Queuing Formulas *

Contents

*Notes prepared by A. Gosavi, Department of Engineering Management and Systems Engineering, Missouri S & T.

1 Notation

- λ: mean rate of arrival. It is equal to $1/E$[Inter-arrival-Time] where $E[.]$ denotes the expectation operator.
- μ: mean service rate. It is equal to $1/E$[Service-Time].
- $\rho = \lambda/\mu$ for single server queues: utilization of the server; also the probability that the server is busy or the probability that someone is being served.
- c: number of servers.
- P_n: probability that there are n customers in the system.
- L: Mean number of customers in the system.
- L_q: mean number of customers in the queue.
- W: mean wait in the system.
- W_q: mean wait in the queue.
- C^2: Coefficient of variation of a random variable; $C^2 = \frac{Variance}{(Mean)^2}$.
- C_s^2: Coefficient of variation of service time.
- C_a^2: Coefficient of variation of inter-arrival time.
- σ_s^2: variance of service time.
- K: capacity of queue.

2 Basic Queueing Formulas

Little's rule provide the following results:

$$L = \lambda W; L_q = \lambda W_q.$$

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

3 Single-Server Queues

The following notation is used for representing queues: A/B/c/K where A denotes the distribution of the inter-arrival time, B that of the service time,c denotes the number of servers, and K denotes the capacity of the queue. If K is omitted, we assume that $K = \infty$.

M stands for Markov and is commonly used for the exponential distribution. Hence an $M/M/1$ queue is one in which there is one server (and one channel) and both the inter-arrival time and service time are exponentially distributed. An $M/G/1$ queue is one with one server in which the inter-arrival time is exponentially distributed and the service time is generally distributed, i.e., the service time has any given distribution. A $G/G/1$ queue is one with one server in which both service and the inter-arrival time have any given distribution.

3.1 Formulas

For the M/M/1 queue, we can prove that

$$L_q = \frac{\rho^2}{1-\rho}.$$

For the M/G/1 queue, we can prove that

$$L_q = \frac{\lambda^2\sigma_s^2 + \rho^2}{2(1-\rho)}$$

The above is called the Pollazcek-Khintichine formula (named after its inventors and discovered in the 1930s).

For the G/G/1 queue, we do not have an exact result. The following *approximation* (derived by Marchal in 1976) is popular:

$$L_q \approx \frac{\rho^2(1+C_s^2)(C_a^2+\rho^2 C_s^2)}{2(1-\rho)(1+\rho^2 C_s^2)}.$$

Notice that if the mean rate of arrival is λ, and σ_a^2 denotes the variance of the inter-arrival time, then:

$$C_a^2 = \frac{\sigma_a^2}{(1/\lambda)^2}.$$

Similarly, if μ denotes the service rate and σ_s^2 denotes the variance of the service time, then:

$$C_s^2 = \frac{\sigma_s^2}{(1/\mu)^2}.$$

3.2 Some additional useful facts

- If the random variable X is uniformly distributed with parameters (a, b), where a is the minimum value and b the maximum value, then the mean of X is $(a+b)/2$ and the variance is $\frac{(b-a)^2}{12}$.
- For the exponential distribution if the mean if $1/\lambda$, the variance is $1/\lambda^2$.
- When a variable is deterministic, e.g., inter-arrival time is fixed, its variance is zero and hence so is its coefficient of variation.
- Consider two random variables, X and Y. Then if $E[.]$ denotes the mean and $V[.]$ denotes the variance, then

  $$E[X+Y] = E[X] + E[Y];$$

  also, if X and Y are independent,

  $$V[X+Y] = V[X] + V[Y].$$

3.3 Examples

Example 1: Consider the following scenario: the inter-arrival time is exponentially distributed with a mean of 10 minutes and the service time is also exponentially distributed with a mean of 8 means, find the (i) mean wait in the queue, (ii) mean number in the queue, (iii) the mean wait in the system, (iv) mean number in the system and (v) proportion of time the server is idle.

We have an M/M/1 system. We also have: $\lambda = 1/10$; $\mu = 1/8$. Hence, $\rho = 8/10$. Then:

$$\text{Number in the Queue } = L_q = \frac{\rho^2}{1-\rho} = \frac{0.8^2}{1-0.8} = 3.2.$$

$$\text{Wait in the Queue } = W_q = L_q/\lambda = 32\ mins.$$

$$\text{Wait in the System } = W = W_q + 1/\mu = 40\ mins.$$

$$\text{Number in the System } = L = \lambda W = 4.$$

$$\text{Proportion of time the server is idle } = 1 - \rho = 0.2.$$

Example 2: Consider the following scenario: the inter-arrival time is exponentially distributed with a mean of 10 minutes and the service time has the uniform distribution with a maximum of 9 minutes and a minimum of 7 minutes, find the (i) mean wait in the queue, (ii) mean number in the queue, (iii) the mean wait in the system, (iv) mean number in the system and (v) proportion of time the server is idle.

We have an M/G/1 system. We also have: $\lambda = 1/10$; the mean service time will be $(7+9)/2 = 8$, i.e., $\mu = 1/8$. The variance of the service time, σ_s^2 will equal $(9-7)^2/12 = 1/3$. Also, $\rho = 8/10$. Then:

$$\text{Number in the Queue } = L_q = \frac{\lambda^2 \sigma_s^2 + \rho^2}{2(1-\rho)} = 1.608.$$

$$\text{Wait in the Queue } = W_q = L_q/\lambda = 16.08 \; mins.$$

$$\text{Wait in the System } = W = W_q + 1/\mu = 24.08 \; mins.$$

$$\text{Number in the System } = L = \lambda W = 2.408.$$

$$\text{Proportion of time server is idle } = 1 - \rho = 0.2.$$

Example 3: Consider the following scenario: the inter-arrival time has a distribution that is unknown with a mean of 10 minutes and a variance of 2 min^2. The service time has the uniform distribution with a maximum of 9 minutes and a minimum of 7 minutes, find the (i) mean wait in the queue, (ii) mean number in the queue, (iii) the mean wait in the system, (iv) mean number in the system and (v) proportion of time the server is idle.

We have a G/G/1 system. We also have: $\lambda = 1/10$; the variance of the inter-arrival time is 2. The mean service time will be $(7+9)/2 = 8$, i.e., $\mu = 1/8$. The variance of the service time, σ_s^2 will equal $(9-7)^2/12 = 1/3$. Also, $\rho = 8/10$. Then:

$$C_a^2 = \frac{\sigma_a^2}{(1/\lambda)^2} = 0.02; C_s^2 = \frac{\sigma_s^2}{(1/\mu)^2} = 1/192.$$

$$\text{Number in the Queue } = L_q = \frac{\rho^2(1+C_s^2)(C_a^2 + \rho^2 C_s^2)}{2(1-\rho)(1+\rho^2 C_s^2)} = 0.037.$$

$$\text{Wait in the Queue } = W_q = L_q/\lambda = 0.373 \; mins.$$

$$\text{Wait in the System } = W = W_q + 1/\mu = 8.373 \; mins.$$

$$\text{Number in the System } = L = \lambda W = 0.8373.$$

$$\text{Proportion of time server is idle } = 1 - \rho = 0.2.$$

4 Multiple-Server Queues

We will only consider the identical server case. Let c denote the number of identical servers. Here

$$\rho = \frac{\lambda}{c\mu}$$

For the M/M/c queue,

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

where

$$P_0 = 1/\left[\sum_{m=0}^{c-1} \frac{(c\rho)^m}{m!} + \frac{(c\rho)^c}{c!(1-\rho)}\right].$$

Note that P_0 denotes the probability that there are 0 customers in the system.

For the G/G/c queue, we have an *approximation* (Kingman, 1964)

$$W_q^{G/G/c} \approx W_q^{M/M/c} \frac{C_a^2 + C_s^2}{2},$$

where $W_q^{A/B/c}$ denotes the waiting time in the queue for the A/B/c queue.