# Chapter 8: Queueing Theory

## 1 Queueing Models and Characteristics

A *queueing system* consists of arrival streams of *customers* and a series of *servers*. When there are more customers than available servers, the remaining customers are said to *wait in queue*. Customers leave the system eventually, being turned away, balking or reneging, or finishing service. Queueing theory is the theory of waiting lines. Draw the queueing model.

Queueing systems are often described by the notation $A/B/s/K$ (originally due to Kendall), where $A$ stands for the arrival distribution and $B$ stands for the service distribution ($D$=deterministic, $M$=exponential (memoryless), and $G$=general). The interarrival times and service times are assumed to form i.i.d. sequences that are independent of each other. The number of servers in parallel is $s$ and $K$ is the number of customers the system can hold. If $K$ is not given then it is assumed to be infinite. Unless otherwise stated, service order is assumed to be First-In-First-Out (FIFO), otherwise known as First-Come-First-Serve (FCFS). Other common service disciplines are Last-Come-First-Serve (LCFS) and Shortest Expected Processing Time (SEPT).

Examples: Since server queues ($M/M/1$, $M/M/1/K$), multiple server queues ($M/M/s$), and call centers $M/M/s/K$).

Questions:

(i.) Average number of customers in the system $L$ ($= \frac{\lambda}{\lambda+\mu}$) and in the queue $L_Q$ ($= \frac{\rho^2}{1-\rho}$), in service $L_s$ ($= \rho$). $L = L_Q + L_s$.

(ii.) Average amount of time a customer spends in the system $W$ ($= \frac{1}{\mu-\lambda}$) and in the queue $W_Q$ ($= \frac{\rho}{1-\rho}$). $W = W_Q + \frac{1}{\mu}$.

(iii.) Probability $W > t$ ($= e^{-(\mu-\lambda)t}$) or $W_Q > t$ ($= \rho e^{-(\mu-\lambda)t}$).

(iv.) Probability a customer does not need to wait $P_0$.

(v.) Probability a customer finds at least 3 customers ahead of him upon arrival $1 - P_0 - P_1 - P_2$.

## 2 The $M/M/1$ Queues

We know that $P_n = \rho^n(1-\rho)$ where $\rho = \frac{\lambda}{\mu}$ is the utilization of the server, or the traffic intensity. The condition $\rho < 1$ is called the *stability condition* for the $M/M/1$ queue. Thus,

$$L = \sum_{n=0}^{\infty} nP_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n.$$

Let $\sum_{n=0}^{\infty} n\rho^n = Z$. Then,

$$\begin{aligned} \rho + 2\rho^2 + 3\rho^3 + \cdots &= Z, \\ \rho^2 + 2\rho^3 + \cdots &= \rho Z, \\ - - - - - - - - - - & - - - - \\ \rho + \rho^2 + \rho^3 + \cdots &= (1-\rho)Z. \end{aligned}$$

Hence, $Z = \frac{\rho}{(1-\rho)^2}$ and $L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$. As $\rho \to 1$, $L \to \infty$. To see the magnitude,

| $\rho$ | 0.8 | 0.85 | 0.9 | 0.95 | 0.98 |
|---|---|---|---|---|---|
| $L = \frac{\rho}{1-\rho}$ | $\frac{0.8}{0.2} = 4$ | $\frac{0.85}{0.15} = 5.7$ | $\frac{0.9}{0.1} = 9$ | $\frac{0.95}{0.05} = 19$ | $\frac{0.98}{0.02} = 49$ |

The fraction of time the server is idle is $P_0 = 1-\rho$ and $\rho$ is the fraction of time the server is busy.

$$\begin{aligned} W &= \sum_{n\to\infty} E(\text{time in system}|\text{you observe } n \text{ in the system upon arrival})P_n \\ &= \sum_{n=0}^{\infty} \frac{n+1}{\mu}\rho^n(1-\rho) = \frac{1-\rho}{\mu}\sum_{n=0}^{\infty}(n+1)\rho^n = \frac{1-\rho}{\mu}\left(\sum_{n=0}^{\infty} n\rho^n + \sum_{n=0}^{\infty}\rho^n\right) \\ &= \frac{1-\rho}{\mu}\left[\frac{\rho}{(1-\rho)^2} + \frac{1}{1-\rho}\right] = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}. \end{aligned}$$

Note that, $L = \lambda W$ which is the well known Little's Law and apply to any system.
With Little's Law, we can calculate the following using $L = \frac{\rho}{1-\rho}$ and $W = \frac{1}{\mu-\lambda}$:

$$\begin{aligned} L_s &= \lambda \times \frac{1}{\mu} = \rho, \qquad \text{exactly the percentage of time the server is busy,} \\ L_Q &= L - L_s = \frac{\rho^2}{1-\rho}, \\ W_Q &= \frac{L_Q}{\lambda} = \frac{\rho}{\mu-\lambda}, \\ W_Q + \frac{1}{\mu} &= \frac{1}{\mu-\lambda} = W. \end{aligned}$$

**Example:** Which cashier to hire? Cashier A is faster with mean service time of 1 min and a salary of \$10, while cashier B is slower with mean service time of 2 min and a salary of \$5. Customers arrive at a rate of $\lambda = 25$ per hr. Assume it costs \$0.02 (including time being served) for each minute (\$1.2/hr) a customer is in the system.

If B is hired, $\mu = 30$ per hr and $\rho = \frac{5}{6}$. $W = \frac{1}{\mu-\lambda} = \frac{1}{30-25} = 0.2hrs = 12min$, $L = \lambda W = 25 \times 0.2 = 5$, $W_Q = \frac{\rho}{\mu-\lambda} = \frac{1}{6} = 10min$, $L_Q = \frac{\rho^2}{1-\rho} = \frac{25}{6} = 4.17$. The cost is $\$5 + \$1.2 \times 0.2hr \times 25/hr = \$11/hr$.
If A is hired, $\mu = 60$ per hr and $\rho = \frac{5}{12}$ (half the utilization). $W = \frac{1}{\mu-\lambda} = \frac{1}{60-25} = \frac{1}{35}hrs = 1.7min$, $L = \lambda W = 25 \times \frac{1}{35} = \frac{5}{7} = 0.7$, $W_Q = \frac{\rho}{\mu-\lambda} = \frac{1}{84}hrs = 0.7min$, $L_Q = \lambda W_Q = \frac{25}{84} = 0.3$. The cost is $\$10 + \$1.2 \times \frac{1}{35}hr \times 25/hr = \$10.86/hr$.
Conditioning on the number of customers the customer sees in the system upon his arrival, $N$, we have

$$\begin{aligned}
P(\text{time in system} > t) &= \sum_{n=0}^{\infty} P(\text{time in system} > t | N = n) P(N = n) \\
&= \sum_{n=0}^{\infty} P_n P(\text{time in system} > t | N = n) \\
&= \sum_{n=0}^{\infty} (1-\rho)\rho^n \sum_{k=0}^{n} e^{-\mu t} \frac{(\mu t)^k}{k!}, \quad Gamma(n+1, \mu) \\
&= (1-\rho)e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \sum_{n=k}^{\infty} \rho^n \\
&= e^{-\mu x} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} \rho^k \\
&= e^{-(\mu-\lambda)t}.
\end{aligned}$$

That is, in steady state, the time a customer spends an exponential amount of time with $\mu - \lambda$ in the system. So the probability a customer spends more than $t$ amount of time in the system is $e^{-(\mu-\lambda)t}$. It can be shown that $P(w_Q > t) = \rho e^{-(\mu-\lambda)t}$.
**Example**: Lead time quotation.

- Quote a fixed lead time $t$ to all customers to achieve a service level $SL$. Note that $\bar{F}(t)$ is the probability that a customer stays more than $t$ amount of time in the system. Then one should quote the shortest lead time such that $\bar{F}(t) = e^{-(\mu-\lambda)t} \leq 1 - SL$ or $t = -\frac{1}{\mu-\lambda}\ln(1 - SL)$.
- Quote lead times based on the work load. Let $t_n$ be the lead time quoted to a customer when there are $n$ customers in the system upon his arrival, $n \geq 0$. Since the amount of time the customer will spend in the system is gamma$(n+1, \mu)$, one should quote the shortest lead time such that $\bar{F}(t_n) = e^{-\mu t_n} \sum_{k=0}^{n} \frac{(\mu t_n)^k}{k!} = 1 - SL$. In this case, the lead time increases as the congestion level increases.

# 3 $M/M/1/K$ Queues

Let $X(t)$ be the number of customers in the system at time $t$. Then, $\{X(t), t \geq 0\}$ is a birth and death process on state space $\{0, 1, \ldots, K\}$ with

$$\begin{aligned} \lambda_n &= \begin{cases} \lambda, & \text{for } 0 \leq n < K, \\ 0, & \text{for } n \geq K, \end{cases} \\ \mu_n &= \mu \quad (n \geq 1), \quad \mu_0 = 0. \end{aligned}$$

Now $\frac{\lambda}{\mu}$ (it is no longer the actual utilization) and can be anything. Solving the balance equations in terms of $P_0$ yields

| State | Balance equation | | |
|---|---|---|---|
| 0 | $\lambda P_0 = \mu P_1$ | $\Rightarrow$ | $P_1 = \frac{\lambda}{\mu} P_0$ |
| 1 | $(\lambda + \mu) P_1 = \lambda P_0 + \mu P_2$ | $\Rightarrow$ | $P_2 = \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu}\right)^2 P_0$ |
| $\vdots$ | | | |
| $K-2$ | $(\lambda + \mu) P_{K-2} = \lambda P_{K-3} + \mu P_{K-1}$ | $\Rightarrow$ | $P_{K-1} = \frac{\lambda}{\mu} P_{K-2} = \left(\frac{\lambda}{\mu}\right)^{K-1} P_0$ |
| $K$ | $\lambda P_{K-1} = \mu P_K$ | $\Rightarrow$ | $P_K = \frac{\lambda}{\mu} P_{K-1} = \left(\frac{\lambda}{\mu}\right)^{K} P_0$ |

So, $P_0 = \left[\sum_{i=0}^{K} \left(\frac{\lambda}{\mu}\right)^i\right]^{-1}$ and $P_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{\sum_{i=0}^{K}\left(\frac{\lambda}{\mu}\right)^i}$. One can calculate $L = \sum_{n=0}^{K} n P_n$.

In steady state, the arrival customer will join the queue with probability $1 - P_K$. So the actual arrival rate is $\lambda' = \lambda(1 - P_K)$. The actual utilization is $\frac{\lambda}{\mu}(1 - P_K) = \frac{\lambda}{\mu} \sum_{n=0}^{K-1} P_n = \sum_{n=1}^{K} P_n = 1 - P_0 < 1$. The average number blocked per unit time is $\lambda P_K$. By Little's Law, for those who enter, $W = \frac{L}{\lambda'} = \frac{L}{\lambda(1-P_K)}$.

**Example** A gas station with a single pump and maximum 3 cars allowed. $\lambda = 60$/hr and average service time 2 min or $\mu = 30$/hr. So $\frac{\lambda}{\mu} = 2$, $P_0 = \frac{1}{1+2+2^2+2^3} = \frac{1}{15}$, $P_1 = \frac{2}{15}$, $P_2 = \frac{4}{15}$ and $P_3 = \frac{8}{15}$. $L = \frac{2}{15} + 2 \times \frac{4}{15} + 3 \times \frac{8}{15} = \frac{34}{15}$ and $W = \frac{L}{\lambda(1-P_3)} = \frac{34/15}{60(1-8/15)} = \frac{17}{210} hrs = \frac{34}{7} min.$

# 4 $M/M/c$ Queues

Let $X(t)$ be the number of customers in the system at time $t$. Then $\{X(t), t \geq 0\}$ is a birth and death process with

$$\begin{aligned} \lambda_n &= \lambda, \quad n \geq 0, \\ \mu_n &= \begin{cases} n\mu, & \text{for } 0 \leq n \leq c, \\ c\mu, & \text{for } n > c. \end{cases} \end{aligned}$$

Let $\rho = \frac{\lambda}{c\mu}$ or $\frac{\lambda}{\mu} = c\rho$. Thus, the stability condition of the queue is $\rho < 1$ or $c > \frac{\lambda}{\mu}$. Solving the balance equations in terms of $P_0$ yields,

| State | Balance equation | | |
|---|---|---|---|
| 0 | $\lambda P_0 = \mu P_1$ | $\Rightarrow$ | $P_1 = \frac{\lambda}{\mu} P_0 = c\rho P_0$ |
| 1 | $(\lambda+\mu)P_1 = \lambda P_0 + 2\mu P_2$ | $\Rightarrow$ | $P_2 = \frac{\lambda}{2\mu} P_1 = \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 P_0 = \frac{1}{2}(c\rho)^2 P_0$ |
| 2 | $(\lambda+2\mu)P_2 = \lambda P_1 + 3\mu P_3$ | $\Rightarrow$ | $P_3 = \frac{\lambda}{3\mu} P_2 = \frac{1}{3}\cdot\frac{1}{2}\left(\frac{\lambda}{\mu}\right)^3 P_0 = \frac{1}{3!}(c\rho)^3 P_0$ |
| $\vdots$ | | | |
| $c-1$ | $(\lambda+c\mu)P_{c-1} = \lambda P_{c-2} + c\mu P_c$ | $\Rightarrow$ | $P_c = \frac{\lambda}{c\mu} P_{c-1} = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c P_0 = \frac{1}{c!}(c\rho)^c P_0$ |
| $c$ | $(\lambda+c\mu)P_c = \lambda P_{c-1} + c\mu P_{c+1}$ | $\Rightarrow$ | $P_{c+1} = \frac{\lambda}{c\mu} P_c = \frac{1}{cc!}\left(\frac{\lambda}{\mu}\right)^{c+1} P_0 = \frac{c^c}{c!}\rho^{c+1} P_0$ |
| $c+1$ | | | $P_{c+2} = \frac{\lambda}{c\mu} P_c = \frac{1}{c^2 c!}\left(\frac{\lambda}{\mu}\right)^{c+2} P_0 = \frac{c^c}{c!}\rho^{c+2} P_0$ |
| $\vdots$ | | | |
| $n-1$ | $(\lambda+c\mu)P_{n-1} = \lambda P_{n-2} + c\mu P_n$ | $\Rightarrow$ | $P_n = P_{c+(n-c)} = \frac{1}{c^{n-c}c!}\left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{c^c}{c!}\rho^n P_0$ |

$$P_n = \begin{cases} \frac{(c\rho)^n}{n!} P_0, & \text{for } 0 \le n \le c-1, \\ \frac{c^c \rho^n}{c!} P_0, & \text{for } n \ge c, \end{cases}$$

and

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

We first calculate

$$\begin{aligned} L_Q &= \sum_{n=c+1}^{\infty}(n-c)P_n = \sum_{n=c+1}^{\infty}(n-c)\frac{c^c\rho^n}{c!}P_0 = \frac{(c\rho)^c}{c!}P_0\sum_{n=c+1}^{\infty}(n-c)\rho^{n-c} \\ &= \frac{(c\rho)^c P_0}{c!}\sum_{n=1}^{\infty} n\rho^n = \frac{(c\rho)^c P_0}{c!}\frac{\rho}{(1-\rho)^2}. \end{aligned}$$

By Little's Law, $L = L_Q + L_s = L_Q + \frac{\lambda}{\mu} = L_Q + c\rho$ and

$$W = \frac{L}{\lambda} = \frac{1}{\mu} + \frac{1}{c\mu-\lambda} \times \frac{(c\rho)^c}{c!(1-\rho)}P_0 = W_Q + \frac{1}{\mu}$$

as $W_s = \frac{1}{\mu}$. The probability an arrival does not need to wait is $\sum_{n=0}^{c-1} P_n$. The probability that an arrival has to wait for $t$ or longer is $\left(1 - \sum_{n=0}^{c-1} P_n\right) e^{-(c\mu-\lambda)t} = e^{-(c\mu-\lambda)t}\frac{(c\rho)^c}{c!(1-\rho)}P_0$.

**Examples**

**Resource pooling at a call center** $\lambda = 10.8/hr$ and $\mu = 6/hr$.

Pooled with one number (two extensions): When $c = 2$,

$$\begin{aligned} P_0 &= \left(1 + 2\rho + \frac{2\rho^2}{1-\rho}\right)^{-1} = \frac{1-\rho}{1+\rho}, \\ L_Q &= \frac{(2\rho)^2 P_0}{2}\frac{\rho}{(1-\rho)^2} = \frac{2\rho^3}{1-\rho^2}. \end{aligned}$$

$\rho = \frac{\lambda}{c\mu} = 0.9$, $L_Q = 7.7$ calls, and $W_Q = \frac{L_Q}{\lambda} = \frac{7.7}{10.8}$hrs = 43min. Examples: bank, airport, call centers.

Two numbers each with a single server queue and $\lambda = 5.4/hr$: $\rho = \frac{\lambda}{\mu} = 0.9$, $L_Q = \frac{\rho^2}{1-\rho} = 8.1$ calls, and $W_Q = \frac{8.1}{5.4}$hrs = 90min. Examples: grocery stores, McDonald's, toll booths.

**One faster server vs. two slower servers**

| | Faster server with $\mu$ | | slower servers with $\frac{\mu}{2}$ |
|---|---|---|---|
| $\rho$ | $\frac{\lambda}{\mu}$ | $=$ | $\frac{\lambda}{2\mu/2} = \frac{\lambda}{\mu}$ |
| $L_Q$ | $\frac{\rho^2}{1-\rho}$ | $>$ | $\frac{2\rho^3}{1-\rho^2}$ |
| $W_Q$ | $\frac{\rho^2}{\lambda(1-\rho)}$ | $>$ | $\frac{2\rho^3}{\lambda(1-\rho^2)}$ |
| $W$ | $\frac{\rho^2}{\lambda(1-\rho)} + \frac{1}{\mu}$ | $<$ | $\frac{2\rho^3}{\lambda(1-\rho^2)} + \frac{2}{\mu}$ |

Which is better? Depends on price and what better means. If you don't mind being served longer, two slow ones may be better. If total time in the system matters, one may be better. It leads to the discussion of the psychological side of waiting.

However, a good approximation (exact when $c = 1$) is

$$L_Q = \frac{\rho\sqrt{2(1+c)}}{1-\rho}$$

and $W_Q = L_Q/\lambda$.

**The psychology of queueing**

- A customer's tolerance for waiting in a queue is proportional to the complexity or quantity of service anticipated by that customer.

  Customers at a bank prefers FCFS that assures social justice, but generally does not mind if those with "12 items or less" join a special express service line in a supermarket. They expect to wait longer. In an airport, a customer tolerates a multi-hour delay in making a connection to an overseas flight while even a fraction of that wait for a New York to Washington Shuttle would not be tolerated. Banks have been neglecting this mixing of high service time and low service time customers. So far many customers opted to go to an ATM, but maybe banks should set up preferred customers accounts (common in Latin America) like first class passengers and everyone understands it. This is not so apparent in a teller line (why should someone came 20 minutes later get served before me). Research should be done before customers started to move away. Besides, a preferred customer may start life as a non-preferred customer.

- It is not the duration of the delay that matters; it is what you experience while you are waiting that matters.

How a customer feels in a queue is dependent on the duration of the wait and the total environment around her. Adding servers will lower the duration, but cost will go up. Disney surrounds its queue lines with entertainment and other diversions. The line-skipping system called FastPass allows guests to book a time for an attraction, leave to do other things, and return at an allotted time. One major initiative of NextGen focuses on what is being called an xPASS, which would allow guests to book rides weeks or months in advance. Visitors planning their trip would go on the xPASS website and use the free service which allows you to reserve experiences, including ride times, exclusive meet-and-greets with Disney characters, even viewing spots for the nightly fireworks. The xPASS system would also help to avoid lines at restaurants by ordering food in advance.

A couple of things about this strike me as interesting. First, arranging meet and greets as well as saving spots for fireworks seems pretty easy to do but rides are something else. They are prone to breakdowns so actually getting everyone with a 2:00 reservation on Pirates of the Caribbean might be tough. There is also a question of how much capacity one makes available for advance reservations. The article says that a concern is that people who book late may be unable to get on popular rides. That obviously is a problem, particularly if a park (like Disneyland in Southern California) gets a fair amount of local visitors. Even for those who plan early, it is hard to know what rides are most desirable. How do you tell a five-year old that they can't go on a ride they loved a second time because dad only booked one reservation for it? Disney has to limit the number of reservations: both to buffer for downtime and to accommodate spontaneity.

A second part of NextGen is the use of a wrist band embedded with RFID, that reads your identity and acts as your ticket. Disney is already experimenting with RFID technology, for example, at Epcot. But the NextGen wrist band concept is expected to go further. Its believed that guests would provide information  such as their names, credit card information and favorite attractions ahead of their arrival. After they enter at the park, the RFID would interact with sensors deployed throughout Disneys resorts and trigger interactive features. So for example, an attraction may greet you and your family and call you by name.

In effect, a Disney park would become a little more like a website, recording where youve been and choice youve made. Note that this could make running a reservation system a little easier. When the Jones party of four has not shown up for their 2:00 reservation by 2:05, it may be possible to see that they were on the other side of the park 2 minutes ago and dont have any chance of getting to the ride soon. Their space could then be given over to stand by customers.

This also a big data boon. Knowing ages of the family and whether this is their first trip to a park could allow Disney to suggest itineraries of age-appropriate rides the next time a family with a similar profile books.

The pivotal research finding in this area dates to the mid-50s in NYC. Complaints started to sore at that time as more and more people found themselves in high-rise buildings, waiting for elevators. It has to be the design of the buildings, so either dynamite the building and put in more elevators or tell people they have to wait. A consultant pointed the problem not to be the duration, but the complaints about the delay which they needed to reduce. The solution was to place floor-to-ceiling mirrors next to each elevator door and the complaints plummeted! How to get happy customers with lower costs at banks? A study on bank teller lines in 1990 at Bank of Boston (now BankBoston) for three weeks, each week the line operating in a different mode: Status quo; Silent Radio, Digital queue wait advisory. Silent Radio: placement of a Time Square type scrolling alphanumeric readout with live news, sports, weather and even advertisements for bank services. Silent with no disruption and easily seen if chosen to. Digital advisory: At the queue entrance, on a poll attached to one of the queue stanchions marking the wait lane, the readout would say Current Wait = 8 minutes (expensive device). Customers loved the Silent Radio so much that several regular customers complained to the bank manager the Monday following the removal of the display. They purchased and leased several also in front of busy congested ATMs. Digital Advisory actually reminded customers of the waiting. Customers repeatedly looked at their wrist watches, trying to play the game of Beat the Clock. In the 1980s the Savings Bank of New York employed not fancy technology but rather live piano recitals each day during the lunch hour. Visiting the bank was so appealing that an enterprising entrepreneur once sold tickets to sidewalk passersby just to get into the bank lobby. At a Toronto bank shows an eight minutes looped videotape to its customers.

## 5 $M/M/\infty$ Queues

Let $X(t)$ be the number of customers in the system at time $t$. $\{X(t), t \geq 0\}$ is a birth and death process with

$$\begin{aligned} \lambda_n &= \lambda, \qquad n \geq 0, \\ \mu_n &= n\mu, \qquad n \geq 0. \end{aligned}$$

Let $\rho = \frac{\lambda}{\mu}$. Solving the balance equations in terms of $P_0$ yields $P_n = \frac{\rho^n}{n!} P_0$ and $P_0 = e^{-\rho}$. So $P_n$ is Poisson with $\rho$ and $L = \rho$. $W = \frac{1}{\mu}$ and $W_Q = 0$.

# 6 Multiple stages

## 6.1 $M/M/1$ queues at each stage

The departure process of an $M/M/1$ queue. If a departing customer leaves behind an empty system with probability $P_0 = 1 - \rho$, the next departure will take place after a time equal in distribution to the sum of two independend exponential random variables with $\lambda$ and $\mu$. If a customer leaves behind at least one customer with probability $\rho$, the next departure will occur at an exponential amount of time with $\mu$. Hence,

$$\begin{aligned}
P(T<t) &= \rho P(T<t|N=0) + (1-\rho)P(T<t|N\geq 0) \\
&= \rho(1-e^{-\mu t}) + (1-\rho)\int_0^t P(X<t-y|Y=y)\lambda e^{-\lambda y}dy \\
&= \rho(1-e^{-\mu t}) + (1-\rho)\int_0^t (1-e^{-\mu(t-y)})\lambda e^{-\lambda y}dy \\
&= \rho(1-e^{-\mu t}) + (1-\rho)\left[\int_0^t e^{-\lambda t}d(\lambda t) - \lambda e^{-\mu t}\int_0^t e^{(\mu-\lambda)y}dy\right] \\
&= \rho(1-e^{-\mu t}) + (1-\rho)\left[e^{-\lambda t}\Big|_t^0 - \frac{\lambda}{\mu-\lambda}e^{-\mu t}\int_0^t e^{(\mu-\lambda)t}d(\mu-\lambda)y\right] \\
&= \rho(1-e^{-\mu t}) + (1-\rho)\left\{1 - e^{-\lambda t} - \frac{\rho}{1-\rho}e^{-\mu t}[e^{(\mu-\lambda)t}-1]\right\} \\
&= \rho(1-e^{-\mu t}) + (1-\rho)(1-e^{-\lambda t}) - \rho(e^{-\lambda t} - e^{-\mu t}) \\
&= 1 - e^{-\lambda t}.
\end{aligned}$$

So the departure time is exponential and each stage is an $M/M/1$ queue.

## 6.2 $G/G/c$ queues at each stage

### 6.2.1 $G/G/c$ queues

Let $C_a = \frac{\text{std of the interarrival time}}{\text{mean of the interarrival time}}$ and $C_s = \frac{\text{std of the service time}}{\text{mean of the service time}}$, the coefficient of variation. Then

$$L_Q \approx \frac{\rho^{\sqrt{2(1+c)}}}{1-\rho}\left(\frac{C_a^2 + C_s^2}{2}\right).$$

For the $M/M/c$ queue, $C_a = C_s = 1$. Then $W_Q = \frac{L_Q}{\lambda}$, $W = W_Q + \frac{1}{\mu}$, $L = \lambda W$, and $L_s = \frac{\lambda}{\mu}$.

### 6.2.2 Multi-stage, each with an $G/G/c$ queue

$$\begin{aligned}
C_d^2 &= 1 + (1-\rho^2)(C_a^2 - 1) + \frac{\rho^2}{\sqrt{c}}(C_s^2 - 1), \\
C_d^2 &= (1-\rho^2)C_a^2 + \rho^2 C_s^2, \quad \text{if } c = 1, \\
C_d^2 &= 1, \quad \text{if } c = 1 \text{ and } M/M/1 \text{ queues.}
\end{aligned}$$

**Insights:**

- Variability propogates.
- Non-bottleneck can cause major problems.
- Variability early in process has more impact on $W$ and $L$.