Improving the Customer Experience

Minimize Wait Time and Improve the Waiting Experience

Lavi Industries

www.lavi.com | (888) 285-8605

Overview

Waiting lines easily become the source of tension between customers and businesses and even cause the loss of revenue. Both research and experience suggest that a customer's evaluation of quality of service strongly depends on the time spent waiting in line: the longer the wait, the less the quality of service. Thus, minimizing the time a customer spends in line is crucial to the customer's perception of quality of service.

> "Waiting is frustrating, demoralizing, agonizing, aggravating, annoying, time consuming and incredibly expensive."
>
> - FedEx Advertisement

Waiting time can be perceived differently in different contexts: waiting five minutes for a waiter in a comfortable setting at a restaurant is acceptable while waiting for five minutes at a checkout counter for a "price-check" can be agonizing. In all cases, businesses must also be able to influence the customer's perception of the waiting experience.

This paper proposes the following measures a business can implement to improve the customer's experience while waiting in line:

- **Minimize the actual waiting time a customer spends in the queue**
- **Improve the perception of the waiting experience**

Minimizing the actual waiting time is achieved by using an efficient queuing configuration to reduce the variation in waiting time of the customer, hence improving the overall customer experience.

Improving the customer's waiting experience depends on a company's ability to influence customer perception of the length of time spent while in line. This can be achieved through in-line entertainment or advertising, information about the wait itself, and fairness of the wait.

When it comes to improving the customer experience, advantage is gained by having a **single-line multiple-server queue configuration.** This configuration is widely used in real-world scenarios, including:

- Retail
- Banking
- Entertainment
- Transportation

Many businesses invest in improving the waiting line experience:

- Amusement Parks
- Hospitality
- Big Box Retailers
- Healthcare
- Airports and Terminals

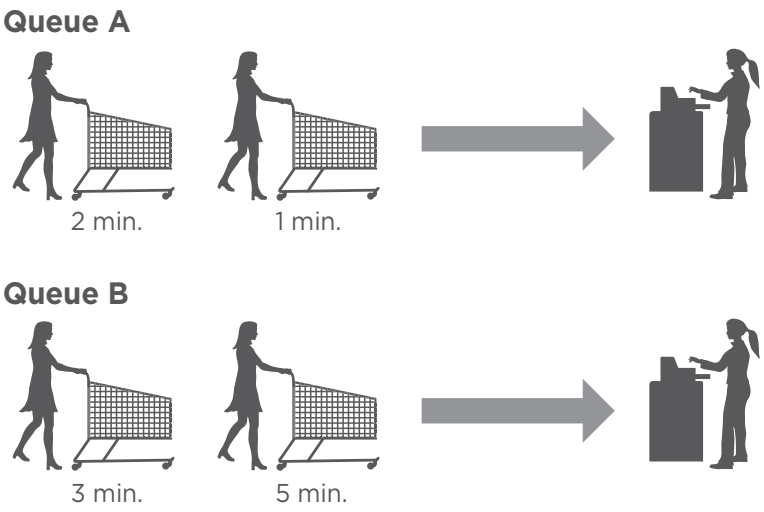
Minimizing Waiting Time

Recently, a nationwide retail chain introduced a single-queue multiple-server configuration (a single line of customers feeding into multiple checkout stations) at its urban stores. This choice was made over the traditional multiple-line, multiple server model (a line of customers at each checkout station) in order to reduce the uncomfortable amount of time customers spent standing in line.

Why is it more efficient to have one line of customers served by multiple stations than multiple lines of customers at each station? The answer lies in controlling the variation of time a customer might spend in a queuing line. In a multiple-queue, multiple-server configuration, time spent in a queue may vary greatly from customer to customer. A single-queue, multiple-server configuration reduces that variance.

Multiple-Queue, Multiple-Server Configuration

Consider a case of two checkout stations, each with its own line of customers, in a multiple-queue, multiple-server configuration (see illustration below):



In queue A, the first customer only has one item, and completing his transaction requires one minute. The next customer requires two minutes to complete checkout. In queue B, the customer closest to the server requires a price check (but the number of items in his cart is not excessive), and thus requires five minutes to complete checkout. The next customer in line requires three minutes to complete checkout.

In this configuration the least amount of time a customer will spend in line and at the checkout is one minute, but the maximum amount of time to complete checkout is eight minutes.

The probability of a customer spending eight minutes standing in line and at the checkout station is equivalent to the probability of the customer requiring three minutes for checkout when in the same queue as the customer requiring five minutes. That probability is ⅓ (because there are four customers and two queues, the chance of one particular customer being in a queue with another particular customer is ⅓) or 33.3%.

To illustrate further, we have charted each permutation into table 1.1. There are a total of 24 different permutations of the 2 queues. Additionally, we have highlighted the permutations that result in a maximum individual process time of 8 minutes. The Permutation Table clearly illustrates that 8 out of 24 permutations, or 33.3%, result in a maximum process time of 8 minutes.

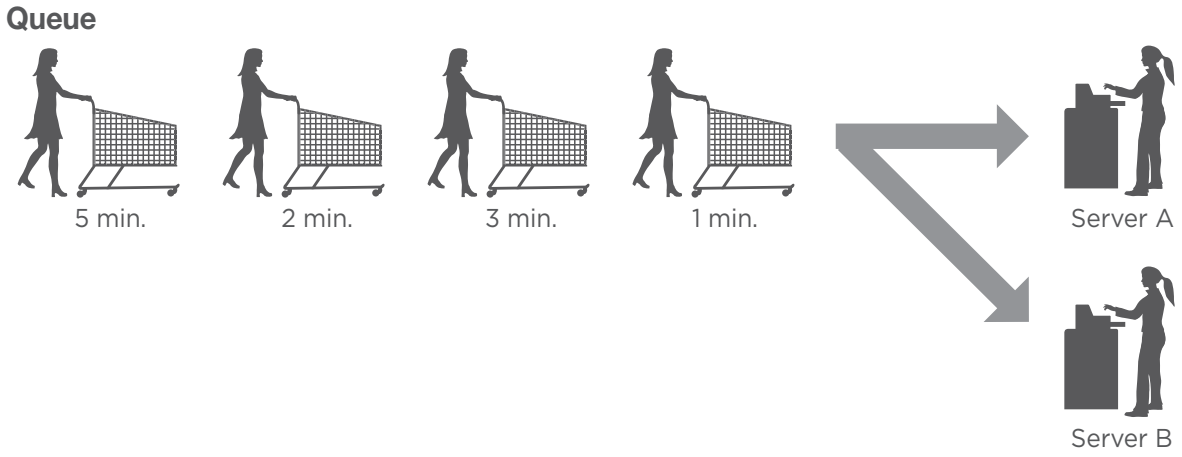
Multiple-Queue Multiple Server

1 2 3 5	2 1 3 5	3 1 5 2	5 1 2 3
1 2 5 3	2 1 5 3	3 1 2 5	5 1 3 2
1 3 2 5	2 3 5 1	3 2 5 1	5 2 3 1
1 3 5 2	2 3 1 5	3 2 1 5	5 2 1 3
1 5 3 2	2 5 3 1	3 5 1 2	5 3 1 2
1 5 2 3	2 5 1 3	3 5 2 1	5 3 2 1

Table 1.1

Single-Queue Multiple-Server Configuration

Consider a similar configuration, in which there is only one queue and two servers: a single-queue multiple-server configuration (see illustration below):

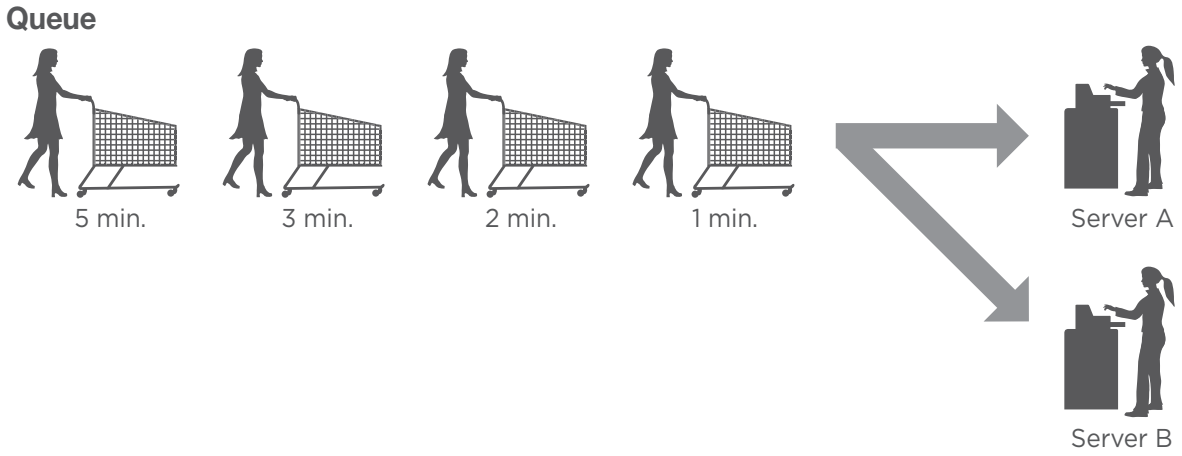


In this case, the minimum amount of time a customer spends in the queue and at the checkout is also one minute, and the maximum is eight minutes. However, a chance of any customer spending eight minutes in the system is significantly less than in a multiple queue – multiple server configuration.

The probability that the maximum waiting time is encountered is equivalent to the chance that the person requiring five minutes is the last in line and not next to the person requiring three minutes. The combined probability is $(\frac{1}{4})(\frac{2}{3}) = \frac{1}{6} = 0.167$ or 16.7%. In any other situation the maximum time a customer will spend in the system is less than eight minutes.

For example, in the situation depicted above, server A will help the customers requiring one or two minutes of processing time, server B will help the customer requiring three minutes of processing time, and the customer requiring five minutes can be helped either by server A or by server B. Thus, the total amount of time the last customer spends in the system is eight minutes. However, if the customer requiring three

minutes were in front of the customer requiring five minutes (see image below), the flow would be different: server A would help the customers requiring one and three minutes of processing time, and server B would help the customers requiring two and five minutes. In this case the maximum amount of time a customer spent in the system would be seven minutes.



In our example of the single-queue multiple-server model there are also 24 permutations of the queue. However, only 3 result in a maximum individual process time of 8 minutes. (See table 2.1) In this new example, the probability of a customer experiencing the maximum process time falls dramatically to just 16.7%.

Single-Queue Multiple Server

1 2 3 5	2 1 3 5	3 1 5 2	5 1 2 3
1 2 5 3	2 1 5 3	3 1 2 5	5 1 3 2
1 3 2 5	2 3 5 1	3 2 5 1	5 2 3 1
1 3 5 2	2 3 1 5	3 2 1 5	5 2 1 3
1 5 3 2	2 5 3 1	3 5 1 2	5 3 1 2
1 5 2 3	2 5 1 3	3 5 2 1	5 3 2 1

Table 2.1

Therefore, by changing the configuration of the queues and checkout stations, the chance of a customer standing in line for the maximum amount of time is reduced significantly (please see Appendix: Queuing Model Comparison for an in-depth discussion).

Summary

Introducing a single-queue multiple-server model reduces the variance of the total time a customer spends in the queue and at the checkout, so fewer customers will experience the maximum processing time. As such, average waiting times are dramatically decreased in a single-queue multiple-server model, resulting in a higher quality of service as perceived by the customer and providing an opportunity for better control of server utilization.

To further improve the customer experience, many retailers add express checkouts for customers with lightly loaded shopping carts.

The single-queue multiple-server model ensures that fewer customers will experience the maximum processing time.

Handling Jockeying

Another factor to be taken into consideration when comparing a multiple-queue, multiple-server configuration to single-queue multiple-server configuration is jockeying.

Customers who are impatient with the time spent in a queue may want to join another queue instead. Theoretically, such an approach adds to the efficiency of the multiple-queue, multiple-server configuration, but in practice customers seldom jockey more than once; and even once is enough to cause tension between customers and decrease the perception of service quality.

Improving the Perception of Service Quality

Although actual waiting time for a customer is reduced, the perceived waiting time may seem "forever." It is important to consider the psychological aspect of waiting to ease the customer's experience. A multitude of work has been published on the subject of wait line psychology. One of the most quoted experts on the subject, David Maister, coined the following propositions relative to our discussion[1]:

- Occupied time feels shorter than unoccupied time
- People want to get started
- Anxiety makes waiting seem longer
- Uncertain waits are longer than known finite waits
- Unexplained waiting is longer than explained waiting
- Unfair waiting is longer than equitable waiting

Occupied Time Feels Shorter than Unoccupied Time

When a customer is distracted from concentrating on the time passage itself, the time in line seems to pass more quickly. Television is the ultimate distraction, and having a television set with a show or a loop of commercials certainly eases the agony of waiting. Also, introducing impulse-buy products along the queuing line adds to customer occupation.

People Want to Get Started

A simple notification that the customer will be helped shortly alleviates anxiety because the servers have acknowledged the customer's presence. Restaurants employ this technique by having their hosts instruct customers that "someone will be with you shortly." The ability to fill out paperwork while in line, to unload the cart while another person is being served – any task that initiates a customer's transaction before the actual exchange begins can be effective in getting people started.

Anxiety Makes Waiting Seem Longer

Anxiety can arise from a customer's perception that he/she chose the "wrong" line, or that "the other line always moves faster." However, in a single-queue multiple-server configuration there are no "slow" lines and this fact reduces the anxiety level.

Uncertain Waits are Longer than Known Finite Waits

The sense of the unknown profoundly increases customer anxiety. Field studies have found that people perceive waiting as acceptable and even relaxing once they know how long the wait will last.

A valid concern exists over the customers' perception of a long line because single-line multiple-server configurations will usually have a line that seems far longer than individual lines feeding into multiple servers. The long lines might discourage customers. However, if the customer is assured that he/she won't spend more than, for example, three minutes in the line, then the wait is perceived as bearable.

Also, it is beneficial to overestimate the expected wait time rather than under estimate it because if the perceived duration of the wait exceeds the expected wait, frustrations will increase.[2]

Unfair Waits are Longer than Equitable Waits

Few situations are worse than knowing that someone who joined another queue later than you has completed the checkout before you. A single-queue multiple-server configuration ensures that customers will be served in a strict First In - First Out (FIFO) fashion.

The principles above are widely and successfully practiced by numerous businesses in order to ease the 'necessary evil' of being in a waiting line. When customers are respected and entertained, the wait may even become an enjoyable part of the shopping experience.

Conclusion

Waiting lines are inevitable. However, by implementing efficient queuing configurations and improving the perception of the waiting process, the quality of service and the profitably of the business can be positively influenced.

Appendix

Queuing Model Comparison

Queuing theory uses specific annotation to describe the models of waiting lines. The annotation consists of letters and numbers separated by "/". The models of interest in this discussion are:

- M/M/1
- M/M/*s*

The first letter denotes the statistical distribution of the customer arrival process. In the models listed above, "M" signifies that the arrival process is random and exponentially distributed. The second letter denotes the statistical distribution of the service process; "M" in this case again signifies that the service process is also random and exponentially distributed. The third letter or number displays how many servers exist in the model. In the first model, only one server is present, and in the second model, *s* is equal to the number of serving channels (an integer value of one or greater).

Additional parameters required for modeling are the arrival rate of the customers and the service rate of the servers. An essential requirement of the model is that the queue discipline is FIFO.

Once a standard model is formulated, a set of pre-derived performance measures can be calculated, including:

- Server utilization
- Time a customer spends in the system (standing in queue and the checkout)

Below are examples of M/M/1 and M/M/s queues which use pre-derived formulas for the results. For a complete discussion on the subject and the derivation of the formulas, please see Gross, Donald and Carl M. Harris, "Fundamental of Queuing Theory."[3]

Single-Server Model: Consider a store configuration where one line of customers feeds into one checkout station. The average arrival rate of customers to the waiting line is 30 customers per hour, and the average service rate of the checkout station is 40 customers per hour. In this case the system can be described as M/M/1 with the arrival rate = 30 and service rate = 40 (see fig.1).

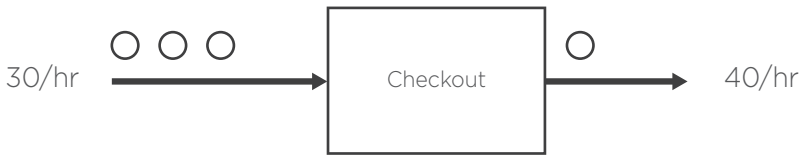


Fig. 1: Single Server Queue

In this model, the average time a customer spends standing in line and being serviced amounts to six minutes. (See Single Server Model section for calculations.)

Multiple-Server Model: To decrease the time a customer spends in line and at the checkout station, a new checkout station is added to the system. This configuration allows either two queues to be formed - one for each checkout station (see fig.2) - or a single queue feeding into both stations (see fig.3).

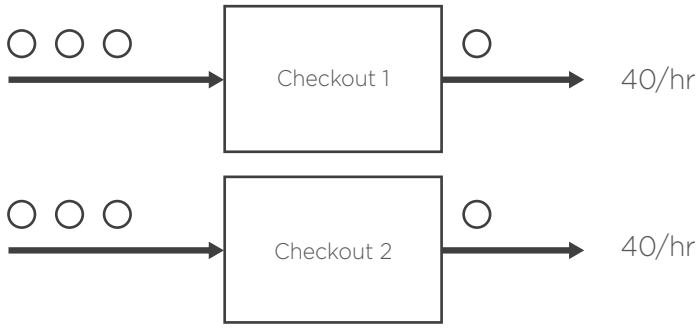


Fig. 2: Multiple Servers Multiple Queues

In the configuration with two queues feeding into two separate checkout stations, each queue will have an arrival rate of 15 (half the original rate of 30 customers per hour). The derived equations for this configuration reveal that a customer joining either of the two lines will, on average, spend **2.4 minutes** in the system. (See Multiple Queue – Multiple Servers Model section for calculations.) However, if the configuration is changed to a single queue feeding into both stations (fig.3), the behavior of the system improves. The customer will only spend, on average, **1.75 minutes** in the system. (See Single Queue – Multiple Servers section for calculations.)

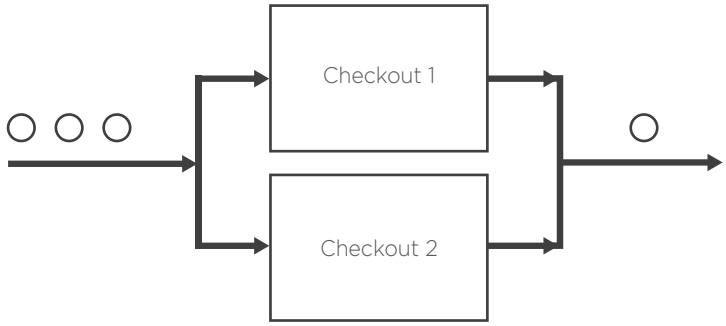


Fig. 3: Multiple Servers Single Queue

The reduction of time the customer spends in the system improves customer morale. Conversely, the increase in the customer flow rate provides an opportunity to maximize server utilization. The efficiencies of the single-queue multiple-server model allows businesses to achieve comparable customer flow and throughput to the multiple-queue multiple-server model using less servers.

Single Server Model

Model: M/M/1
λ: 30/hr
μ: 40/hr

Average time a customer spends in the system:

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{40 - 30} = 0.1 \text{ hrs} = 6 \text{ mins}$$

Server utilization: $\rho = \frac{\lambda}{\mu} = \frac{30}{40} = 0.75$

Annotation:

λ: arrival rate

μ: service rate

ρ: server utilization

s: number of servers

W_s: average time a customer spends in the system

L_q: average queue size a customer spends in the queue

p_0: probability of an empty system

Multiple Queue - Multiple Servers Model

Both queues are identical:

Model: M/M/1
λ: 15/hr
μ: 40/hr

Average time a customer spends in the system:

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{40 - 15} = 0.04 \text{ hrs} = 2.4 \text{ mins}$$

Server utilization: $\rho = \frac{\lambda}{\mu} = \frac{15}{40} = 0.375$

Single Queue - Multiple Servers Model

Both queues are identical:

Model: M/M/s
λ: 30/hr
μ: 40/hr
s: 2

Server utilization: $\rho = \frac{\lambda}{s\mu} = \frac{30}{80} = 0.375$

Probability of no customers in the queue:

$$p_0 = \left[\sum_{n=0}^{s-1} \left(\frac{(\frac{\lambda}{\mu})^n}{n!} \right) + \frac{(\frac{\lambda}{\mu})^s}{s!(1-\rho)} \right] -1 = \frac{1}{2.2}$$

Average length of the queue: $L_q = \frac{\rho(\frac{\lambda}{\mu})^s \, p_0}{s!(1-\rho)^2} = \frac{0.27}{2.2}$

Average time a customer spends in the system: $W_s = \frac{L_q}{\lambda} + \frac{1}{\mu} \approx 0.029 \text{ hrs} \approx 1.75 \text{ mins}$

References:

1. Maister, David. “The Psychology of Waiting Lines.” http://davidmaister.com/articles/5/52/

2. Houston, et all. “The Relationship Between Waiting in a Service Queue and Evaluation of Service Quality: A Field Theory Perspective.” Psychology and Marketing, Volume 15, Issue 8, p. 735-753

3. Gross, Donald and Carl M. Harris. “Fundamental of Queueing Theory.” Wiley Inter-Science, 1998