Tutorial for Use of Basic Queueing Formulas *

Contents

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

Please use as is, and I can't help with your homework unless you're taking a course with me :-)!!

1 Notation

We assume that we have a *single-channel* queue, i.e., there is only one waiting line. See Figure 1.

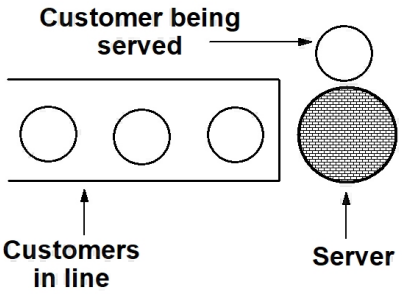

Figure 1: A single-channel, single-server queue, which has three customers waiting in the queue (line) and one being served at the instant this photo is shot

- λ: mean rate of arrival and equals $1/E$[Inter-arrival-Time], where $E[.]$ denotes the expectation operator.
- μ: mean service rate and equals $1/E$[Service-Time]
- c: number of servers in parallel
- $\rho = \lambda/(c\mu)$: utilization of the server; also the probability that the server is busy or the proportion of time the server is busy
- 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 waiting time in the system
- W_q: mean waiting time in the queue
- C^2: squared coefficient of variation of a random variable; $C^2 = \frac{Variance}{(Mean)^2}$
- C_s^2: squared coefficient of variation of service time
- C_a^2: squared coefficient of variation of inter-arrival time
- σ_s^2: variance of service time

2 Two Moment Approximations

This tutorial is written to explain the basics of two-moment approximations that are very popular in industry for obtaining queueing estimates, i.e., the mean waiting time in a queue and the mean length of a queue. These approximations can usually only provide means of outputs, i.e, waiting times and queue lengths, based on three inputs in a standard queue: (i) the mean and variance of the inter-arrival time, (ii) the mean and variance of the service time, and (iii) the number of servers. This situation arises frequently in factories, airports, and hospitals, where limited data, i.e., only means and variances of the inputs, are available.

Note that the mean is the so-called first moment. Thus, if a random variable is denoted by X, the first moment, $E[X]$, is the mean, while the variance is $E[X^2] - (E[X])^2$, where $E[X^2]$ is the so-called second moment. Thus, the variance is **not** the second moment, but rather the second moment minus the square of the mean. While the approximations studied in this tutorial are technically called *two-moment approximations*, we really only need the mean and the variance, and the calculation of the second moment is not needed.

3 Basic Queueing Formulas

Little's rule provides the following results:

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

the first of the above applies to the system and the second to the queue, which is a part of the system. Another useful relationship in the queue is:

$$W = W_q + \frac{1}{\mu}; \tag{1}$$

the above is intuitive (we prove it later): it says the mean wait in the *system* is the sum of the mean wait in the *queue* and the service time $(1/\mu)$.

4 Queueing Notation

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.

5 Single-Server Queues

We first consider single-server queues first where $c = 1$. They arise in many manufacturing and service systems.

5.1 Formulas

For the M/M/1 queue, we can prove that (Ross, 2014)

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

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; see Ross (2014)).

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

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

In the above, 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:

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

Another approximation from Kraemer and Langenbach-Belz (1976) is also quite powerful:

$$L_q \approx \frac{\rho^2(C_a^2 + C_s^2)}{2(1-\rho)} g \tag{3}$$

where

$$g = \exp\left(\frac{-2(1-\rho)(1-C_a^2)^2}{3\rho(C_a^2 + C_s^2)}\right) \text{ when } C_a^2 \leq 1; \tag{4}$$

$$g = \exp\left(\frac{(1-\rho)(1-C_a^2)}{C_a^2 + 4C_s^2}\right) \text{ when } C_a^2 > 1. \tag{5}$$

5.2 Useful Facts

- If $\rho \geq 1$ in a queue where either the inter-arrival or service time or both are random, the queue becomes unstable, i.e., the length of the queue and the wait become infinity. If both are constants, $\rho > 1$ implies instability. Such queues need additional servers for stability.
- 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 $\frac{(b-a)^2}{12}$.
- 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

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

  thus if X is the wait in the queue and Y is the service time, we have $W = W_q + E[Y] = W_q + \frac{1}{\mu}$, which was Equation (1).

5.3 Examples

Example 1: Consider the following single-server queue: 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 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.

Solution: 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 single-server queue: 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.

Solution: 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:

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

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

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

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

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

Example 3: Consider the following single-server queue: the inter-arrival time has a gamma distribution with a mean of 10 minutes and a variance of 20 min^2. The service time has the normal distribution with a mean of 8 minutes and a variance of 25 min^2 , 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. Simulation results indicate W_q to be about 8.1 minutes.

We have a G/G/1 system. We also have: $\lambda = 1/10$; the variance of the inter-arrival time is 20. The mean service time will be 8, i.e., $\mu = 1/8$. The variance of the service time, σ_s^2 is

25. Also, $\rho = 8/10$. Then,

$$C_a^2 = \frac{\sigma_a^2}{(1/\lambda)^2} = 0.2; C_s^2 = \frac{\sigma_s^2}{(1/\mu)^2} = 0.3906.$$

Now using Marchal's approximation:

$$\text{Number in the Queue via Equation (2)} \ = L_q = \frac{\rho^2(1+C_s^2)(C_a^2+\rho^2 C_s^2)}{2(1-\rho)(1+\rho^2 C_s^2)} = 0.8010.$$

$$\text{Wait in the queue} \ = W_q = L_q/\lambda = 8.01 \, mins \approx 8.1 \, mins, \ \text{which is the simulation estimate.}$$

$$\text{Wait in the system} \ = W = W_q + 1/\mu = 16.01 \, mins.$$

$$\text{Number in the system} \ = L = \lambda W = 1.601.$$

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

Using the Kramer-Langenbach-Belz approximation in Equation (3), we have:

$$L_q \approx \frac{\rho^2(C_a^2+C_s^2)}{2(1-\rho)} g = 0.9450g$$

where since $C_a^2 < 1$, via Equation (4),

$$g = \exp\left(\frac{-2(1-\rho)(1-C_a^2)^2}{3\rho(C_a^2+C_s^2)}\right) = 0.8348.$$

Then, $L_q = 0.9450(0.8348) = 0.7889$, which implies $W_q = L_q/\lambda = 0.7889(10) = 7.889$ mins, which is also reasonably close to the simulation estimate of 8.1 mins.

6 Multiple-Server Queues

We will only consider the identical (homogenous) server case in which there are c identical servers in parallel and there is just one waiting line (i.e., the queue is a single-channel queue). Let c denote the number of identical servers. Here

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

For the M/M/c queue (Ross, 2014),

$$L_q = \frac{P_0(\frac{\lambda}{\mu})^c \rho}{c!(1-\rho)^2} \tag{6}$$

where

$$P_0 = 1/\left[\sum_{m=0}^{c-1} \frac{(c\rho)^m}{m!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]. \tag{7}$$

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

Hence, W_q can be obtained as follows:

$$W_q = L_q/\lambda.$$

Then, for the $G/G/c$ queue, we have the following *approximation* (Whitt, 1976; Medhi, 2003):

$$W_q^{G/G/c} \approx W_q^{M/M/c} \frac{C_a^2 + C_s^2}{2}, \tag{8}$$

where $W_q^{A/B/c}$ denotes the waiting time in the queue for the A/B/c queue. The above works well for M/G/c queues, but does not always work well when the inter-arrival time is not exponentially distributed. For multi-server queues, it has been shown that data on two moments is usually not sufficient to generate good approximations for the mean waiting time or queue length (Gupta et al , 2010). When the distributions are known, it is often possible to deduce expressions for these metrics, but they often involve calculus and computational methods (see Kahraman and Gosavi (2011) for one such situation in bulk queues and Medhi (2003) for general discussions, including the Lindley equation).

Example 4: Consider the following scenario: the inter-arrival time has an exponential distribution with a mean of 10 minutes. There are two servers, and the service time of each server has the uniform distribution with a maximum of 20 minutes and a minimum of 10 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. Results from discrete-event simulation, which are known to be very accurate, show that the mean waiting time in the queue is 9.5693 minutes. Compute the error in the G/G/c approximation.

Solution: This is an $M/G/2$ system. We have $\lambda = 1/10$; the $C_a^2 = 1$ as a result. The mean service time will be $(10 + 20)/2 = 15$, i.e., $\mu = 1/15$. The variance of the service time, σ_s^2 will equal $(20 - 10)^2/12 = 8.33$. Also, $\rho = 15/(2 \times 10) = 0.75$. Then:

$$C_s^2 = \frac{\sigma_s^2}{(1/\mu)^2} = 8.33/(15)^2 = 0.03.$$

Using the $G/G/c$ approximation, we first assume the queue to be an $M/M/c$ queue and compute its L_q: Now using the formula above in Eqn. (7): $P_0 = 0.1453$. Then, using Eqn. (6), we have that $L_q = \frac{0.1453(\frac{1/10}{1/15})^2 0.75}{2!(1-0.75)^2} = 1.929$. Then, $W_q = L_q/\lambda = 1.929 \times 10 = 19.29$.

Now, we need to transform this to an $G/G/2$ queue using the approximation in Eqn. (8):

$$W_q^{G/G/c} \approx W_q^{M/M/c} \frac{C_a^2 + C_s^2}{2} = (19.29)(1 + 0.03)/2 = 9.93.$$

Then, $L_q^{G/G/c} = W_q^{G/G/c} \times \lambda = 9.93 \times 1/10 = 0.993$. The error in the approximation is:

$$\frac{|9.9376 - 9.5693|}{9.9376} \times 100\% = 3.07\%.$$

$$\text{Wait in the System } = W = W_q + 1/\mu = 9.93 + 15 = 24.93 \; mins.$$

$$\text{Number in the System } = L = \lambda W = 2.493.$$

References

V. Gupta, M. Harchol-Balter, J.G. Dai and B. Zwart. On the inapproximability of $M/G/K$: why two moments of job size distribution are not enough. *Queueing Systems*, 64(1): 5–48, 2010.

A. Kahraman and A. Gosavi. On the distribution of the number stranded in bulk-arrival, bulk-service queues of the M/G/1 form. *European Journal of Operational Research*, 212(2):352–360, 2011.

W. Kraemer and M. Langenbach-Belz. Approximate formulae for the delay in the queueing system $GI/G/1$. In *Proceedings of the 8th International Telegraphic Congress*, volume 2(3), page 235/1 235/8, Melbourne, Australia, 1976.

W.G. Marchal. An approximation formula for waiting times in single-server queues. *AIIE Transactions*, 8: 473, 1976.

J. Medhi. *Stochastic Models in Queueing Theory*. Academic Press, Amsterdam, Second edition, 2003.

S. M. Ross. *Introduction to Probability Models*. Academic Press, San Diego, CA, USA, Eleventh edition, 2014.

W. Whitt. The queueing network analyzer. *Bell System Technical Journal*, 62(9):2779-2815, 1983.

Exercises:

1. Consider a single-server queue with a gamma distributed inter-arrival time which has a mean of 10 minutes and variance of 20 minutes-squared. The service time is uniformly distributed between 6 and 7 minutes. From simulations, W_q has been estimated to be 0.612 minutes. Compute W_q using (a) Marchal's approximation and (b) the Kraemer-Langenbach-Belz approximation. Compute the error in each estimate from simulation.
2. Consider a multi-server queue with 5 servers in parallel. Each server has a normally distributed service time with a mean of 16.67 minutes and a standard deviation of 1 minute. The inter-arrival time is exponentially distributed with a mean of 5 minutes. The simulator provides an estimate of 1.99 minutes for W_q. Use the queueing approximation discussed above to generate a value for W_q and compare it to the simulation estimate.