Queueing Systems (2022) 100:273–275
https://doi.org/10.1007/s11134-022-09745-y

Queues with no-show customers

Serhan Ziya[1]

Received: 24 January 2022 / Accepted: 28 February 2022 / Published online: 15 March 2022

1 Introduction

There are different ways of modeling appointments made for a service facility. One useful way is to view them as jobs in a queue. This may seem unnatural at first. After all, an appointment implies a commitment made by the service provider to serve a particular customer at some specific time in the future, and if service times are stochastic it is not possible to know, at the time of the "arrival of an appointment," when exactly the service associated with that appointment will start. However, one can assume deterministic service times, in which case customers would be told exactly when their services will start at the time they make their appointments, or allow stochastic service times, but assume that customers somehow show up at their random appointment times (see, e.g., [1]). One particular approach would be by viewing the appointment queue as a time-slotted batch-service system where a single service period in the model corresponds to a predetermined time period, say a single day, during which multiple customers are served (see, e.g., [2]). For example, such a model would be a good fit for appointments made for a healthcare clinic, which has the capacity to see some fixed number of patients on each day. In this system, batch size would correspond to the daily capacity of the clinic and each service period would last exactly one day. In this model, patients can be told exactly which day their services would take place at the time they schedule their appointments (i.e., the appointment's arrival at the queue). They could presumably be also told the time of their appointment since the queueing model does not need to capture the times of individual services within one service period.

One complicating issue, however, is that, in practice, many service systems that work with appointments suffer from customer no-shows. A no-show is like an abandonment, but traditional queueing models with customer abandonments do not capture the no-show phenomenon well. This is because in these models when a customer abandons their position in the queue that position is filled with others, whereas in the case of no-show models, there are no abandonments from the appointment queue. All

✉ Serhan Ziya
ziya@unc.edu

[1] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC, USA

appointments stay in the queue until their scheduled times and whether or not they are no-show appointments is revealed only then. When an appointment is a no-show, the time allocated for that appointment is essentially wasted. If one assumes that each customer is a no-show with some fixed probability, independently of other customers as well as other queueing dynamics, then the analysis of this queue would not be too different from those of the standard models. However, empirical studies suggest that such an independence assumption would be questionable. For example, several papers reported that a customer is more likely to not show-up for their appointment when their appointment delay (e.g., the time spent in the appointment queue) is longer (see, e.g., [1]). Incorporating such dependence without rendering mathematical analysis impossible appears to be a significant challenge that is in need of queueing theorists' attention. In the next section, we formally describe a queueing model to layout the basic research question more clearly, but future work might consider alternative, potentially more promising modeling approaches that still capture the same no-show dynamics. (Note that our focus here is the long-run, primarily steady-state, analysis of appointment queues with infinite stream of arrivals, not transient dynamics that are relevant in models that deal with scheduling of finite number of appointments over a finite horizon.)

2 Problem statement

Suppose that customers schedule new appointments according to a Poisson process with rate λ. We assume that the service provider always offers the next appointment slot available and customers always accept the appointment time offered to them. Therefore, new appointments join the appointment queue from the back of the queue and service is performed in a First-Come-First-Served manner. The daily service capacity is K ($1 \leq K < \infty$), i.e., on each day there are K appointment slots available. The queue is modeled as a time-slotted batch-service queue where each service period corresponds to a single day during which at most K customers are served. Number of customers served on a given day would be less than K if there are fewer than K appointments in the queue or some of the K appointments end up being no-shows.

We assume that the probability that a customer will be a no-show depends on how many days the customer will need to wait for their appointment to arrive, i.e., the appointment's waiting time in the queue before service, which we call the *appointment delay*. However, given this appointment delay whether or not the customer will show-up is assumed to be independent of the behavior of other customers as well as anything else in the system. Let p_i ($0 \leq p_i \leq 1$) denote the probability that a customer with an appointment delay of i days will show-up for their appointment. Suppose that at the time a customer calls to schedule an appointment, there are j appointments in the queue (excluding any appointments in the batch service at the time of arrival). Then, the appointment delay for the customer is $\lceil j/K \rceil$ days, where $\lceil z \rceil$ is the smallest integer that is greater than z for any $z \in \mathbb{R}$. When a customer does not show up for an appointment, the customer calls to make another appointment at the end of the day with some independent fixed probability q ($0 \leq q \leq 1$). Let X_n denote the number of appointments in the queue at the beginning of day n including appointments for day n,

A_n denote the number of new appointment calls (excluding rescheduled appointments due to no-shows) received during day n, and R_n denote the number of appointments rescheduled due to no-shows on day n. Then,

$$X_{n+1} = X_n + A_n + R_n - \min(X_n, K). \tag{1}$$

For this queue, it would be of interest to develop closed-form expressions or numerical methods for computing throughput, i.e., number of customers who receive service per day in the long-run, long-run fraction of customers who show-up for their appointments, expected length of the appointment queue in steady state, etc.

3 Discussion

The main difficulty in analyzing the queue described in Sect. 2 is due to the term R_n in (1). Precise characterization of this random variable is possible if one knows the no-show probabilities for the $\min(X_n, K)$ number of customers, who are served on day n, and that would necessitate knowing the appointment delays for those customers. (If $q = 0$, R_n disappears, and this simplifies the analysis substantially allowing computation of some of the key performance measures like throughput. This is because in this model, appointment delays are known with certainty and as a result no-show probabilities can be computed at the time of arrivals. In a model with stochastic service times and possibly non-batch service, no-show probabilities cannot be determined at the time of arrivals.) One could enlarge the state description and keep track of the no-show probability for each appointment (which can be computed at the time of arrival) along with the number of appointments in the system, but it is not clear if the analysis of such a complex model will lead to useful results. One approach used first by [1] is using the queue size at the time of service as a proxy for the queue size at the time the appointment joined the queue, i.e., letting the no-show probability for a customer not depend on that customer's appointment delay but instead on the delay that would be associated with a hypothetical appointment that joins the queue at the time of the customer's appointment. It is reasonable to expect that this would be a good approximation, but to the best of our knowledge, there has not been any mathematical work that investigates this question. More broadly, queueing analysis of appointment-based service systems, particularly by building novel models that are more amenable to mathematical analysis, developing alternative approaches that lead to exact results, or devising new approximations with performance guarantees is an interesting, understudied research direction.

References

1. Green, L.V., Savin, S.: Reducing delays for medical appointments: a queueing approach. Oper. Res. **56**(6), 1526–1538 (2008)
2. Izady, N.: Appointment capacity planning in specialty clinics: a queueing approach. Oper. Res. **63**(4), 916–930 (2015)