QUEUEING THEORY

MADELYN ESTHER CRUZ

1. Introduction

A *queueing model* describes the ability of a system to meet demands of customers whose occurrences and durations are random. In the models we will tackle, we will consider three factors: the arrival process, the service mechanism, and the queue discipline.

We will use the notation $A/B/m$ to describe the queuing models, where A, B, and m stand for the arrival process, service time distribution, and number of servers, respectively.

The arrival process describes the amount of time between any consecutive arrivals. The service time is the time taken to complete one service. Note that there is at least one server and if there is more than one free server, the customers choose any of the servers at random. If all the servers are busy, they will join a queue and the first customer in that queue will be served first. We will also assume that the arrivals of the customers are independent, the service times are independent, and there is no maximum queue length.

We will be focusing on $M/M/c$ models, which deal with Poisson arrival process and exponential service time distribution, both of which denoted by M. In a Poisson arrival process, arrivals look like random arrivals. The processes we will look into are: 1) the *number of customers at a given time*, 2) the *waiting time of each customer*, and 3) the *busy period process*, which will be defined later.

2. M/M/1 Queue

Definition 2.1. An *M/M/1 queue* is a collection of random variables $L_0, L_1, L_2, \ldots$ whose state space is the set $\{0, 1, 2, 3, ...\}$, where each value denote the number of customers in the system. The state space diagram for this chain is shown in Figure 1. Let $L_t = L_{t-1} + X_t$, where X_t is a Poisson process with parameter λt, where λ is the arrival rate. Moreover, the service times are exponentially distributed with parameter μ, which is the service rate. In this model, there is only one server.

An *interarrival time* is the time between the time of arrival of one customer and the time of the arrival of the next customer, and a *service time* is the time it takes for one server to complete a customer service. With this definition, we have

$$\mathbb{E}[\text{interarrival time}] = \frac{1}{\lambda}$$

and

$$\mathbb{E}[\text{service time}] = \frac{1}{\mu}.$$

Proposition 2.2. *The rate of the process entering a state is the same as the rate of the process departing that same state.*

Date: November 21, 2020.

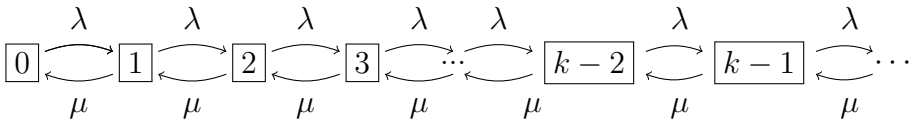

Figure 1. A pictorial representation of the $M/M/1$ queue

Proof. See proof of Theorem 1.4 of [1]. ■

Let C_k denote the k^{th} customer, which arrives at the system at time t_k, N_t be the number of arrivals during $(0,t]$, and L_t be the total number of customers in the system at time t. So, we have $N_t = \max\{k : t_k \leq t\}$. Also, suppose customer C_k spends $W_k \geq 0$ time units in the system, then C_k leaves at time $t_k^d = W_k + t_k$. Hence, we have

$$L_t = \sum_{k=1}^{\infty} \mathbb{1}(t)_{\{t_k \leq t < t_k^d\}} = \sum_{k:t_k \leq t} \mathbb{1}(t)_{\{t < t_k^d\}} = \sum_{k:t_k \leq t} \mathbb{1}(t)_{\{t-t_k < W_k\}} = \sum_{k=1}^{N_t} \mathbb{1}(t)_{\{W_k > t-t_k\}} \tag{2.1}$$

Theorem 2.3. *(Little's result) The average number of customers L in a stationary system is equal to the product of the average arrival rate λ and the average time W a customer spends in the system, given that λ, W, and L exist and λ and W are finite.*

Proof. This proof is based on [2]. First, we have the following equations (if the limits exist):

$$\lambda = \lim_{t\to\infty} \frac{N_t}{t} \tag{2.2}$$

$$W = \lim_{k\to\infty} \frac{1}{k} \sum_{j=1} W_j \tag{2.3}$$

$$L = \lim_{t\to\infty} \frac{1}{t} \int_0^t L_s \mathrm{d}s \tag{2.4}$$

By (2.1), we get

$$\int_0^t L_s \mathrm{d}s = \int_0^t \left[\sum_{j:t_j \leq s \leq t} \mathbb{1}(s)_{\{W_j > s-t_j\}} \right] \mathrm{d}s = \sum_{j:t_j \leq t} \int_{t_j}^t \mathbb{1}(s)_{\{W_j > s-t_j\}} \mathrm{d}s = \sum_{j:t_j \leq t} \min\{W_j, t-t_j\}.$$

Since $\min\{W_j, t-t_j\} \leq W_j$, we have

$$\int_0^t L_s \mathrm{d}s \leq \sum_{j:t_j \leq t} W_j.$$

Also,

$$\begin{aligned} \sum_{j:t_j \leq t} \min\{W_j, t-t_j\} &= \sum_{j:W_j \leq t-t_j} W_j + \sum_{j:t-t_j \geq 0, W_j > t-t_j} (t-t_j) \\ &\geq \sum_{j:W_j \leq t-t_j} W_j = \sum_{j:t_j^d \leq t} W_j. \end{aligned}$$

Hence,

$$\sum_{j:t_j^d \le t} W_j \le \int_0^t L_s \mathrm{d}s$$

and

$$\sum_{j:t_j^d \le t} W_j \le \int_0^t L_s \mathrm{d}s \le \sum_{j:t_j \le t} W_j = \sum_{j=1}^{N_t} W_j.$$

Therefore,

$$\frac{1}{t}\sum_{j:t_j^d \le t} W_j \le \frac{1}{t}\int_0^t L_s \mathrm{d}s \le \frac{N_t}{t} \cdot \frac{1}{N_t}\sum_{j:t_j \le t} W_j = \sum_{j=1}^{N_t} W_j.$$

Taking the limit as $t \to \infty$, we get $L = \lambda W$. ■

Many articles show that the Little's Result is a general property of queueing models, so we will use this to get some performance measures. We are now ready to get the number of customers and the waiting time of each customer. For practicality, we will study the stationary behavior of the chains, and we denote π_k to be the limiting probability that there are k customers in the system.

Proposition 2.4. *The average number of customers in the system is* $\frac{\lambda}{\mu-\lambda}$ *and the average time a customer spends in the system is* $\frac{1}{\mu-\lambda}$.

Proof. We will first get the probability of having k customers in the queue, which we denote by π_k. Notice that the rate of entering state 0 is $\mu\pi_1$ (from state 1) and the rate of leaving 0 is $\lambda\pi_0$. By Proposition 2.2,

$$\lambda\pi_0 = \mu\pi_1.$$

Now, consider state k, where $k = 1, 2, \ldots$. The rate of entering state k is $\lambda\pi_{k-1} + \mu\pi_{k+1}$ and the rate of leaving k is $\lambda\pi_k + \mu\pi_k$. Thus, we have $\lambda\pi_k + \mu\pi_k = \lambda\pi_{k-1} + \mu\pi_{k+1}$. Solving these systems of equations, we get that

$$\pi_k = \left(\frac{\lambda}{\mu}\right)^k \pi_0.$$

Since $\sum_{k=1}^{\infty} \pi_k = 1$, we have $\frac{\pi_0}{1-\frac{\lambda}{\mu}} = 1$. Hence,

$$\pi_k = \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right).$$

Let Q be the number of customers in the system. Then,

$$\begin{aligned} E[Q] &= \sum_{k=1}^{\infty} k \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right) \\ &= \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} = \frac{\lambda}{\mu - \lambda}. \end{aligned}$$

It is necessary that $\lambda < \mu$ since if $\frac{\lambda}{\mu} > 1$, the sum diverges and the system will not have a stationary distribution. By Theorem 2.3, we have the average time a customer spends in the system is

$$T = \frac{E[Q]}{\lambda} = \frac{1}{\mu - \lambda}.$$

■

Definition 2.5. Suppose $X_0 = k$, then the time it takes until there are 0 customers is called the busy period initiated by k customers. The busy period initiated by 1 customer is the *busy period*. The end of a busy period is the start of the *idle period*.

Let B and I be the busy and idle periods, respectively Then $\frac{E[B]}{E[B]+E[I]}$ is the proportion of time a server is busy, so this is equal to $\frac{\lambda}{\mu}$. Also, $E[I] = \frac{1}{\lambda}$ since the mean time until a new customer arrives is exponentially distributed with parameter λ. Solving for $E[B]$, we get $E[B] = \frac{1}{\mu-\lambda}$, where $\mu > \lambda$. Notice that the mean of the busy period is equal to the time a customer spends in the system.

The time a customer spends in the system is equal to the sum of the waiting time for service T_q of an arriving customer and the service time. Thus,

$$T_q = T - \frac{1}{\lambda} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)}.$$

Example. An operator finds that the time ey spent on fixing a phone is exponentially distributed with mean 20 minutes. Ey operates in a first in first out manner and the arrival of the customers is Poisson distributed with an average rate of 50 per day.

The arrival rate is $\lambda = 50$ per day and the service rate is $\mu = 72$ per day. The M/M/1 queue reaches a steady state since $\frac{50}{72} < 1$. The expected idle time per day is the proportion of time spent doing nothing which is $1 - \frac{50}{72} = \frac{11}{36}$. The expected number of customers is $\frac{\lambda}{\mu-\lambda} = \frac{50}{72-50} = \frac{25}{11}$. The average time a customer spends in the system (waiting for service + service time) and the mean of the busy period is $\frac{1}{22}$ day and the average time of a customer waiting for service is $\frac{50}{72 \cdot 22} = \frac{25}{792}$ day.

3. M/M/∞ Queue

If the mean arrival rate is larger than the service rate, we will not reach a stable distribution, so more servers are needed. In the M/M/c queue, there are c servers, and the properties of the M/M/1 queue still hold. We will also assume that the servers provide service independently, and that the customers form a single queue. If c is ∞, then there is an infinite number of servers.

Definition 3.1. An *M/M/∞ queue* is a collection of random variables $L_0, L_1, L_2, \ldots$ whose state space is the set $\{0, 1, 2, 3, ...\}$, where each value denote the number of customers in the system. The state space diagram for this chain is shown in Figure 2. In an $M/M/\infty$ queue, there is an infinite number of servers. Moreover, customers arrive randomly for service in a Poisson process and the service times are exponentially distributed, similar to the $M/M/1$ queue. Let λ be the rate of arrival rate, and μ_i be the service rate when there are i customers. Then, we have $\mu_i = i\mu$.

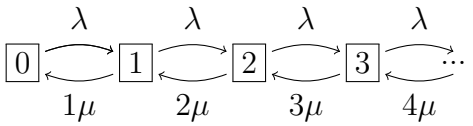

Figure 2. A pictorial representation of the $M/M/\infty$ queue

Proposition 3.2. *The average number of customers in the system is* $\frac{\lambda}{\mu}$ *and the average time a customer spends in the system is* $\frac{1}{\mu}$*. Furthermore, the waiting time for service of a customer is* 0*.*

Proof. The probability of having k customers is

$$\pi_k = \pi_0 \prod_{i=1}^{n} \frac{\lambda}{\mu_i} = \pi_0 \prod_{i=1}^{n} \frac{\lambda}{i\mu} = \pi_0 \frac{\lambda^k}{\mu^k k!}.$$

So, we have

$$\sum_{k=0}^{\infty} \pi_k = \sum_{k=0}^{\infty} \frac{\lambda^k}{\mu^k k!} \pi_0 = e^{\frac{\lambda}{\mu}} \pi_0 = 1.$$

Thus,

$$\pi_k = \frac{\lambda^k}{\mu^k k!} e^{\frac{-\lambda}{\mu}}.$$

Clearly, this is Poisson distributed so $E[Q] = \frac{\lambda}{\mu}$. By Little's Theorem, the average time a customer spends in the system is

$$T = \frac{E[Q]}{\lambda} = \frac{\lambda}{\mu\lambda} = \frac{1}{\mu}.$$

Moreover, the waiting time for service T_q of a customer is

$$T_q = \frac{1}{\mu} - \frac{1}{\mu} = 0.$$

This makes sense because there are infinite servers so each customer will be served immediately. This also explains the following proposition. ■

Proposition 3.3. *In a* $M/M/\infty$ *queue, there always exists a stationary distribution* π*.*

Example. Consider a company with 1000 employees. Suppose the employment rate and the retirement rate is constant and that there are no people who are going to enter or leave the job, unless they are retiring. Also, assume that the number of years employees stay at a company is exponentially distributed. It is given that the average number of years employees stay at a company is 30 years.

We will use the $M/M/\infty$ model. We are given that $\mathbb{E}[Q] = 1000$ and $\mu^{-1} = 30$. Since $\mathbb{E}[Q] = \frac{\lambda}{\mu}$, we get that $\lambda = \frac{100}{3}$. Thus, the employment rate is $\frac{100}{3}$ new employees per year.

4. M/M/c Queue

Definition 4.1. In a *M/M/c queue*, customers arrive randomly for service in a Poisson process and the service times are exponentially distributed like the previous models. There

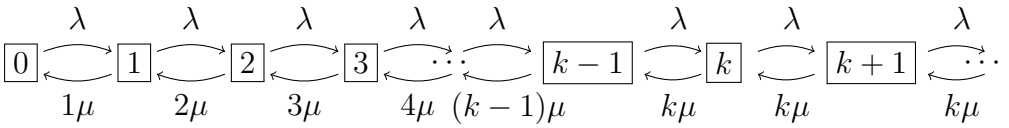

Figure 3. A pictorial representation of the $M/M/c$ queue

are k servers in this model. Also, let λ be the rate of arrival rate, and μ_i be the service rate when there are i customers. Define

$$\mu_i = \begin{cases} i\mu & 0 \leq i \leq c \\ c\mu & i > c. \end{cases}$$

The state space diagram for this chain is shown in Figure 3. The probability of having k customers is

$$\pi_k = \begin{cases} \frac{\lambda\lambda\cdots\lambda}{(\mu)(2\mu)\cdots(k\mu)}\pi_0 = \frac{\lambda^k}{\mu^k k!}\pi_0 & 0 \leq k \leq c \\ \frac{\lambda\lambda\cdots\lambda}{(\mu)(2\mu)\cdots(c\mu)(c\mu)\cdots(c\mu)}\pi_0 = \frac{\lambda^k}{\mu^k c! c^{k-c}}\pi_0 & k > c \end{cases}.$$

Thus, we have

$$\sum_{k=0}^{\infty}\pi_k = \sum_{k=0}^{c-1}\frac{\lambda^k}{\mu^k k!}\pi_0 + \sum_{k=c}^{\infty}\frac{\lambda^k}{\mu^k c! c^{k-c}}\pi_0 = 1.$$

So,

$$\pi_0 = \left[\sum_{k=0}^{c-1}\frac{\lambda^k}{\mu^k k!} + \sum_{k=c}^{\infty}\frac{\lambda^k}{\mu^k c! c^{k-c}}\right]^{-1} = \left[\sum_{k=0}^{c-1}\frac{\lambda^k}{\mu^k k!} + \frac{\frac{\lambda^c}{\mu^c}}{c!(1-\frac{\lambda}{\mu c})}\right]^{-1}$$

and it is necessary that $\lambda < \mu$ since if $\frac{\lambda}{\mu} > 1$, the sum diverges and the system will not have a stationary distribution.

Theorem 4.2. *(Erlang's C formula) The proportion of time that all k servers are busy is*

$$C_k\left(\frac{\lambda}{\mu}\right) = \mathbb{P}\left[\sum_{n=k}^{\infty}\pi_n\right] = \frac{(\frac{\lambda}{\mu})^k k\pi_0}{k!(k-\frac{\lambda}{\mu})}.$$

Since the busy period for an $M/M/c$ queue is complicated , we will only get the expected number of busy servers.

Proposition 4.3. *The mean number of busy servers is $\frac{\lambda}{\mu}$.*

Proof. Let the number of busys servers be S. Then, we have

$$\begin{aligned}\mathbb{E}[S] &= \sum_{k=0}^{c-1} k\pi_k + \sum_{k=c}^{\infty} c\pi_k = \sum_{k=0}^{c-1} k\frac{\lambda^k}{\mu^k k!}\pi_0 + \sum_{k=c}^{\infty} c\frac{\lambda^k}{\mu^k c! c^{k-c}}\pi_0 \\ &= \frac{\lambda}{\mu}\sum_{k=1}^{c-1}\frac{\lambda^{k-1}}{\mu^{k-1}(k-1)!}\pi_0 + \frac{c\pi_0\frac{\lambda^c}{\mu^c c!}}{1-\frac{\lambda}{\mu c}} = \frac{\lambda}{\mu}\sum_{k=0}^{c-2}\frac{\lambda^k}{\mu^k k!}\pi_0 + \frac{c\pi_0\frac{\lambda^c}{\mu^c}}{c!(1-\frac{\lambda}{\mu c})} \\ &= \frac{\lambda}{\mu}\sum_{k=0}^{c-1}\frac{\lambda^k}{\mu^k k!}\pi_0 + \frac{c\pi_0\frac{\lambda^c}{\mu^c}}{c!(1-\frac{\lambda}{\mu c})} - \frac{\lambda}{\mu}\cdot\frac{\lambda^{c-1}}{\mu^{c-1}(c-1)!}\pi_0 \\ &= \frac{\lambda}{\mu}\left[\sum_{k=0}^{c-1}\frac{\lambda^k}{\mu^k k!} + \frac{\frac{\lambda^c}{\mu^c}}{c!(1-\frac{\lambda}{\mu c})}\right]^{-1}\pi_0 = \frac{\lambda}{\mu}.\end{aligned}$$

■

Proposition 4.4. *The average number of customers in the system is*

$$\frac{\lambda}{\mu c-\lambda}C_c\left(\frac{\lambda}{\mu}\right)+\frac{\lambda}{\mu}$$

and the average time a customer spends in the system is

$$\frac{C_c\left(\frac{\lambda}{\mu}\right)}{\mu c-\lambda}+\frac{1}{\mu}.$$

Furthermore, the waiting time for service of a customer is

$$\frac{C_c\left(\frac{\lambda}{\mu}\right)}{\mu c-\lambda}.$$

Proof. Let $Q' = Q + c$ be the total number of customers in the system (c customers are being served, Q customers are in the queue). Then,

$$\mathbb{E}[Q'] = \mathbb{E}[Q] + \mathbb{E}[B].$$

First, we compute for $\mathbb{E}[Q]$. We get

$$\begin{aligned}\mathbb{E}[Q] &= \sum_{k=c}^{\infty}(k-c)\pi_k = \sum_{k=c}^{\infty}(k-c)\frac{\lambda^k}{\mu^k c! c^{k-c}}\pi_0 \\ &= \frac{\pi_0\lambda^c}{c!\mu^c}\sum_{k=c}^{\infty}(k-c)\frac{\lambda^{k-c}}{\mu^{k-c}c^{k-c}} = \frac{\pi_0\lambda^c}{c!\mu^c}\cdot\frac{\frac{\lambda}{\mu c}}{(1-\frac{\lambda}{\mu c})^2} \\ &= \frac{\frac{\lambda}{\mu c}}{1-\frac{\lambda}{\mu c}}\cdot\frac{(\frac{\lambda}{\mu})^c c\pi_0}{c!(c-\frac{\lambda}{\mu})} = \frac{\lambda}{\mu c-\lambda}C_c\left(\frac{\lambda}{\mu}\right).\end{aligned}$$

Then, we have

$$\mathbb{E}[Q'] = \frac{\lambda}{\mu c-\lambda}C_c\left(\frac{\lambda}{\mu}\right)+\frac{\lambda}{\mu}.$$

Thus, the waiting time for service T_q of a customer or the waiting time in the queue is

$$\frac{\mathbb{E}[Q]}{\lambda} = \frac{\frac{\lambda}{\mu c-\lambda} C_c\left(\frac{\lambda}{\mu}\right)}{\lambda} = \frac{C_c\left(\frac{\lambda}{\mu}\right)}{\mu c-\lambda}$$

and the average time T a customer spends in the system is

$$\frac{\mathbb{E}[Q']}{\lambda} = \frac{\frac{\lambda}{\mu c-\lambda} C_c\left(\frac{\lambda}{\mu}\right) + \frac{\lambda}{\mu}}{\lambda} = \frac{C_c\left(\frac{\lambda}{\mu}\right)}{\mu c-\lambda} + \frac{1}{\mu}.$$

■

Example. Consider an $M/M/1$ queue with a server operating at $c\mu$ and an $M/M/c$ queue with a server operating at μ. We will compare the average time a customer spends in the system for the two queues.

For the $M/M/c$ queue, we have $\mathbb{E}[T_c] = \frac{C_c\left(\frac{\lambda}{\mu}\right)}{\mu c-\lambda} + \frac{1}{\mu}$ and for the $M/M/1$ queue, we have $\mathbb{E}[T_1] = \frac{1}{\mu c-\lambda}$. As $\lambda \ll \mu$, i.e. under few tasks, we have

$$\frac{E[T_1]}{E[T_c]} \approx c.$$

If the load is heavy, we have $\mu c - \lambda \ll \mu$,

$$\frac{E[T_1]}{E[T_c]} \approx 1.$$

References

[1] Jyotiprasad Medhi. "Stochastic models in queueing theory". In: Elsevier, 2002. Chap. 1-3.

[2] Karl Sigman. *Notes on Little's law.* URL: `http://www.%20columbia.%20edu/%E2%88%BC%20ks20/stochastic-I/stochastic-I-LL.%20pdf`. (accessed: 11.09.2020).