Performance Assessment of Physically-Constrained Fixed-Lane Queueing Systems

Simon Anthony Lorenzo
Assistant Professor
Department of Industrial Engineering and Operations Research
University of the Philippines Diliman
Metro Manila, Philippines
sdlorenzo@up.edu.ph

Lizabeth Ann Franco
Lecturer
Department of Industrial Engineering and Operations Research
University of the Philippines Diliman
Metro Manila, Philippines
ldfranco@up.edu.ph

Abstract

The study of queueing systems has been around for decades. Since, at its most basic level, queueing systems are considered as continuous-time Markov Chains, the study of queueing theory must adhere to numerous assumptions. Most of these assumptions deal with the distribution and behavior of customer arrivals and service times, as these are often illustrated as Poisson processes. Service prioritization, capacities, and customer behaviors such as balking and reneging are also often considered in model assumptions. Often not considered are the physical constraints associated with real-world queueing systems. In several real-world queues, customer movement is constrained by the available physical space, such as consecutive toll booths or grocery counters. There are multi-server queues with only one lane wherein a customer that has been served by the initial server cannot move out of the system if subsequent servers are still busy, thus rendering the initial server still occupied. In addition, servers located after the initial one cannot be accessed even if they are vacant if the initial server is still occupied. Due to this, the behavior of the system greatly differs from those of similar queueing systems where physical constraints are assumed to be non-existent. Through simulation, the study has evaluated that the performance of the physically-constrained queueing setup can lead to waiting times and queue lengths of up to 6 times greater than that of a traditional M/M/2 queue, especially for a relatively high level of customer arrivals. Improvements to these metrics are also measured upon the implementation of proposed corrective schemes that aim to ease the physical restrictions of queueing.

Keywords
Queueing Theory, Blocking, Simulation

1. Introduction

Queueing systems are characterized by arrival of entities that are in need of some type of servicing. Once these entities are served, they exit the system. The queue serves as a waiting area for the entities until they can transfer to a server. Queueing theory is the study of these systems. The Kendall-Lee notation is used in queueing theory to describe various aspects of a queue, such as customer arrival and service time distribution, queue capacity, number of servers, and service behavior (Kendall 1953). Two main metrics are used to assess queue performance: the average time spent by customers in the system W_s and the number of customers in the system L_s.

In the process of evaluating the behavior of queues, numerous assumptions are made. These assumptions often include (but are not limited to) a fixed distribution and magnitude for customer arrivals/service times, singular customer arrivals into the system, and absence of customer behavior such as jockeying, reneging and balking.

Another key assumption that is rarely addressed in theoretical computations in queueing theory is a customer's freedom to enter and exit a queueing system. It often goes without saying that customers can freely enter the queueing system whenever they want to (assuming the system is not full) and freely leave upon completion of service. However, there are real-world occurrences that hinder this ease to enter and exit the system, such as the physical layout of queues. It is not uncommon for two servers to be laid-out sequentially in a single-lane queueing system such as in toll booths. This setup was observed in a commercial establishment's parking lot where the payment toll booths were sequentially arranged. This is illustrated in the Figure 1 below.

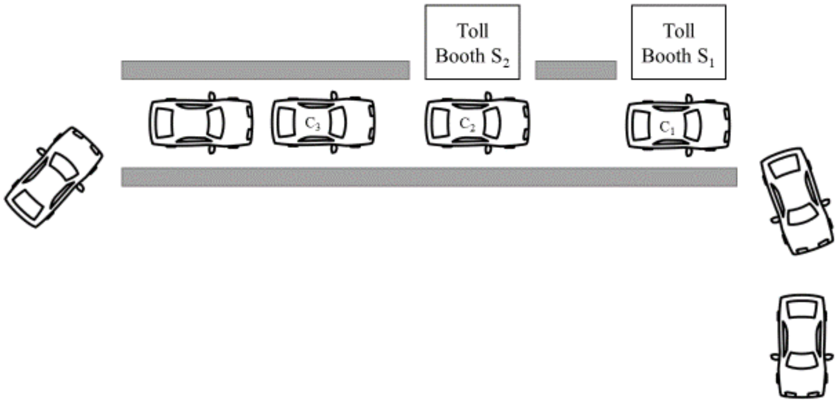

Figure 1. Sequential queueing setup with blocking.

As seen in Figure 1, the queue is physically-constrained by the exit lane. Ideally, the server in Toll Booth 1 (S_1) and Toll Booth 2 (S_2) can serve two customers (C_1 and C_2) simultaneously. However, a problem arises when service has been completed for C_2 but not yet for C_1, as C_2 cannot exit the system, making their time in the system longer. Moreover, S_2 becomes idle and cannot serve the next customer, lowering the server's utilization and making the succeeding customers' waiting times longer. Another problem arises with the opposite situation wherein service has been completed for C_1 but not yet for C_2. As seen in Figure 2, C_1 can readily exit the system, but S_1 becomes idle, and C_3 cannot skip ahead of C_2 and must wait until C_2 moves and exits the system. Similar to the previous scenario, this lowers the server's utilization and the succeeding customers' waiting times become longer. The situations described above fall under the phenomenon in queueing theory called *blocking*.

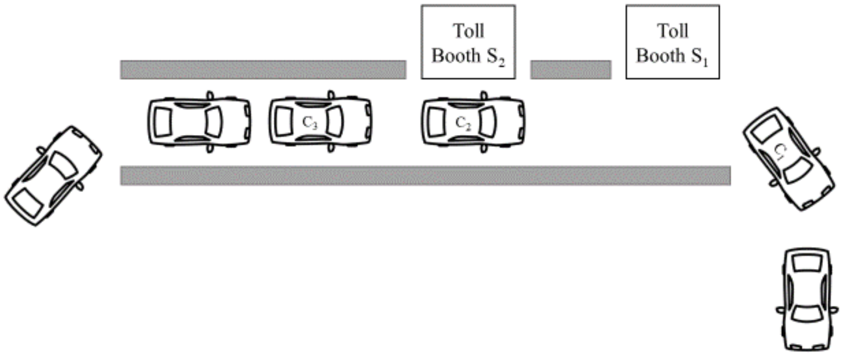

Figure 2. Illustration of blocking phenomenon.

While it is widely accepted that a single multiple-server queue outperforms multiple single-server parallel queues assuming identical performance for the servers (Do et al. 2015), the queueing system described above does not behave exactly like a multiple-server queue due to blocking caused by physical constraints. It is therefore worth investigating how the physically-constrained system performs with respect to two single-server queues and a single multi-server queue without blocking, and if there is an incentive to remove the physical constraints to improve the system's performance.

1.1 Objectives

As identified in the previous section, the study aims to investigate and assess the performance of physically-constrained two-server queues. Specifically, the objectives of the research are:

- To obtain metrics that would assess the queueing performance of physically-constrained two-server queues
- To compare the performance of the identified queue setup to similar two-server variations such as a traditional M/M/2 queue and two parallel M/M/1 queues

2. Literature Review

The study of queues has been around for a long time, with the first published study on queues being credited to (Erlang, 1909). Since then, several papers and textbooks have discussed the concepts involving queues, most often in the context of Poisson processes and Continuous-time Markov Chains (see Kendall (1953), Ross (2014), Stewart (2009), etc).

Studying queueing systems as basic Poisson processes makes various simplifying assumptions that make it, to a certain extent, non-descriptive of real-world queues. Phenomena such as jockeying, balking, reneging, batch processing, etc., are not captured in queueing theory. Thus, studies exploring these special cases arose following the onset of basic queueing theory. Studies on jockeying ((Koenigsberg 1966), (Zhao & Grassmann 1995), (Xu & Zhao 1996)), server psychology and behavior ((Shunko et al. 2015), (Do et al. 2015), (Do et al. 2017)), and balking ((Haight 1957), (Economou & Manou 2011), (Wang et al. 2014)), for example, have been abundant in the past decades.

Another special case in queueing, as identified and explained earlier, is blocking. Blocking is described as the phenomenon in a finite queueing network where the flow of customers through a node may be momentarily stopped when another node in the network reaches its capacity (Balsamo et al. 2001). Different mechanisms of blocking have also been described such as blocking after service and blocking before service. Blocking after service occurs when a customer upon completion of its service attempts to enter a destination node but it cannot because the destination is occupied. On the other hand, blocking before service occurs when a server is not served until the customer's destination node is available. The system described is unique in that blocking after service occurs for a customer who is done being served but is stuck behind the other customer being served and is blocking the exit of the system. Meanwhile, blocking before service occurs for a customer stuck in the queue behind a customer being served that is blocking the next available server.

Blocking in queueing systems is not a new concept. This phenomenon can be observed in computer systems (Konham and Reiser 1976), production systems (Foster and Perros 1980), and even in cafeteria-type queues (Weber and Weiss 1994). The setup of two sequential servers has been explored decades ago in the works of Avi-Itzhak & Yadin (1965), and the specific values for queueing system metrics were theoretically computed. Other studies have also explored the effect of blocking on two stations in a series with each station possibly having more than one server ((Neuts 1968), (Latouche & Neuts 1980), (Gomez-Corral 2004)). These studies, however, focused on a queues where the customer requires service from both stations before it can exit the system. These studies did not address the case where only one service instance is required. In addition, no comparison to other common queueing system setups was done.

3. Methods

3.1 Modelling Setup and Assumptions

To evaluate the performance of the queueing system being investigated, it is compared with two separate queueing systems, namely two single-server parallel queues and one two-server queue. To ensure the accuracy of the comparisons, the same assumptions are made about the three queueing systems. First, it is assumed that the arrival rate of customers and service rates of the two servers are identical. The simulator used also allows for the exact replication of the generated date for the simulations, which ensures an "apples to apples" or more accurate comparison of the systems. Second, the assumed customer behavior is that jockeying, balking, or reneging is not present.

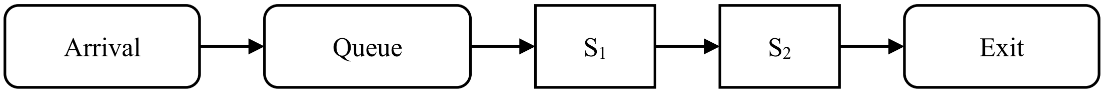

Figure 3. Physically-constrained queue behavior

Figure 3 illustrates the behavior of the physically-constrained system to be simulated. A customer would arrive and enter the queue if the system is not empty. If the system is empty, the customer proceeds to S_2 for service. If S_2 is occupied, the customer is served by S_1 and passes through the serving area of S_2 to be able to exit the system, provided that S_2 is no longer occupied. If S_1 is occupied but S_2 is not, the customer would remain in the queue until the service is completed for the customer at S_1 and moves to S2 then to the exit. This order of the servers is maintained for the succeeding discussions.

As previously mentioned, two queueing systems are used as the comparison for performance or the physically-constrained system. The first queueing system evaluated is two single-server parallel queues. For this scenario, it is assumed that when the customers arrive, they choose to line up in the queue with a shorter queue length. This setup is prevalent in the real world, in common queue locations such as groceries with individual counters and other commercial establishments such as fast food restaurants or public restrooms.

The next queueing system evaluated is a single two-server queue. This assumes that when customers arrive in the system, they enter the queue, and are served by the next available server. In general, multi-server queues are becoming more prevalent in real world systems, which can be attributed to better performance of the queue due to the pooling effect (Do et al. 2015). This is the system used in parking lots where there is one long queue of cars waiting for a slot to open up, for example.

3.2 Simulation Model

All three queueing systems are evaluated through a simulation model constructed using the MATLAB-based simulator Simulink. Only a single simulation model is created, where the three identified setups run in parallel using the exact same customer arrivals and service distribution times. To account for uncertainty of the behavior of both customers and servers, the study includes various combinations of distributions, values, and relationships for both customer arrivals and service distributions. The following Table 1 summarizes the variable input parameters used in the simulation:

Table 1. Simulation parameter values.

Customer Arrival Distribution		**Server Service Distribution**	
Distribution	**Value**	**Relative Values**	
Exponential Distribution	Low customer arrivals	Exponential Distribution	Server 1 Service Time > Server 2 Service Time
	Intermediate customer arrivals		Server 1 Service Time < Server 2 Service Time
	High customer arrivals		Server 1 Service Time = Server 2 Service Time

For customer arrival and service, a low, average, and high level of intensity is tested, to account for the variation of arrivals in queueing systems due to various factors such as peak and lean periods, etc. In addition, since the study assumes the same type of service provided by both servers, it would be acceptable to assume the same service distribution for the two. However, the study explores the case where the two servers have varying degrees of service delivery of the same service, which is not uncommon in the real world. Three scenarios are explored: one where the first server is slower than the second, one where the second server is slower than the first, and one where the service distributions are equal. Exhaustive combinations of these parameter sets were run, summing up to 3x3 = 9 simulation setups. Values for both customer inter-arrival time and service times are obtained from observations made on the toll booths at the parking lot of a mall. Service includes receiving a ticket from and providing payment to the toll booth personnel.

To assess each of the three queueing systems, two main and four supporting metrics are used. The average queue length (L) and average customer waiting time (W) are the main metrics for assessment, while the percent utilization of servers (U_1 and U_2), and number of customers served per server (n_1 and n_2) are also computed to gain some insights on the behavior of the identified queues.

4. Results and Discussion

4.1 Simulation Results

From the simulation runs, various metrics for the queueing setups can be obtained. Namely, the average customer waiting time (W) in seconds, average number of customers in the system (L), and server utilization. The following Table 2 compares these metrics for the physically-constrained setup and the standard M/M/2 setup:

Table 2. Comparison of metrics.

Customer Arrival Type	Service Relation Type	Physically-constrained setup		M/M/2 setup		2 x M/M/1 setup	
		L	W	L	W	L	W
High Arrival Rate	S_2 slower than S_1	4.413	93.657	0.625	13.010	0.548	27.769
High Arrival Rate	S_1 slower than S_2	4.065	86.334	0.730	15.185	0.851	37.715
High Arrival Rate	S_1 and S_2 equal	0.817	17.099	1.553	32.603	0.824	34.343
Intermediate Arrival Rate	S_2 slower than S_1	0.228	7.692	0.099	3.309	0.127	10.594
Intermediate Arrival Rate	S_1 slower than S_2	0.183	6.096	0.116	3.854	0.266	15.488
Intermediate Arrival Rate	S_1 and S_2 equal	0.131	4.420	0.204	6.813	0.259	16.276
Low Arrival Rate	S_2 slower than S_1	0.085	3.861	0.032	1.469	0.065	7.373
Low Arrival Rate	S_1 slower than S_2	0.066	3.008	0.051	2.341	0.184	12.784
Low Arrival Rate	S_1 and S_2 equal	0.046	2.109	0.053	2.405	0.108	9.635

From the results, it can be seen that for both L and W, the M/M/2 setup performs much better than both the physically-constrained and 2 x M/M/1 setup, as is somewhat expected. What is notable is that for high customer arrivals, the physically-constrained setup actually performs much worse than the 2 x M/M/1 setup. This is primarily due to the blocking phenomenon present in the system, which is exacerbated when a large number of customers are present in the singular system queue. The M/M/1 setup actually has a queue-correction step, albeit minor, in the form of queueing up on the shorter queue. For the intermediate customer arrival rate, however, we can see that the physically-constrained setup performs better than the 2 x M/M/1 setup. For the low customer arrival, there are instances where the physically-constrained setup performs better than the M/M/2 setup, but this is attributed to the blocking not occurring for a low number of customers in the system, thus the physically-constrained system approaches the behavior of the M/M/2 queue.

4.2 Other Discussion and Recommendations

Upon analysis of the physically-constrained queue and M/M/2, we can see that the latter is strictly better than the former. This is because blocking only serves as a constraint from a base M/M/2 setup, which only serves to retain or worsen its performance, by definition.

In addition to this, in the two-server, physically-constrained queue setup, its performance is best when the service rate is equal for both servers, which may be due to blocking being minimized. An interesting phenomenon occurs when there is a significant difference in the service times of the two servers. If the first server is faster than the second, the system generally performs worse than if the second server is faster than the first. One reason for this is that the server utilization in this setup is much higher for the second server, primarily because it is the first server to be chosen if both are available. If the slower server is utilized more than the faster one, the queueing system would obviously be worse off. Another reason is that a slower second server would lead to more instances of blocking when the first server is done and the second server is still serving. To avoid this scenario, it is recommended that the queueing system operator/manager implement the quick and easy solution of putting the faster server on the second position.

Based on the results, we can see that converting a physically-constrained queue to a standard M/M/2 setup drastically improves performance, especially for high customer arrivals. In real-world scenarios, if the space and financial considerations permit, it would be a very valid option to construct solutions to reach an M/M/2 setup, such as physical channels for free travel of customers within the queueing system.

It is important to note that for customers stuck with S_1 while waiting for S_2 to complete servicing, the simulation model is currently unable to capture the duration of blocking (i.e. time from service completion to S_2 being available). This time is also important to record as it contributes to the total time the customer spends in the system. On the other hand, for customers that are waiting behind S_1, who in an otherwise unblocked setup can be already served by S_2, their waiting times capture the time where they could have already been served by S_2. Being able to separate this time from the recorded waiting time will help provide insight into how much of the customer waiting time is solely due to the presence of the physical constraint.

5. Conclusion and Areas for Future Work

The study was able to compare a physically-constrained two-server queueing system to traditional 2 M/M/1 queues and a single M/M/2 queueing system. The difference was quantified and it was found that W and L are both significantly larger (up to 600% greater) with the physically-constrained setup compared to the M/M/2. This significant performance decrease should be put into consideration when designing physically-constrained queueing systems, and a simple alternative of creating a departure channel specifically for customers at server 1 would be of great benefit to system performance.

As future work, it is recommended that further analysis be done on the blocking time of the servers to provide more insight into the queueing system. This study could also be extended to scenarios with multiple servers beyond two servers.

References

Avi-Itzhak, B., & Yadin, M., A sequence of two servers with no intermediate queue. *Management Science*, vol. 11, no. 5, pp. 553–564, 1965.

Balsamo, S., De Nitto Persone, V., & Onvural, R., *Analysis of Queueing Networks with Blocking.* 1st Edition, Kluwer Academic Publishers, 2001.

Do, H., Shunko, M., Lucas, M., & Novak, D., *On the Pooling of Queues: How Server Behavior Affects Performance, 2015*. https://doi.org/10.13140/RG.2.1.2606.1688

Do, H., Shunko, M., Lucas, M., & Novak, D., Impact of behavioral factors on performance of multi-server queueing systems. *SSRN Electronic Journal, 2017*. https://doi.org/10.2139/ssrn.3080700

Economou, A., & Manou, A., *Equilibrium balking strategies for a clearing queueing system in alternating environment* (arXiv:1112.5555). arXiv, 2011. http://arxiv.org/abs/1112.5555

Erlang, A., The theory of probabilities and telephone conversations. *Nyt. Tidsskr. Mat. Ser. B*, *20*, 33–39, 1909.

Gomez-Corral, A., Sojourn times in a two-stage queueing network with blocking. *Naval Research Logistics*, vol. 51, no. 8, pp. 1068-1089, 2004.

Foster, F. G., & Perros, H. G., On the blocking process in queue networks. *European Journal of Operational Research*, vol. 5, no. 4, pp. 276–283, 1980.

Haight, F. A., Queueing with balking. *Biometrika,* vol. 44, no. 3/4, pp. 360-369, 1957.

Kendall, D. G., Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, vol. 24, no.3, pp.338–354, 1953.

Koenigsberg, E., On jockeying in queues. *Management Science*, vol. 12, no. 5, pp.412–436, 1966.

Konham, A., & Reiser, M., A queueing model with finite waiting room and blocking. *Journal of the ACM*, vol. 23, no. 2, pp. 328-41, 1976.

Latouche, G., & Neuts, M.F., Efficient algorithmic solutions to exponential tandem queues with blocking. *SIAM Journal on Algebraic Discrete Methods,* vol. 1, no. 1, pp. 93-106, 1980.

Neuts, M.F., Two queues in series with a finite, intermediate waiting room. *Journal of Applied Probability*, vol 5, pp. 123-142, 1968.

Ross, S. M., *Introduction to Probability Models*. Academic Press, 2014.

Shunko, M., Niederhoff, J., & Rosokha, Y., Humans are not machines: the behavioral impact of queueing design on service time. *Management Science*, vol. 64, no. 1, pp. 453-473, 2015.

Stewart, W. J., *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*. Princeton University Press, 2009.

Wang, Y., Guo, J., Ceder, A. (Avi), Currie, G., Dong, W., & Yuan, H., Waiting for public transport services: Queueing analysis with balking and reneging behaviors of impatient passengers. *Transportation Research Part B: Methodological*, vol. 63, pp. 53–76, 2014.

Weber, R.R., & Weiss, G., The cafeteria process-tandem queues with 0-1 dependent service times and the bowl shape phenomenon. *Operations Research,* vol. 42, no. 5, pp. 895-912, 1994.

Xu, S. H., & Zhao, Y. Q., Dynamic routing and jockeying controls in a two-station queueing system. *Advances in Applied Probability*, vol. *28*, no. 4, pp.1201–1226, 1996.

Zhao, Y., & Grassmann, W. K., Queueing analysis of a jockeying model. *Operations Research*, vol. 43, no. 3, pp. 520–529, 1995.

Biographies

Simon Anthony Lorenzo is an Assistant Professor of the Department of Industrial Engineering and Operations Research of UP Diliman. He obtained his B.S. Industrial Engineering and M.S. Industrial Engineering degree (with a specialization in Operations Research) from the University of the Philippines Diliman. He has been involved in consulting projects with banking and financial institutions, and has conducted training on analytics and statistics. His research focuses on optimization studies, particularly in network, queueing and simulation models.

Lizabeth Ann Franco is a Lecturer of the Department of Industrial Engineering and Operations Research of UP Diliman. She obtained her B.S. Industrial Engineering from the University of the Philippines Diliman and is currently completing her M.S. Industrial Engineering degree from the same university. She has been involved in consulting projects with banking and financial institutions, and companies in the fast-moving consumer goods industry. Her research focuses on production systems and process improvement.