

RESEARCH ARTICLE

Overlap times in the infinite server queue

Sergio Palomo¹ and Jamol Pender^{2*} 

¹ Systems Engineering, Cornell University, Ithaca, NY, USA

² School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA.

*Corresponding author. E-mail: jjp274@cornell.edu

Keywords: Applied probability, Queueing theory, Stochastic modeling

Abstract

Imagine, you enter a grocery store to buy food. How many people do you overlap with in this store? How much time do you overlap with each person in the store? In this paper, we answer these questions by studying the overlap times between customers in the infinite server queue. We compute in closed form the steady-state distribution of the overlap time between a pair of customers and the distribution of the number of customers that an arriving customer will overlap with. Finally, we define a residual process that counts the number of overlapping customers that overlap in the queue for at least δ time units and compute its distribution.

1. Introduction

Have you ever wondered how many people you overlap with in a store or retail shop? This is fascinating question on its own, but it is especially important given the current COVID-19 pandemic. Much of the work on COVID-19 has focused on using deterministic compartmentalized style models to estimate the infection rate and dynamics of the spread, see for example Dandekar *et al.* [2], Nguemdjo *et al.* [20], and Kaplan [14]. However, we know that stochastic effects can play an important role in determining the spread, see for example Drakopoulos *et al.* [9], Palomo *et al.* [22], Pang and Pardoux [23], Forien *et al.* [11], and Moein *et al.* [19]. As shoppers crowd a store to stock up on water or large amounts of non-perishable items, it is inevitable that the virus would spread. To combat the spread of the virus, many service facilities and systems have installed new air filters and transparent barriers, and have asked that patrons wear facial masks. Moreover, these service systems also have implemented various forms of social/physical distancing in order to minimize close proximity of one customer to another Bove and Benoit [1].

However, there are some places where people work or shop that limiting distance is not feasible. In this case, we really care about how much customers overlap with one another. Recently, there has been new work by Kang *et al.* [13] and Palomo and Pender [21] that explores how one can calculate the overlap times of customers in multi-server and single server queues, respectively. More specifically, Kang *et al.* [13] show how to use the overlaps to compute a new R_0 value for understanding infection rates in compartmentalized epidemic models and Palomo and Pender [21] prove that the overlap distribution is exponential for the M/M/1 queue and show via simulation that a similar result holds for the non-Markovian setting as well. Our analysis is important because it can demonstrate, exactly, how much overlap occurs and can provide distributional information or prediction intervals for possible overlap. Moreover, it can be used as a tool to prevent large overlaps and as a design tool to construct appropriate overlap by restricting the arrival rate or service distributions.

In addition to understanding lightly loaded systems with our infinite server queue overlap analysis, there are many other applications where studying overlap times is essential. The first application is from queues with advanced reservations, see for example [17]. Companies such as AirBnB, HomeAway, and

Vrbo all have systems where the arrival process is an incoming stream of customer booking requests from customers and an important object of study is the maximum number of overlapping customers. Moreover, in the context of spot welding in production lines, it is known that multiple welders trying to access the same power source is detrimental to the welding process. In order to understand the interference that is caused when multiple welders are using the same power source, it is important to derive the distribution of the maximum number of overlapping welders at any given instant in a simple closed network, see for example [26]. Thus, the study of overlap times is critical to developing a thorough understanding of these applications.

In this paper, we extend the overlap time analysis to the setting of the infinite server queue. At first glance, the infinite server queue analysis might not seem relevant, however, for service entities such as grocery stores, outlets, restaurants, and retail shops, an infinite server queue is quite relevant as there is not much waiting or the waiting to check out may be insignificant when compared with the shopping time. Moreover, the overlap times in the infinite server queue serve as a lower bound for the overlap times that a customer might experience in systems where there is significant waiting or in a multi-server setting. What also makes the infinite server queue important is that we are able to derive explicit formulas for the overlap distribution and residual overlap distributions as well as the number of people that a customer will overlap with during the duration of their service experience.

1.1. Contributions and organization of the paper

In Section 2, we describe the stochastic model that we will use in this work. We derive an equation for describing the overlap times for customers in the infinite server queue. We use this equation to compute the steady-state distribution of the overlap time of customers such that exactly $k - 1$ other customers arrived between their arrival times. In Section 3, we compute the mean and variance of the number of customers a customer will overlap with during their time in the queue. We also show how to compute a residual version of the overlap time where a customer must overlap at least δ units of time. Finally, in Section 4, we provide a conclusion and some future research directions.

2. Infinite server overlap times

In this section, we study the infinite server queue with the intention of understanding how much time consecutive customers spend in the system together. A similar type of analysis has been completed [13,21] albeit in Markovian systems.

2.1. Overlap time distribution

In this section, we consider the $GI/GI/\infty$ queue starting with 0 customers at time 0. Let A_i be the arrival time of the i th customer and we define the inter-arrival time between the i th and $(i + 1)$ th customers to be $A_{i+1} - A_i$, which are i.i.d. random variables with cumulative distribution function (cdf) $F(x)$. We also assume that S_i is the service time of the i th customer and the service times are i.i.d. with cdf $G(x)$. In the infinite server queue, by definition, no customer will wait. Thus, the departure time for the n th customer is given by the following equation

$$D_n = S_n + A_n.$$

We can now construct an equation for the overlap time between consecutive customers. The overlap time between the n th and $(n + k)$ th customers is given by

$$\begin{aligned} O_{n,n+k} &= (\min(D_n, D_{n+k}) - A_{n+k})^+ \\ &= (\min(A_n + S_n, A_{n+k} + S_{n+k}) - A_{n+k})^+ \\ &= ((S_n - (A_{n+k} - A_n))^+ \wedge S_{n+k}). \end{aligned} \tag{1}$$

It is important to observe that the overlap time between the n th and $(n + k)$ th customers can be decomposed into two parts. The first part (the left term in the minimum) is the time that the n th customer overlaps with the $(n + k)$ th customer given that the n th customer stays longer. The second part is the service time of the $(n + k)$ th customer if the service time of the $(n + k)$ th stay is shorter than the n th customer's service time minus the inter-arrival time gap. We will leverage this representation when considering the steady-state overlap time and the fact that all of the random variables are independent from one another in the $GI/GI/\infty$ queue.

Theorem 2.1. *Let O_k have the steady-state distribution of $O_{n,n+k}$ in the $GI/GI/\infty$ queue, and let S and $\tilde{S}$ be two independent service times with cdf $G(x)$, then the tail distribution of $O_k = \lim_{n \rightarrow \infty} O_{n,n+k}$ is given by*

$$\mathbb{P}(O_k > t) = \bar{G}(t) \int_0^\infty \bar{G}(t+x) h_k(x) dx,$$

where $h_k(x)$ is the density of the sum of k i.i.d. inter-arrival times.

Proof. First, we need to decompose the overlap probability into two probabilities by using a property of the minimum of two independent random variables, that is,

$$\begin{aligned} \mathbb{P}(O_k > t) &= \mathbb{P}(((S - \mathcal{A}_k)^+ \wedge \tilde{S}) > t) \\ &= \mathbb{P}((S - \mathcal{A}_k)^+ > t) \cdot \mathbb{P}(\tilde{S} > t) \\ &= \mathbb{P}((S - \mathcal{A}_k)^+ > t) \cdot \bar{G}(t) \\ &= \bar{G}(t) \int_0^\infty \mathbb{P}(S > t+x) h_k(x) dx \\ &= \bar{G}(t) \int_0^\infty \bar{G}(t+x) h_k(x) dx. \end{aligned}$$

This completes the proof. □

Corollary 2.2. *Let O_k have the steady-state distribution of $O_{n,n+k}$ in the $GI/M/\infty$ queue with service rate μ and let $\mathcal{A}_1$ be an inter-arrival time. Then, the tail distribution of $O_k = \lim_{n \rightarrow \infty} O_{n,n+k}$ is given by*

$$\mathbb{P}(O_k > t) = e^{-2\mu t} \mathbb{E}[e^{-\mu \mathcal{A}_1}]^k.$$

It is clear from this result that the functional form of the tail distribution of the overlap time is given by the service time distribution, which is exponential in this case. The inter-arrival times do not impact the decay rate of the tail distribution, which implies that it does not depend on the time parameter t . Moreover, the inter-arrival time distribution only emerges as a constant that determines how many of the overlap times are zero. Thus, the service time governs the tail behavior of the overlap time while the inter-arrival distribution controls the probability of a specific overlap time being equal to zero.

3. Computing the number of overlaps

In addition to knowing how much time consecutive customers will spend together in a service system, it is important to also know how many customers one expects to overlap with. In the context of epidemics, the more customers that one actually overlaps with the greater chance that one might contract the disease. In this section, we restrict our analysis to that of the $M_t/G/\infty$ queue and leverage results from Eick *et al.* [10] to compute the exact distribution of overlapping customers. In what follows, we assume without

loss of generality that all queues start with zero customers. Under this assumption, we know that the number in the $M_t/G/\infty$ queue at time t is given by the following expression

$$Q(t) = \sum_{j=1}^{N(t)} \{A_j < t < A_j + S_j\},$$

where $N(t)$ is the number of arrivals in the interval $(0, t]$ and we define $\{\mathcal{A}\}$ as an indicator function of the set $\mathcal{A}$. Thus, the number of customers in the system at the time of arrival of the k th customer is given by

$$Q(A_k-) = \sum_{j=1}^{N(A_k-)} \{A_j < A_k < A_j + S_j\}.$$

In order to compute the total number of overlapping customers, we also need to calculate the number of customers that arrive during the service time of the k th customer, which is

$$N(A_k + S_k) - N(A_k).$$

Adding the number of customers upon arrival and the number that arrive during service, we arrive at an expression for M_k , the total number of customers that the k th customer will overlap with.

$$M_k = N(A_k + S_k) - N(A_k) + \sum_{j=1}^{N(A_k-)} \{A_j < A_k < A_j + S_j\}. \quad (2)$$

Theorem 3.1. *The mean number of overlaps for the k th arrival in the $M/G/\infty$ queue is equal to*

$$\mathbb{E}[M_k] = \lambda \mathbb{E}[S] + \sum_{j=1}^{k-1} \frac{\lambda^{k-j}}{\Gamma(k-j)} \int_0^\infty \bar{G}(x) e^{-\lambda x} x^{k-j-1} dx.$$

Proof.

$$\begin{aligned} \mathbb{E}[M_k] &= \mathbb{E}[N(A_k + S_k) - N(A_k)] + \mathbb{E}\left[\sum_{j=1}^{N(A_k-)} \{A_j < A_k < A_j + S_j\}\right] \\ &= \lambda \mathbb{E}[S] + \mathbb{E}\left[\sum_{j=1}^{k-1} \{A_j < A_k < A_j + S_j\}\right] \\ &= \lambda \mathbb{E}[S] + \sum_{j=1}^{k-1} \mathbb{E}[\bar{G}(A_k - A_j)] \\ &= \lambda \mathbb{E}[S] + \sum_{j=1}^{k-1} \frac{\lambda^{k-j}}{\Gamma(k-j)} \int_0^\infty \bar{G}(x) e^{-\lambda x} x^{k-j-1} dx. \end{aligned}$$

This completes the proof. □

Corollary 3.2. *When the service distribution is given by an exponential with rate μ , we have that the mean number of overlaps for the k th arrival is equal to*

$$\mathbb{E}[M_k] = \frac{\lambda}{\mu} + \frac{\lambda}{\mu} \cdot \left(1 - \left(\frac{\lambda}{\lambda + \mu}\right)^{k-1}\right).$$

It is important to note that the total number of overlapping customers for the k th arrival can be only computed when the k th arrival departs the queue and it is not known at the time of arrival. Thus, the representation of Eq. (2) for the number of overlaps provides a methodology for computing the number of overlaps for each customer via simulation. However, this representation is customer-centered and not time-centered. In what follows, we provide a time-centered perspective of the number of overlaps.

If a customer arrives at time t , then we define $O(t)$ as the number of customers that the arriving customer will overlap with. $O(t)$ has the following expression

$$O(t) = \underbrace{N(t+S) - N(t)}_{\text{\#of customer arrivals during service}} + \underbrace{Q(t)}_{\text{queue length at time } t}$$

where S is a generic service time, which is independent of the arrival process.

It is important to note here that the number of overlapping customers $O(t)$ is not completely known at time t . It is only fully known after the service of the customer that arrives at time t . However, we can use the expression of $O(t)$ to compute the distribution of the number of customers that an arrival will overlap with at time t . In this way, it is time-centered as the expression depends on time and not the specific number of the customer arriving. In addition to knowing the exact number of customers, an arrival at time t would overlap with for each sample path, we can also describe the distribution of the number of customers that an arrival at time t would overlap. One important ingredient to describing the exact distribution is knowing that the number of customers upon arrival is independent of the number of arrivals that arrive during service. We provide the exact distribution in the following theorem.

Theorem 3.3. *The distribution of overlapping customers for a customers arriving at time t in the $M_t/G/\infty$ queue is*

$$O(t) \stackrel{D}{=} \text{Poisson} \left(\int_t^{t+S} \lambda(s) ds + \int_0^t \lambda(u) \bar{G}(t-u) du \right).$$

This result follows easily from the representation given in Eq. (2).

Corollary 3.4. *The overlap distribution at time t in the $M/M/\infty$ queue is equal to*

$$\mathbb{P}(O(t) = k) = \rho^k (1 - \rho) e^{(\mu/\lambda)q(t)} \frac{\Gamma(k+1, \frac{\lambda+\mu}{\lambda}q(t))}{\Gamma(k+1)}$$

and $O(t)$ can be decomposed into a sum of geometric and Poisson random variables, that is,

$$O(t) \stackrel{D}{=} \text{Geometric} \left(\frac{\lambda}{\lambda + \mu} \right) + \text{Poisson}(q(t)),$$

where $q(t) = \lambda \int_0^t \bar{G}(t-u) du$.

In addition to understanding the distribution of the number of overlapping customers at any point in time, one might be interested in counting the number of customers that overlap by at least δ amount of time. In this case, we can define the thinned arrival process counting only the customers with service times that exceed δ so $\lambda_\delta(t) = \lambda(t) \cdot \bar{G}(\delta)$ and $\bar{G}_\delta(x) = G(x)/G(\delta)$ for $x \geq \delta$. Then, the thinned process $O(t, \delta)$ has the same distribution as Eq. (3.4) where S is replaced by $(S - \delta)^+$, λ is replaced by λ_δ , and $G(x)$ is replaced by $G_\delta(x)$.

4. Conclusion and future work

In this paper, we consider the overlap times for customers in an infinite server queue. The infinite server model is appropriate in retail settings where the time a customer waits is small relative to their shopping experience. We derive the steady-state distribution for the overlap time of customers that are k arrivals apart. We also compute explicitly the distribution of the number of customers that a randomly arriving customer will overlap with as the sum of a Poisson random variable and Poisson with a random arrival rate. Finally, we compute an expression for the number of customers that a random arrival will overlap at least δ time units. We compute the mean and variance for this quantity and are able to provide a prediction interval for a randomly arriving customer at any time. Our analysis has implications for understanding the interaction time between customers in a pandemic setting and sheds light on the interactions of customers.

Despite our analysis, there are many avenues for additional research. First, we would like to complete our analysis by analyzing the $G/G/\infty$ queue in explicit detail. Issues like dependent arrivals or service times like in Pang and Whitt [24,25], Daw and Pender [5,6], Koops *et al.* [15,16], and Daw *et al.* [3] would be interesting to explore as well. Second, we would like to extend our analysis to more complicated queueing systems like the Erlang-A, see for example Daw and Pender [7], Massey and Pender [18], and Hampshire *et al.* [12] where customers can abandon the system. Abandoning customers clearly reduces the number of overlapping customers, but by how much? It would also be great to extend our analysis to queueing systems with batch arrivals. In this case, the scaled Poisson decomposition of Daw and Pender [8] and Daw *et al.* [4] might be helpful in replicating our analysis in the batch setting. Finally, we would like to extend our analysis to spatial point processes and think about the overlap in terms of not only time, number, but spatial distance as well.

We are also interested in knowing the overlap distribution for more than two customers. For example, if one considers the overlap time of three customers $(n, n+j, n+k)$ where $1 \leq j < k$, then one obtains the following overlap time for the n th, $(n+j)$ th, and the $(n+k)$ th customers as

$$\begin{aligned} O_{n,n+j,n+k} &= (\min(D_n, D_{n+j}, D_{n+k}) - A_{n+k})^+ \\ &= (\min(A_n + S_n, A_{n+j} + S_{n+j}, A_{n+k} + S_{n+k}) - A_{n+k})^+ \\ &= (S_n + A_n - A_{n+k})^+ \wedge (S_{n+j} + A_{n+j} - A_{n+k})^+ \wedge S_{n+k} \\ &= (S_n + A_n - A_{n+k})^+ \wedge (S_{n+j} + A_n + (A_{n+j} - A_n) - A_{n+k})^+ \wedge S_{n+k}. \end{aligned}$$

Unlike the two customer situation, it is clear here that the random variables are not independent anymore. This presents a new issue that must be resolved in future work.

Acknowledgments. J.P. would like to acknowledge the gracious support of the National Science Foundation DMS Award # 2206286. S.P. would like to acknowledge the gracious support of the Sloan Foundation for supporting his graduate studies.

References

- [1] Bove, L.L. & Benoit, S. (2020). Restrict, clean and protect: signaling consumer safety during the pandemic and beyond. *Journal of Service Management*.
- [2] Dandekar, R., Henderson, S.G., Jansen, H.M., McDonald, J., Moka, S., Nazarathy, Y., Rackauckas, C., Taylor, P.G., & Vuorinen, A. (2021). Safe blues: The case for virtual safe virus spread in the long-term fight against epidemics. *Patterns* 2(3): 100220.
- [3] Daw, A., Castellanos, A., Yom-Tov, G.B., Pender, J., & Gruendlinger, L. (2020). The co-production of service: Modeling service times in contact centers using Hawkes processes. *arXiv preprint arXiv:2004.07861*.
- [4] Daw, A., Fralix, B., & Pender, J. (2020). Non-stationary queues with batch arrivals. *arXiv preprint arXiv:2008.00625*.
- [5] Daw, A. & Pender, J. (2018). Exact simulation of the queue-Hawkes process. In *Proceedings of the 2018 Winter Simulation Conference*, pp. 4234–4235.
- [6] Daw, A. & Pender, J. (2018). Queues driven by Hawkes processes. *Stochastic Systems* 8(3): 192–229.
- [7] Daw, A. & Pender, J. (2019). New perspectives on the Erlang-A queue. *Advances in Applied Probability* 51(1): 268–299.
- [8] Daw, A. & Pender, J. (2019). On the distributions of infinite server queues with batch arrivals. *Queueing Systems* 91(3): 367–401.

- [9] Drakopoulos, K., Ozdaglar, A., & Tsitsiklis, J.N. (2017). When is a network epidemic hard to eliminate?. *Mathematics of Operations Research* 42(1): 1–14.
- [10] Eick, S.G., Massey, W.A., & Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research* 41(4): 731–742.
- [11] Forien, R., Pang, G., & Pardoux, É. (2020). Epidemic models with varying infectiosity. *arXiv preprint arXiv:2006.15377*.
- [12] Hampshire, R.C., Bao, S., Lasecki, W.S., Daw, A., & Pender, J. (2020). Beyond safety drivers: Applying air traffic control principles to support the deployment of driverless vehicles. *PLoS ONE* 15(5): e0232837.
- [13] Kang, K., Doroudi, S., Delasay, M., & Wickeham, A. (2021). A queueing-theoretic framework for evaluating transmission risks in service facilities during a pandemic. *arXiv preprint arXiv:2103.13441*.
- [14] Kaplan, E.H. (2020). OM forum—COVID-19 scratch models to support local decisions. *Manufacturing & Service Operations Management* 22(4): 645–655.
- [15] Koops, D.T., Boxma, O.J., & Mandjes, M.R.H. (2017). Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities. *Queueing Systems* 86(3): 301–325.
- [16] Koops, D.T., Saxena, M., Boxma, O.J., & Mandjes, M. (2018). Infinite-server queues with Hawkes input. *Journal of Applied Probability* 55(3): 920–943.
- [17] Maillardet, R.J. & Taylor, P.G. (2016). Queues with advanced reservations: An infinite-server proxy for the bookings diary. *Advances in Applied Probability* 48(1): 13–31.
- [18] Massey, W.A. & Pender, J. (2018). Dynamic rate Erlang-A queues. *Queueing Systems* 89(1): 127–164.
- [19] Moein, S., Nickaeen, N., Roointan, A., Borhani, N., Heidary, Z., Javanmard, S.H., Ghaisari, J., & Gheisari, Y. (2021). Inefficiency of sir models in forecasting COVID-19 epidemic: A case study of Isfahan. *Scientific Reports* 11(1): 1–9.
- [20] Nguemdjo, U., Meno, F., Dongfack, A., & Ventelou, B. (2020). Simulating the progression of the COVID-19 disease in cameroon using sir models. *PLoS ONE* 15(8): e0237832.
- [21] Palomo, S. & Pender, J. Measuring the overlap with other customers in the single server queue. In K.H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, & R. Thiesing (eds), *Submitted to the Proceedings of the 2021 Winter Simulation Conference*.
- [22] Palomo, S., Pender, J., Massey, W., & Hampshire, R.C. (2020). Flattening the curve: Insights from queueing theory. *arXiv preprint arXiv:2004.09645*.
- [23] Pang, G. & Pardoux, É. (2020). Functional limit theorems for non-Markovian epidemic models. *arXiv preprint arXiv:2003.03249*.
- [24] Pang, G. & Whitt, W. (2012). The impact of dependent service times on large-scale service systems. *Manufacturing & Service Operations Management* 14(2): 262–278.
- [25] Pang, G. & Whitt, W. (2013). Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems* 73(2): 119–146.
- [26] Rumsewicz, M. & Taylor, P. (1988). A spot welding reliability problem. *The ANZIAM Journal* 29(3): 257–265.