customers in an eight hour day,
    - μ Service rate per hour?

    - $\frac{1}{\mu}$ Mean service time (hour)?

Please note: the mean service rate μ and the mean service time 1/μ should initially have the same time units, an hour, for example.

- The arrival rate λ and the service rate μ:

- $\lambda < \mu$ in steady state
- $\lambda \approx \mu$ or λ is close to μ, what would happen if the server has to wait for an arrival to come? A server can't get any lost time back.

- Number of servers: single versus multiple

- Number of queue positions: infinite versus finite

- Source population: infinite versus finite (machine repairs)

- Queueing rules (disciplines / job sequencing):
    - FIFS or First In and First Serve: Bank teller
    - LIFS or Last In and First Serve : Elevator
    - Shortest Processing Time: Express lane at Grocery store
    - Earliest Due Date: Order processing to reduce customer complaints
    - Balking, reneging, jockeying queue positions: Special permit / ticket

What are Operating Characteristics of Queuing Theory?

λ = mean arrival rate (mean number of arrivals per time unit)

$1/\lambda$ = mean inter-arrival time for arrivals

μ = mean service rate (mean number of services per time unit)

$1/\mu$ = mean service time per customer or job

L_q = average queue length or number of units in line waiting for service

W_q = average waiting time a unit spent in queue before being served

$$L_q = \lambda W_q$$

- ✓ The average queue length is the arrival rate multiplies by the average time spent waiting in the queue.
- ✓ Jobs blocked and refused entry to the system are not counted in λ.

L = average number of units in the system (Lq in queue plus being served)

W = average time a unit spent in the system (in queue plus being served)

$$L = \lambda W$$

- ✓ The average queue length plus the one being served is the arrival rate multiplies by the average time spent waiting in the queue plus the time being served.
- ✓ Jobs blocked and refused entry to the system are not counted in λ.

s = number of parallel or equivalent servers in the system

ρ (Rho) or U = server utilization factor = the proportion of time the server is busy

P_w = Probability of an arriving unit to wait in the queue before being served

P_0 = Probability of no unit in the system (empty) (neither in queue nor being served)

P_n = Probability of having n units in the system (in queue plus being served)

Operating Characteristics of Basic Single-Server M/M/1 Queueing Model with FCFS, Infinite queue and source:

λ $\frac{1}{\lambda}$ μ $\frac{1}{\mu}$ $\rho = U = P_w = \frac{\lambda}{\mu} = P(n \geq 1) = 1 - P_0$

$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$ $P_n = P_0 \left[\frac{\lambda}{\mu}\right]^n = P_0 \rho^n$ $W = \frac{1}{\mu - \lambda} = W_q + \frac{1}{\mu} = \frac{L}{\lambda} = \frac{1}{\mu(1-\rho)}$

$L = \lambda W = \frac{\lambda}{\mu - \lambda} = L_q + \frac{\lambda}{\mu} = \frac{\rho}{1-\rho}$ $W_q = \frac{\lambda}{\mu(\mu-\lambda)} = W - \frac{1}{\mu} = \frac{L_q}{\lambda} = \frac{\rho}{\mu(1-\rho)}$ $L_q = \lambda W_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho}$

Operating Characteristics of Basic Double-Server M/M/2 Queueing Model with FCFS, Infinite queue and source:

λ $\frac{1}{\lambda}$ $\mu_n = \begin{cases} n\mu, for\ n \leq 2 \\ 2\mu, for\ n > 2 \end{cases}$ $\frac{1}{2\mu}$ $\rho = U = \frac{\lambda}{2\mu}$ $P_0 = \frac{2\mu - \lambda}{2\mu + \lambda} = \frac{1-\rho}{1+\rho}$

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0, for\ n \leq 2 \\ \frac{(\lambda/\mu)^n}{2^{(n-1)}} P_0, for\ n > 2 \end{cases}$$

$L_q = \lambda W_q = 2\rho P_0 \left(\frac{\rho}{1-\rho}\right)^2 = \frac{P_0 (\lambda/\mu)^3}{(2-\lambda/\mu)^2}$ $W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda}$ $L = \lambda W = L_q + \frac{\lambda}{\mu} = L_q + 2\rho$

$W = \frac{L}{\lambda} = W_q + \frac{1}{\mu}$ $P_w = 2\rho^2 \left(\frac{1}{1-\rho}\right) P_0 = \left(\frac{\lambda^2}{\mu(2\mu-\lambda)}\right) P_0$

Operating Characteristics of Basic Multiple-Server M/M/s Queueing Model with FCFS, Infinite queue and source:

λ $\frac{1}{\lambda}$ $\mu_n = \begin{cases} n\mu, for\ n \leq s \\ s\mu, for\ n > s \end{cases}$ $\frac{1}{s\mu}$ $\rho = U = \frac{\lambda}{s\mu}$

$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{s\mu}{s\mu - \lambda}\right)\right]^{-1}$ $P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0, for\ n \leq s \\ \frac{(\lambda/\mu)^n}{s! s^{(n-s)}} P_0, for\ n > s \end{cases}$ $L_q = \lambda W_q = \frac{(\lambda/\mu)^{s+1}}{(s-1)!(s-\lambda/\mu)^2} P_0$

$W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda}$ $L = \lambda W = L_q + \frac{\lambda}{\mu}$ $W = \frac{L}{\lambda} = W_q + \frac{1}{\mu}$ $P_w = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{s\mu}{s\mu - \lambda}\right) P_0$

Operating Characteristics of Basic Single-Server M/G/1 Queueing Model with FCFS, Infinite queue and source:

λ $\frac{1}{\lambda}$ μ $\frac{1}{\mu}$ $\rho = U = P_w = \frac{\lambda}{\mu}$ $P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$ $P_n = P_0 \left[\frac{\lambda}{\mu}\right]^n = P_0 \rho^n$

$W = W_q + \frac{1}{\mu} = \frac{L}{\lambda}$ $L = \lambda W = L_q + \frac{\lambda}{\mu}$ $W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda}$ $L_q = \lambda W_q = \frac{\lambda^2 \sigma^2 + (\lambda/\mu)^2}{2(1-\lambda/\mu)}$