QUEUING THEORY

RYAN BERRY

Abstract. This paper defines the building blocks of and derives basic queuing systems. It begins with a review of some probability theory and then defines processes used to analyze queuing systems, in particular the birth-death process. A few simple queues are analyzed in terms of steady-state derivation before the paper discusses some attempted field research on the topic.

1. Introduction

Queuing theory is a branch of mathematics that studies and models the act of waiting in lines. This paper will take a brief look into the formulation of queuing theory along with examples of the models and applications of their use. The goal of the paper is to provide the reader with enough background in order to properly model a basic queuing system into one of the categories we will look at, when possible. Also, the reader should begin to understand the basic ideas of how to determine useful information such as average waiting times from a particular queuing system.

The first paper on queuing theory, "The Theory of Probabilities and Telephone Conversations" was published in 1909 by A.K. Erlang, now considered the father of the field. His work with the Copenhagen Telephone Company is what prompted his initial foray into the field. He pondered the problem of determining how many telephone circuits were necessary to provide phone service that would prevent customers from waiting too long for an available circuit. In developing a solution to this problem, he began to realize that the problem of minimizing waiting time was applicable to many fields, and began developing the theory further.

Erlang's switchboard problem laid the path for modern queuing theory. The chapters on queuing theory and its applications in the book "Operations Research: Applications and Algorithms" by Wayne L. Winston illustrates many expansions of queuing theory and is the book from which the majority of the research of this paper has been done.

In the second section of this paper, we will begin defining the basic queuing model. We will begin by reviewing the necessary probabilistic background needed to understand the theory. The we will move on to discussing notation, queuing disciplines, birth-death processes, steady-state probabilities, and Little's queuing formula. In the next section we will begin looking at particular queuing models. We will study the population size, the customer capacity, the number of servers, self-service queues, and the machine repair model, to name a few. We will be calculating steady-state probabilities and waiting times for the models when possible, while also looking at examples and applications. We will conclude the paper by taking a peek at some field research studying the queuing system at a bank.

2. The Basic Queuing Model

To begin understanding queues, we must first have some knowledge of probability theory. In particular, we will review the exponential and Poisson probability distributions.

2.1. **Exponential and Poisson Probability Distributions.**

The exponential distribution with parameter λ is given by $\lambda e^{-\lambda t}$ for $t \geq 0$. If T is a random variable that represents interarrival times with the exponential distribution, then $P(T \leq t) = 1 - e^{-\lambda t}$ and $P(T > t) = e^{-\lambda t}$.

This distribution lends itself well to modeling customer interarrival times or service times for a number of reasons. The first is the fact that the exponential function is a strictly decreasing function of t. This means that after an arrival has occurred, the amount of waiting time until the next arrival is more likely to be small than large. Another important property of the exponential distribution is what is known as the no-memory property. The no-memory property suggests that the time until the next arrival will never depend on how much time has already passed. This makes intuitive sense for a model where we're measuring customer arrivals because the customers' actions are clearly independent of one another.

It's also useful to note the exponential distribution's relation to the Poisson distribution. The Poisson distribution is used to determine the probability of a certain number of arrivals occurring in a given time period. The Poisson distribution with parameter λ is given by

$$\frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

where n is the number of arrivals. We find that if we set $n = 0$, the Poisson distribution gives us

$$e^{-\lambda t}$$

which is equal to $P(T > t)$ from the exponential distribution.

The relation here also makes sense. After all, we should be able to relate the probability that zero arrivals will occur in a given period of time with the probability that an interarrival time will be of a certain length. The interarrival time here, of course, is the time between customer arrivals, and thus is a period of time with zero arrivals.

With these distributions in mind, we can begin defining the input and output processes of a basic queuing system, from which we can start developing the model further.

2.2. **The Input Process.** To begin modeling the input process, we define t_i as the time when the ith customer arrives. For all $i \geq 1$, we define $T_i = t_{i+1} - t_i$ to be the ith interarrival time. We also assume that all T_i's are independent, continuous random variables, which we represent by the random variable A with probability density $a(t)$. Typically, A is chosen to have an exponential probability distribution with parameter λ defined as the arrival rate, that is to say, $a(t) = \lambda e^{-\lambda t}$.

It is easy to show [W 1045] that if A has an exponential distribution, then for all nonnegative values of t and h,

$$P(A > t + h | A \geq t) = P(A > h)$$

This is an important result because it reflects the no-memory property of the exponential distribution, which is an important property to take note of if we're modeling interarrival times.

Another distribution the can be used to model interarrival times (if the exponential distribution does not seem to be appropriate) is the Erlang distribution. An Erlang distribution is a continuous random variable whose density function relies on a rate parameter R and a shape parameter k. The Erlang probability density function is

$$f(t) = \frac{R(Rt)^{k-1}e^{-Rt}}{(k-1)!}$$

2.3. **The Output Process.** Much like the input process, we start analysis of the output process by assuming that service times of different customers are independent random variables represented by the random variable S with probability density $s(t) = \mu e^{-\mu t}$. We also define μ as the service rate, with units of customers per hour. Ideally, the output process can also be modeled as an exponential random variable, as it makes calculation much simpler. Imagine an example where four customers are at a bank with three tellers with exponentially distributed service times. Three of them receive service immediately, while the fourth has to wait for one position to clear. What is the probability that the fourth customer will be the final one to complete service?

Due to the no-memory property of the exponential distribution, when the fourth customer finally steps up to a teller, all three remaining customers have an equal chance of finishing their service last, as the service time in this situation is not governed by how long they have already been served. Thus, the answer to the question is 1/3.

Unfortunately, the exponential distribution does not always represent service times accurately. For a service that requires many different phases of service (for example, scanning groceries, paying for groceries, and bagging the groceries), an Erlang distribution can be used with the parameter k equal to the number of different phases of service.

2.4. **Birth-Death Processes.** We define the number of people located in a queuing system, either waiting in line or in service, to be the state of the system at time t. At $t = 0$, the state of the system is going to be equal to the number of people initially in the system. The initial state of the system is noteworthy because it clearly affects the state at some future t. Knowing this, we can define $P_{ij}(t)$ as the probability that the state at time t will be j, given that the state at $t = 0$ was i. For a large t, $P_{ij}(t)$ will actually become independent of i and approach a limit π_j. [W 1053] This limit is known as the steady-state of state j. Generally, if one is looking at the steady-state probability of j, it is incredibly difficult to determine

FIGURE 1. In a single-server birth-death process, births add one to the current state and occur at rate λ. Deaths subtract one from the current state and occur at rate μ.

the steps of arrivals and services that led up to the steady state. Likewise, starting from a small t, it is also very difficult to determine when exactly a system will reach its steady state, if it exists. Thus, for simplicity's sake, when we study a queuing system, we begin by assuming that the steady-state has already been reached.

A birth-death process is a process wherein the system's state at any t is a nonnegative integer. The variable λ_j is known as the birth rate at state j and symbolizes the probability of an arrival occurring over a period of time. The variable μ_j is known as the death rate at state j and symbolizes the probability that a completion of service occurs over a period of time. Thus, births and deaths are synonymous with arrivals and service completions respectively. A birth increases the state by one while a death decreases the state by one. We note that $\mu_0 = 0$, since it must not be possible to enter a negative state. Also, in order to officially be considered a birth-death process, birth and deaths must be independent of each other. A simple birth-death process is illustrated in Figure 1.

The probability that a birth will occur between t and $t + \Delta t$ is $\lambda_j \Delta t$, and such a birth will increase the state from j to $j + 1$. The probability that a death will occur between t and $t + \Delta t$ is $\mu_j \Delta t$, and such a birth will decrease the state from j to $j - 1$.

2.5. **Steady-state Probabilities.** In order to determine the steady-state probability π_j, we have to find a relation between $P_{ij}(t + \Delta t)$ and $P_{ij}(t)$ for a reasonably sized t. We begin by categorizing the potential states at time t from which a system could end up at state j at time $t + \Delta t$. In order to achieve this, the state at time t must be j, $j - 1$, $j + 1$, or some other value. Then, to calculate π_j, all we have to do is add up the probabilities of the system ending at state j for each of these beginning categories.

To reach state j from state $j - 1$, we need one birth to occur between t and Δt. To reach j from $j + 1$, we need one death. To remain at j, we need no births or deaths to occur. To reach j from any other state we will need multiple births or deaths. Since we will be eventually letting Δt approach zero, we find that it is impossible to reach state j from these other states because births and deaths are independent of each other, and won't occur simultaneously. Hence, we only need to sum the probabilities of these first three situations occurring. That will give us

$$P_{ij}(t+\Delta t) = [P_{i,j-1}(t)(\lambda_{j-1}\Delta t)] + [P_{i,j+1}(t)(\mu_{j+1}\Delta t)] + [P_{ij}(t)(1-\mu_j\Delta t - \lambda_j\Delta t)]$$

which can be rewritten as

$$P'_{ij}(t) = \lim_{\Delta t\to\infty} \frac{P_{ij}(t+\Delta t) - P_{ij}}{\Delta t} = \lambda_{j-1}P_{i,j-1}(t) + \mu_{j+1}P_{i,j+1}(t) - P_{ij}(t)\mu_j - P_ij(t)\lambda_j$$

Since we're trying to calculate steady-state probabilities, it is appropriate to allow t to approach infinity, at which point $P_{ij}(t)$ can be thought of as a constant. Then $P'_{ij}(t) = 0$ Defining the steady-state probability $\pi_j = \lim_{t\to\infty} P_{ij}(t)$, we can substitute further.

$$\lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1} - \pi_j\mu_j - \pi_j\lambda_j = 0$$

$$\lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1} = \pi_j(\lambda_j + \mu_j) \text{ for } j = 1, 2, ...$$

$$\mu_1\pi_1 = \lambda_0\pi_0 \text{ for } j = 0$$

These results are known as the flow balance equations. You may notice that they suggest that the rate at which transitions occur into a particular state equal the rate at which transitions occur out of the same state. At this point, each steady-state probability can be determined by substituting in probabilities from lower states, shown in greater detail on pages 1058-9 of the Winston text. Starting with $\pi_1 = \frac{\pi_0\lambda_0}{\mu_1}$, we can get the general equation

$$\pi_j = \pi_0 c_j$$

where

$$c = \frac{\lambda_0\lambda_1...\lambda_{j-1}}{\mu_1\mu_2...\mu_j}$$

2.6. **Queuing Disciplines.** It is easy for one to think of all queues operating like a grocery checkout line. That is to say, when an arrival occurs, it is added to the end of the queue and service is not performed on it until all of the arrivals that came before it are served in the order they arrived. Although this a very common method for queues to be handled, it is far from the only way. The method in which arrivals in a queue get processed is known as the queuing discipline. This particular example outlines a first-come-first-serve discipline, or an FCFS discipline. Other possible disciplines include last-come-first-served or LCFS, and service in random order, or SIRO. While the particular discipline chosen will likely greatly affect waiting times for particular customers (nobody wants to arrive early at an LCFS discipline), the discipline generally doesn't affect important outcomes of the queue itself, since arrivals are constantly receiving service regardless.

2.7. **Kendall-Lee Notation.** Since describing all of the characteristics of a queue inevitably becomes very wordy, a much simpler notation (known as Kendall-Lee notation) can be used to describe a system. Kendall-Lee notation gives us six abbreviations for characteristics listed in order separated by slashes. The first and second characteristics describe the arrival and service processes based on their respective probability distributions. For the first and second characteristics, M represents an exponential distribution, E represents an Erlang distribution, and G represents a general distribution. The third characteristic gives the number of servers working together at the same time, also known as the number of parallel servers. The fourth describes the queue discipline by it's given acronym. The fifth gives the maximum number of number of customers allowed in the system. The sixth gives the size of the pool of customers that the system can draw from. For example, $M/M/5/FCFS/20/\inf$ could represent a bank with 5 tellers, exponential arrival times, exponential service times, an FCFS queue discipline, a total capacity of 20 customers, and an infinite population pool to draw from.

2.8. **Little's Queuing Formula.** In many queues, it is useful to determine various waiting times and queue sizes for Particular components of the system in order to make judgments about how the system should be run. Let us define L to be the average number of customers in the queue at any given moment of time assuming that the steady-state has been reached. We can break that down into L_q, the average number of customers waiting in the queue, and L_s, the average number of customers in service. Since customers in the system can only either be in the queue or in service, it goes to show that $L = L_q + L_s$.

Likewise, we can define W as the average time a customer spends in the queuing system. W_q is the average amount of time spent in the queue itself and W_s is the average amount of time spent in service. As was the similar case before, $W = W_q + W_s$. It should be noted that all of the averages in the above definitions are the steady-state averages.

Defining λ as the arrival rate into the system, that is, the number of customers arriving the system per unit of time, it can be shown that

$$L = \lambda W$$

$$L_q = \lambda W_q$$

$$L_s = \lambda W_s$$

This is known as Little's queuing formula [W 1062].

3. Queuing Models

With our foundation laid for the study of important characteristics of queuing systems, we can begin to analyze particular systems themselves. We will begin by looking at one of the simplest systems, the M/M/1/GD/∞/∞ system.

3.1. The M/M/1/GD/∞/∞ Queuing System.

An M/M/1/GD/∞/∞ system has exponential interarrival times, exponential service times, and one server. This system can be modeled as a birth-death process where

$$\lambda_j = \lambda \text{ for } (j = 0, 1, 2...)$$

$$\mu_0 = 0$$

$$\mu_j = \mu \text{ for } (j = 1, 2, 3...)$$

Substituting this in to the equation for the steady-state probability, we get

$$\pi_j = \frac{\lambda^j \pi_0}{\mu^j}$$

We will define $p = \lambda/\mu$ as the traffic intensity of the system, which is a ratio of the arrival and service rates. Knowing that the sum of all of the steady state probabilities is equal to one, we get

$$\pi_0(1 + p + p^2 + ... + p^j) = 1$$

If we assume $0 \leq p \leq 1$ and let the sum $S = (1 + p + p^2 + ... + p^j)$, then $S = \frac{1}{1-p}$ and $\pi_0 = 1 - p$. This yields

$$\pi_j = p^j(1 - p)$$

as the steady-state probability of state j [W 1058]. Note that if $p \geq 1$, S approaches infinity, and thus no steady state can exist. Intuitively, if $p \geq 1$, then it must be that $\lambda \geq \mu$, and if the arrival rate is greater than the service rate, then the state of the system will grow without end.

With the steady-state probability for this system calculated, we can now solve for L. If L is the average number of customers present in this system, we can represent it by the formula

$$L = \sum_{j=0}^{\infty} j\pi_j = (1 - p)\sum_{j=0}^{\infty} jp^j$$

Let $S = \sum_{j=0}^{\infty} = p + 2p^2 + 3p^3 + ...$ Then $pS = p^2 + 2p^3 + 3p^4 + ...$ If we subtract, we get

$$S - pS = p + p^2 + p^3... = \frac{p}{1 - p}$$

And $S = \frac{p}{(1-p)^2}$. Substituting this into the equation for L will get us

$$L = (1 - p)\frac{p}{(1 - p)^2} = \frac{p}{1 - p} = \frac{\lambda}{\mu - \lambda}$$

To solve for L_s, we have to determine how many customers are in service at any given moment. In this particular system, there will always be one customer in service except for when there are no customers in the system. Thus, this can be calculated as

$$L_q = 0\pi_0 + 1(\pi_1 + \pi_2 + \pi_3 + ...) = 1 - \pi_0 = 1 - (1 - p) = p$$

From here, L_q is an easy calculation.

$$L_q = L - L_s = \frac{p}{1 - p} - p = \frac{p^2}{1 - p}$$

Using Little's queuing formula, we can also solve for W, W_s, and W_q by dividing each of the corresponding L values by λ.

3.2. **The M/M/1/GD/c/∞ Queuing System.** An M/M/1/GD/c/∞ queuing system has exponential interarrival and service times, with rates λ and μ respectively. This system is very similar to the previous system, except that whenever c customers are present in the system, all additional arrivals are excluded from entering, and are thereafter no longer considered. For example, if a customer were to walk up to a fast food restaurant and see that the lines were too long for him to want to wait there, he would go to another restaurant instead.

A system like this can be modeled as a birth-death process with these parameters:

$$\lambda_j = \lambda \text{ for } j = 0, 1, ..., c-1$$

$$\lambda_c = 0$$

$$\mu_0 = 0$$

$$\mu_j = \mu \text{ for } j = 1, 2, ..., c$$

The restriction $\lambda_c = 0$ is what sets this apart from the previous system. It makes it so that no state greater than c can ever be reached. Because of this restriction, a steady state will always exist. This is because even if $\lambda \geq \mu$, there will never be more than c customers in the system.

Looking at formulas derived from the study of birth-death processes and once again letting $p = \frac{\lambda}{\mu}$, we can derive the following steady-state probabilities [W 1068]:

$$\pi_0 = \frac{1-p}{1-p^{c+1}}$$

$$\pi_j = p^j \pi_0 \text{ for } j = 1, 2, ..., c$$

$$\pi_j = 0 \text{ for } j = c+1, c+2, ..., \infty$$

A formula for L can be found in a similar fashion, but is omitted because of the messy calculations. The technique is similar to the one used in the previous section.

Calculating W is another issue. This is because in Little's queuing formula, λ represents the arrival rate, but in this system, not all of the customers who arrive will join the queue. In fact, $\lambda\pi_c$ arrivals will arrive, but leave the system. Thus, only $\lambda - \lambda\pi_c = \lambda(1-\pi_c)$ arrivals will ever enter the system. Substituting this into Little's queuing formula gives us

$$W = \frac{L}{\lambda(1-\pi_c)}$$

3.3. **The M/M/s/GD/∞/∞ Queuing System.** An M/M/s/GD/∞/∞ queuing system, like the previous system we looked at, has exponential interarrival and service times, with rates λ and μ. What sets this system apart is that there are s servers willing to serve from a single line of customers, like perhaps one would find in a bank. If $j \leq s$ customers are present in the system, then every customer is being served. If $j > s$ customers are in the system, then s customers are being served and the remaining $j - s$ customers are waiting in the line.

To model this as a birth-death system, we have to observe that the death rate is dependent on how many servers are actually being used. If each server completes service with a rate of μ, then the actual death rate is μ times the number of customers actually being served. Parameters for this system are as follows:

$$\lambda_j = \lambda \text{ for } j = 0, 1, ..., \infty$$

$$\mu_j = j\mu \text{ for } j = 0, 1, ..., s$$

$$\mu_j = s\mu \text{ for } j = s+1, s+2, ..., \infty$$

In solving the steady-state probabilities, we will define $p = \frac{\lambda}{s\mu}$. Notice that this definition also applies to the other systems we looked at, since in the other two systems, $s = 1$. The steady-state probabilities can be found in this system in the same manner as for other systems by using the flow balance equations [W 1071]. I will also omit these particular steady-state equations because they are rather cumbersome.

3.4. **The M/G/∞/GD/∞/∞ and GI/G/∞/GD/∞/∞ Queuing Systems.** These systems are set apart in that they have an infinite number of servers, and thus, a customer never has to wait in a queue for their service to begin. One way to think of this is as a self-service, like shopping on the internet, for example. In this system, it can be shown that $W = \frac{1}{\mu}$ and $L = \frac{\lambda}{\mu}$ [W 1076]. It can also be shown that the steady state probability at state j is

$$\pi_j = \frac{(\frac{\lambda}{\mu})^j e^{-(\frac{\lambda}{\mu})}}{j!}$$

3.5. **The Machine Repair model.** The machine repair model is a M/M/R/GD/K/K queue system, where R is the number of servers, and K is both the size of the customer population and the maximum number of customers allowed in the system. This model can explain a situation where there are K machines that each break down at rate λ and R repair workers who can each fix a machine at rate μ. This means that both λ and μ are dependent on either how many machines are remaining in the population or how many repair workers are in service.

Let's model this as a birth-death process. Since λ_j depends on the number of machines left in the population that are not in service, we can say that

$$\lambda_j = (K - j)\lambda$$

We can calculate μ_j by looking at the number of repair workers currently in service. If a machine breaks down when all servers are busy, it waits in a queue to be served. We can calculate μ_j as follows:

$$\mu_j = j\mu \text{ for } j = 0, 1, ..., R$$

$$\mu_j = R\mu \text{ for } j = R + 1, R + 2, ..., K$$

The steady-state probability for a machine repair system, derived from page 1081 of the Winston text is

$$\pi_j = \binom{K}{j} p^j \pi_0 \text{ for } J = 0, 1, ..., R$$

$$\pi_j = \frac{\binom{K}{j} j! p^j \pi_0}{R! R^{j-R}} \text{ for } J = R + 1, R + 2, ..., K$$

3.6. **The M/G/s/GD/s/∞ Queuing System.** Another reasonable model of a queue is one where if a customer arrives and sees all of the servers busy, then the customer exits the system completely without receiving service. In this case, no actual queue is ever formed, and we say that the blocked customers have been cleared. Since no queue is ever formed, $L_q = W_q = 0$. If λ is the arrival rate and $1/\mu$ is the mean service time, then $W = W_s = 1/\mu$.

In this system, arrivals are turned away whenever s customers are present, so π_s is equal to the fraction of all arrivals who are turned away by the system. This means that an average of $\lambda\pi_s$ arrivals per unit of time will never enter the system, and thus, $\lambda(1-\pi_s)$ arrivals per unit of time will actually enter the system. This leads us to the conclusion based on Little's queuing formula that $L = L_s = \frac{\lambda(1-\pi_s)}{\mu}$.

4. Field Research

Trying to put this knowledge of queuing theory to some use, I took a trip to a local Bank of America to study their lines and see if their system could be fitted to a particular model. The Bank of America system proved to be a simple one to keep track of particularly because all of the customers waited in a single queue instead of several separate ones. That made it possible to accurately keep track of how long each customer had been waiting in line. However, the number of servers on duty during the period of study kept changing, and to keep the model from becoming overcomplicated, a reasonable value for s had to be chosen. There were 77 customers kept track of with a chosen value of 5 servers over the course of one hour of observation.

On average, a customer arrived every 46 seconds, or every 0.77 minutes. Thus, the sample interarrival rate was 1.30 customers per minute. Also, service completions took an average of 146 seconds, or 2.44 minutes. This means that the sample service rate was 0.41 service completions per minute. The average waiting time in the queue was 43.6 seconds, or 0.73 minutes. The average total time a customer spent in the system was 3.17 minutes. Ideally, this system could be modeled as a M/M/5/FCFS/∞/∞ system.

It is important to remember that a system such as this one is assumed to follow exponential distributions for interarrival times and service times. Setting $\lambda_a = 1.30$ and $\lambda_s = 0.41$, the possible exponential distributions for the interarrival and service times respectively were

$$f(t) = 1.3e^{-1.3t} \text{ and } f(t) = .41e^{-.41t}$$

In order to see if these distributions fit appropriately with the data, they were subjected to a chi-square goodness of fit test, the method of which can be found on page 1091 of the Winston text. For the interarrival times, the test resulted with $\chi^2(obs) = 2.026$, which was low enough to conclude that the exponential distribution can properly model the interarrival times. However, for the service times, the test resulted in $\chi^2(obs) = 36.57$, which was well out of the range for the service times to be properly modeled by the exponential distribution. It seems that this could be a consequence of using a relatively small sample size.

Chi-square tests were performed with other distributions as well, but nothing seemed to fit reasonably. Thus, we were unable to fully connect this study with one of the previously studied models. If data had been taken from the bank for a longer period of time, it seems likely that outliers would have had lesser effect on the chi-square tests. Given that the arrivals fit the exponential distribution, service times probably could have been exponential as well. To confirm that would require further observation.

5. Conclusion

With the knowledge of probability theory, input and output models, and birth-death processes, it is possible to derive many different queuing models, including but not limited to the ones we observed in this paper. Queuing theory can be applicable in many real world situations. For example, understanding how to model a multiple-server queue could make it possible to determine how many servers are actually needed and at what wage in order to maximize financial efficiency. Or perhaps a queuing model could be used to study the lifespan of the bulbs in street lamps in order to better understand how frequently they need to be replaced.

The applications of queuing theory extend well beyond waiting in line at a bank. It may take some creative thinking, but if there is any sort of scenario where time passes before a particular event occurs, there is probably some way to develop it into a queuing model. Queues are so commonplace in society that it is highly worthwhile to study them, even if only to shave a few seconds off one's wait in the checkout line.

6. References

[W] Wayne L Winston, Operations Research: Applications and Algorithms, 2nd edition, PWS-Kent Publishing, Boston, 1991.