Signaled Queueing

Laura Brink
Yale University
New Haven, USA
laura.brink@yale.edu

Robert Shorten
IBM Ireland Research Lab
Dublin, Ireland
robshort@ie.ibm.com

Jia Yuan Yu
IBM Ireland Research Lab
Dublin, Ireland
jiayuanyu@ie.ibm.com

ABSTRACT

Burstiness in queues where customers arrive independently leads to rush periods when wait times are long. We propose a simple signaling scheme to decrease wait times by distributing customer arrivals more uniformly. Agents receive one of several signals with suggestions on what time to join the queue. We quantify the efficiency gains, both analytically and empirically, with respect to a number of parameters of the proposed signaled queue, such as burstiness of arrivals, number of distinct signals, and propensity of customers to follow suggestions.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

General Terms

Management; Design; Performance

Keywords

Queueing; Signaling; Social welfare

1. INTRODUCTION

Queueing problems are encountered throughout multiagent systems, operations research, economics, management, and telecommunications. Many queueing models have been proposed and analyzed in the literature. Some areas that have been examined so far are the equilibrium pattern of customer and operator's interactions [13], the pattern that customers abandon queues based on wait times [5], and applying games models to dynamic queueing problems [3]. While the literature on queueing theory is very large, the use of signals to decrease wait times appears to be under explored.

Our starting point is the observation that while queues are found in many places in every-day life, not all queues should behave in the same way. Most queues operate according to a simple FIFO principle. Customers that arrive first are served first. Example of this type of queueing can be found at check-in in airports, in the grocery store, and in the doctor's surgery. In each case, there exists a scope for applying signaling to improve the queue performance. First, in some situations, the FIFO (first-in first-out) paradigm is not optimal. Although operators have more information than customers, they do not take advantage of it, e.g., airlines know when each upcoming flight will leave. In other situations, the use of FIFO queues is simply not appropriate. For example, in the doctor's surgery, one should prioritize patients according need (or likelihood of infecting other patients, or of being infected by others). Such prioritization can be implemented via a signaling scheme. Finally, even in the FIFO queue, bursty arrivals can lead to poor queue performance. For example, at airports, customers arrive in bursts, i.e., the arrival rate fluctuates highly over time. These bursts cause long wait times because many people are entering the queue at the same time. A signaling system can also be used at an airport to make wait times shorter. When customers check-in online for their flights, they receive a signal suggesting them what time to arrive at the queue, along with a conditional reward as incentive to comply. The signals will redistribute customers arrival times such that arrivals are spaced out and arrival bursts do not occur. This approach can be generalized to any system that has a method to deliver signals to the customers, flexibility in customer arrival times, and bursts of customer arrivals. To encourage customers to act as intended by the signals, incentives based on the difference between their signals and their actual arrival times can be used.

In this paper, we start from a multiagent FIFO queueing system (Section 3). We then augment the FIFO queue with a signaling scheme and a corresponding model of customer-response (Section 4). The operator sends a distinct signal to each customer. Each signal is simply a set of suggested arrival times. This is done only once before the arrival of the first customer. In turn, each customer can slightly modify or modulate its arrival time according to the received signal. To make our analysis tractable, we make an important behavioral assumption on how customers modulate their arrival times in response to the signals received. The modulated arrival time is a convex combination of the nominal arrival time (when there is no signaling) and the closest suggested time in the customer's signal set. Although this behavior is simplistic, it does reflect the reality of human response to signals [4] (and can be modified accordingly to include more complex behaviors). By its linearity, it is even more reflective of machine response.

Our proposed signaling scheme works in many queueing settings, as long as these settings share the following characteristics. A producer of a resource (e.g., a commodity or

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May, 4–8, 2015, Istanbul, Turkey.*

service), a number of consumers. First, the consumers communicate an intent to consume the resource (e.g., prepayment, registration), the producer communicates signals to the consumers along with incentives for following these signals, the consumers choose when to consume the resource— We assume that customers have some flexibility as to when they arrive. Our signaling solution is especially useful when the environment is rapidly changing, i.e., where agents do not have sufficient time to learn or adapt their arrival times to those of other agents. Moreover, just as signaling reduces wait times in queues, it is easy to see that signaling can also reduce peak-demand in power networks and congestion in road networks.

Through both probabilistic analysis (Section 4.2) and simulation (Section 5), we quantify the performance improvement of the signaled queue as a function of a number of parameters of the signaling scheme, of the arrival pattern of customers, and of the responsiveness of customers to the signals. In many cases, the performance improvement is significant.

Example 1 (Airport). *Our signaling solution can be applied to an airport check-in setting. Upon booking a flight, each customer receives an additional signal. This signal indicates a number of suggested arrival time at check-in. For example, a customer may receive the signal "15" and have the option of checking out at 10:15am, 11:15am, etc. When the customer arrives at check-in, he receives a reward based on how close he arrived to one of the suggested times of his signal. The customer could also receive an extra reward based on the length of the queue at arrival. The rewards can be monetary or otherwise.*

2. RELATED WORKS

Our work builds upon queueing systems and multi-agent systems. In particular, the notion of wait times that we take as our performance metric has been characterized in the case of FIFO queues in [10]. Alternative modes of priority to arrival time in a queue have been proposed in numerous works (e.g., [9]). These modes include: agent-priority [12], shortest-job first [11], and multiple queues [16]. All of these do not exhibit signaling between an operator and the customers. The problem of optimization of multiple queues or queueing networks is usually modeled as a Markov decision problem [18]. In contrast to our work, these lines of work do not exhibit signals sent by the operator as control variables.

Our signaling approach is an alternative to other existing approaches to managing queue, such as advanced booking, reservations, and differential pricing deployed at theme parks (e.g., Disney parks, Sea World). Optimization of a single queue has been studied when the service rate (μ in our notation) or the arrival rate (λ_i in our notation) of customers can be controled [7]. This is however not the case in our model, where the operator only controls the signals sent to customers. So-called Active Queue Management algorithms [1] have been proposed to deal with burstiness in communication networks by dropping packets in real-time. However, such approaches are useless when we exclude the option of turning customers away. A potential feature of our proposed solution is that it not employ any real-time control action: all signals are generated and sent at a single time instant. This feature also sets our work apart from the studies of queueing systems with feedback [2].

In our work, we adopt a model of a heterogeneous population of customers characterized by samples from a probability distribution, which is commonly used in statistics to model aggregate effects [8]. The intricate behavioural aspects of queueing customers, although widely studied [15, 20], is not a focus in this paper.

The works most closely related to ours concern multi-agent queues, where the agents can interact strategically (as in the case of queueing games [15]), or by following fixed policies as in our model. In particular, two notions of signals have been introduced in such queues. [13] gives an overview of game theoretic aspects of queueing among utility-maximizing agent. In contrast to the notions of signaling of [13]—where the signal process is typically the length of the queue at every time instant, we consider a set of signals that are sent only once at the start of time. [5] examines how wait time announcements affect the actions of customers in a multi-server system. These announcements are updated in real-time, whereas the signals in our proposed system are sent once before the first customer arrival.

Beyond obvious applications in managing queues of human customers in cities [14], our solution approach can be applied in a number of areas, notably communication network queues [17], transportation network queues [19]. The queueing liteature is extensive and our survey is not exaustive. However, this is one of the first times that signaling and queues have been combined to regulate arrival processes.

3. FIRST-IN FIRST-OUT QUEUES

Most queues operate on a first come, first served basis. In a first-in first-out (FIFO) queue, the person who comes in earlier is served earlier. In such situations, burstiness can cause the queueing system to be overwhelmed, and to excessively long wait times.

We examine a simple FIFO system; there are $i \in \{1, ..., N\}$ customers and only one server. The *service rate* is μ, where $1/\mu$ is the mean service time. For simplicity of presentation, we will examine a special case of the M/M/1 queue [6] where the service time is the same for each customer, i.e., a constant service time $a = \frac{1}{\mu}$. Note that these assumptions are to aid exposition and that our results generalize to i.i.d. service times in a straightforward fashion.

Each customer arrives at the end of the queue at a *nominal arrival time* x_i. We will assume that $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_N$. Let $e_1, e_2, \ldots$ denote independent exponential random variables with parameters $\lambda_1, \lambda_2, \ldots$, respectively. As in the case of M/M/1 queues, we assume the time between consecutive arrivals are independent and exponentially distributed, such that

$$x_i = \sum_{j=1}^{i-1} e_j, \quad \text{for all } i = 2, \ldots, N.$$

The parameter λ_i is the *arrival rate* of the i-th customer; a large value of λ_i indicates that more customers are arriving per unit time. Although μ is a constant, the arrival rate λ_i varies over the the customer index i and hence over time. This varying arrival rate models arrival bursts, when many customers join the queue in a short time period, causing long wait times.

Each customer leaves the queue after some waiting period, or *nominal wait time*, d_i. The time that each customer leaves the queue is then $x_i + d_i$. The wait time of each customer

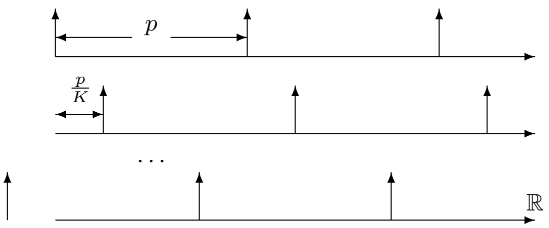

Figure 1: Illustration of K sequences of suggested times corresponding to different signals. There is a shift of p/K between sequences corresponding to consecutive signals, and a period of p between suggested times within each sequence.

depends on the wait times of the customers in front of him. In general, every other customer after the first customer has a (random) wait time of

$$d_i = \max\{a + x_{i-1} + d_{i-1} - x_i, 0\}.$$

The traditional FIFO queue operator has no control over the wait times $\{d_i\}$ due to the lack of actuation capabilities. Our objective is now to reduce the total wait times of all customers by introducing a new queueing system that involves signals sent by the operators to the customers before they arrive.

4. NEW QUEUEING SYSTEM WITH SIGNALS

We aim to decrease wait times by having the operator send signals to all of the customers incentivising them to arrive at suggested times. We call the resulting system a *signaled queueing system* or *signaled queue.* For example, when airline customers check-in online, they may be offered a discount on future flights, if they arrive at the check-in desk at a certain time.

At time $t = 0$, the operator picks two parameters: a real number $p > 0$ called the period, and an integer K called the number of signals. The operator samples N i.i.d. random variables $s_1, \dots, s_N$ from an uniform probability distribution over support $\{0, 1, \dots, K-1\}$. Then, for every customer $i = 1, \dots, N$, the operator sends the *signal* s_i.

The value of the signal corresponds to one of the following sequences of *suggested arrival times*:

$$\begin{aligned}
&0, p, 2p, \dots, \quad \text{if } s_i = 0 \\
&\delta, p+\delta, 2p+\delta, \dots, \quad \text{if } s_i = 1 \\
&\vdots \\
&(K-1)\delta, p+(K-1)\delta, \dots, \quad \text{if } s_i = K-1,
\end{aligned}$$

where for simplicity, we write $\delta = p/K$ to denote the *interval* between consecutive signal values. The suggested arrival times corresponding to different signals are also illustrated in Figure 1.

4.1 Customer response

In our proposed queueing system, each customer's new or *adjusted arrival time* at the end of the queue, y_i, will be a function of their original arrival time or the nominal arrival time, x_i, the signal they received, s_i, and a parameter σ_i that models how much customer i pays attention to the signal:

$$y_i = f(x_i, s_i, \sigma_i).$$

We now propose a model for how customers will respond to the signals. The parameter σ_i represents customer i's tendency to follow the signal, i.e., arrive at a time close to a suggested arrival time. To model a heterogeneous population of customers, we assume that $\sigma_1, \dots, \sigma_N$ are i.i.d. random variables with the interval $[0, 1]$ as support. When $\sigma_i = 0$, the customer pays no attention to the signal and behaves as if he had never received a signal, and therefore, $y_i = x_i$. When $\sigma_i = 1$, the customer arrives at a suggested arrival time close to its nominal arrival time x_i. For simplicity, we take this suggested arrival time as $(\lfloor\frac{x_i}{p}\rfloor + \frac{s_i}{K})p$, which is either the closest suggested arrival time below or above the nominal arrival time x_i.

In sum, we consider the following instance of this function $y_i = f(x_i, s_i, \sigma_i)$ as a model of customer response:

$$y_i = (1-\sigma_i)x_i + \sigma_i\left(\lfloor\frac{x_i}{p}\rfloor + \frac{s_i}{K}\right)p, \quad \text{for } i = 1, \dots, N,$$

where $\lfloor\cdot\rfloor$ denotes the round-down-to-the-nearest-integer operator. The equation satisfies our constraints that $y_i = x_i$ when $\sigma_i = 0$ and $y_i = (\lfloor\frac{x_i}{p}\rfloor + \frac{s_i}{K})p$ when $\sigma = 1$. The set of y_i is not necessarily sorted in ascending order. Let $y_{(1)} \leq y_{(2)} \dots \leq y_{(N)}$ denote the sequence of sorted adjusted arrival times. We define the corresponding *adjusted wait times* $h_1, \dots, h_N$. Assuming that $h_1 = 0$, for every $i = 2, 3, \dots, N$, the wait time for the customer arriving at time $y_{(i)}$ is

$$h_i = \max\{a + y_{(i-1)} + h_{i-1} - y_{(i)}, 0\}.$$

4.2 Performance Comparison

In this section, we compare the performances of the signaled system and the FIFO system by examining the wait times experienced by customers. Our objective is to show that these wait times are shorter in the signaled case. The following theorem gives a performance guarantee on signaled queues, relative to FIFO queues. It says that we expect that the wait-time experienced by customers in the FIFO queue to be larger than in the signaled queue. In particular, the theorem bounds the probabilistic performance improvement. To set notation, recall that h_i is the wait-time of the i-th customer to arrive in the signaled queue, whereas d_i is the wait-time of the i-th customer to arrive in the FIFO queue.

Remark 1. *The customers corresponding to d_i and h_i need not the same physical customer.*

Remark 2 (Comparing d_i and h_i). *How meaningful is a comparison between d_i and h_i? Arguably, customers can comply with the suggested arrival times and wait outside the queue before joining the queue. In this case, such customers experience two wait times: inside the queue and outside the queue. We argue that the wait time outside the queue can be put to better use than time in the queue: having a coffee or shopping in the airport check-in example.*

Theorem 1. *Let $\mu_1 = \mathbb{E}\sigma_i$ and $\mu_2 = \mathbb{E}s_i/K$. Let $v = \mathbb{V}(\sigma_i s_i/K) = \mathbb{E}\sigma_i^2\mathbb{E}(s_i/K)^2 - (\mathbb{E}\sigma_i)^2(\mathbb{E}s_i/K)^2$. For every customer $i \in \{1, \dots, N\}$, every $\zeta > 0$, every $z \geq 0$, and*

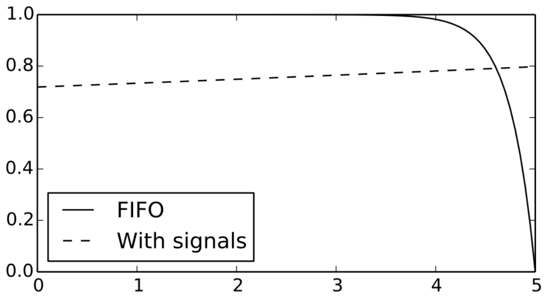

Figure 2: Plots of $\mathbb{P}(d_i > \zeta + z \mid d_{i-1} = z)$ in solid line and the bound on $\mathbb{P}(h_i > \zeta + z \mid h_{i-1} = z)$ in broken line versus ζ, as derived in Theorem 1. We used the following set of values of the parameters: $a = 5$, $p = 5$, $\lambda = 4$, $\zeta \in [0, a]$, γ such that the three fractions on the right-hand side of (2) are approximately equal.

every $\gamma \geq 0$, we have:

$$\mathbb{P}(d_i > \zeta + z \mid d_{i-1} = z) = 1 - e^{-\lambda(a-\zeta)}, \tag{1}$$

$$\begin{aligned} \textit{and} \quad & \mathbb{P}(h_i > \zeta + z \mid h_{i-1} = z) \\ & \leq \frac{N\lambda^{-2}}{(\gamma - N/\lambda)^2} + \frac{N\lambda^{-2}}{((\zeta - a - \gamma)/2 - N/\lambda)^2} \\ & + \frac{v}{(\frac{\zeta - a - \gamma}{2p} - \mu_1\mu_2)^2}. \end{aligned} \tag{2}$$

The proof of Theorem 1 appears in the Appendix.

The probabilities of Theorem 1 are further illustrated in Figure 2. We observe that indeed for all values of ζ, the upper bound on $\mathbb{P}(h_i > \zeta + z \mid h_{i-1} = z)$ is less than $\mathbb{P}(d_i > \zeta + z \mid d_{i-1} = z)$.

Theorem 1 shows that customers in the signaled queues are less likely to wait for a long time than customers in the FIFO queues. As further illustration, Figure 3 shows empirically that the signaled system's wait times are more likely to be small than the FIFO system's wait times.

5. EMPIRICAL RESULTS

This section presents in-depth simulations to illustrate the performance of signaled queues when compared to FIFO queues. Thus, this section complements empirically the previous section.

Specifically, we examine how the values of different parameters affect wait times in the signaled queueing system. Questions to be examined include: how many distinct signals K there should be; what should the period p be; and finally, how much should customers adhere to their suggested times for the system to work adequately. To this end, we define the *relative performance* of the signaled queue compared to the traditional FIFO queue as the ratio between the sums of adjusted and nominal wait times: $\frac{\sum_{i=1}^{N} h_i}{\sum_{j=1}^{N} d_j}$. If this ratio is significantly smaller than one, then signaling is indeed decreasing the aggregated customers' wait times in a noticeable manner.

Throughout our simulations, we set the number of customers N to 5000, the service time a to 5. To model bursty arrivals, the arrival rate sequence $\{\lambda_i\}$ is piecewise-constant, alternating between a high value and a low value every $N/4$ consecutive customers (cf. top of Figure 8). Each simulation scenario is repeated over 100 trials, and average relative performance is plotted along with error-bars corresponding to one standard deviation.

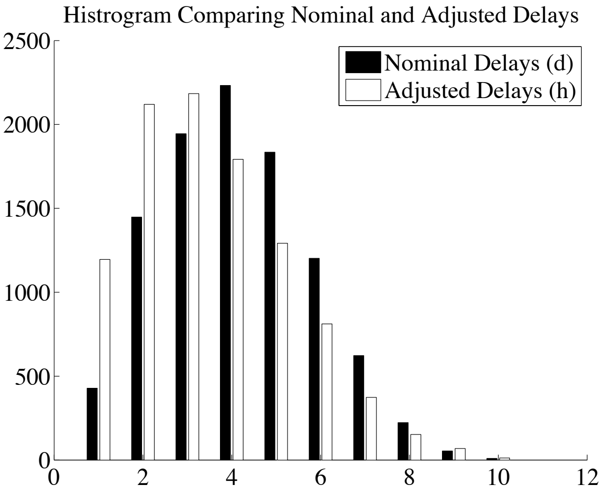

Figure 3: Histogram comparing the probability distributions of the sum of FIFO wait times, $\sum_{j=1}^{N} d_i$, and the sum of adjusted wait times, $\sum_{j=1}^{N} h_i$, for 10,000 trials and one hundred customers, $N = 100$. There is a higher probability that the sum of the adjusted wait times will be within the first bin than the sum of FIFO wait times to be within the first bin.

5.1 Relative Performance versus Interval

The operator can vary the length of time between two consecutive signals, the interval $\delta = p/K$. We set $\sigma_i = 1$ for every i (all customers arrive at the nearest suggested time).

First, we investigate the total time $y_{(N)} - y_{(1)}$ in the signaled queue compared to the total time $x_N - x_1$ without signals. For the case $\delta = a$, Figure 4 shows the ratio $\frac{y_{(N)} - y_{(1)}}{x_N - x_1}$ as a function of the number of distinct signals K. As expected, this ratio is close to 1 for small values of K. However, as K increases, so does the period $p = \delta K$, which leads to large values of $(\lfloor \frac{x_N}{p} \rfloor + \frac{s_N}{K})p$.

Figure 5 compares the relative performances for three values of δ. A larger value of δ leads to a better performance. This is expected because the more time between customer arrivals, the lower their wait times. However, a large δ also leads to a large period p and a large total time $y_{(N)} - y_{(1)}$.

The number of signals per period, K, affects the difference between a customer's nominal arrival time and the nearest of his suggested arrival times. This means that the customer would have to make a large adjustment to follow the suggestion. Figure 6 shows that when K is large, the average relative adjustment required from the customer can become non-negligible.

5.2 Relative Performance versus Burstiness

We now investigate the relative performance for different burst patterns, as modeled by the sequence $\{\lambda_i\}$. During a burst, many customers arrive in a short amount of time, according to a high arrival rate, $\bar{\lambda}$. We consider the burst

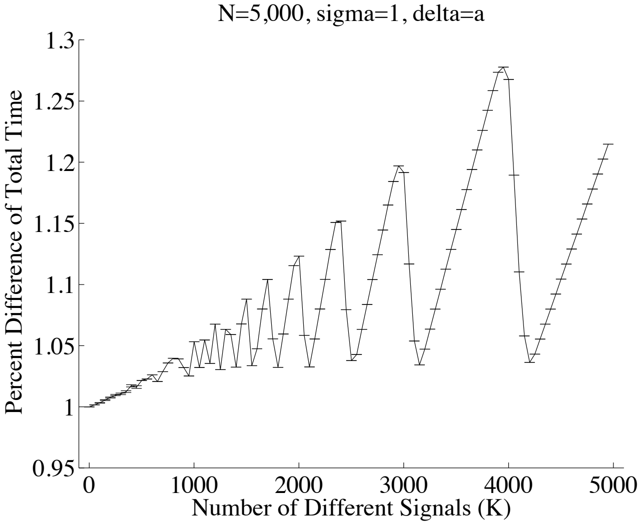

Figure 4: The relative difference between the nominal total time and the adjusted total time, $\frac{y_{(N)}-y_{(1)}}{x_N-x_1}$, depends on the number of signals K. As K increases, the signaled system takes more time than the FIFO system.

patterns of Figure 7. In each pattern, there is the same number of customers with a high arrival rate, who arrive in one or multiple bursts. Figure 8 shows that the relative performance improves as bursts (rush periods) become shorter (and more numerous).

Finally, in Figure 9, we examine how the relative performance depends on the value of the arrival rate $\bar{\lambda}$ during bursts, relative to the normal one $\underline{\lambda}$. The larger the difference between the two, the better is the relative performance of the signaled queue.

5.3 Relative Performance versus σ_i

In the previous simulations, the distribution of the random variable, σ_i, that determines how carefully each customer follows its suggested arrival times, has been ignored—we have set $\sigma_i = 1$ uniformly. We had assumed that all customers follow their suggested arrival times exactly. In this section, we take $\{\sigma_i\}$ as i.i.d. samples from a probability distribution over the support $[0, 1]$. To make things more realistic, we choose the *beta* distribution with shape parameters α and β (cf. Figure 11), and corresponding probability density function

$$g(x) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in [0,1].$$

where the B denotes the beta function. Figure 10 shows how the shape parameters of the beta distribution affect the relative performance.

In Figure 10, we compare the different beta distributions with the ideal case, $\sigma = 1$, and the worst case, $\sigma = 0$. When $\sigma = 1$, all of the customers follow their signals with some error. When $\sigma = 0$, none of the customers follow their signals and they all arrive at their original arrival times, thus the system performance is always 1 because none of their arrival times change. When σ is determined by a beta distribution, there are different probabilities for the degree a customer listens to the suggested arrival times. The probability density functions of the beta distributions used in Figure 10 are shown in Figure 11.

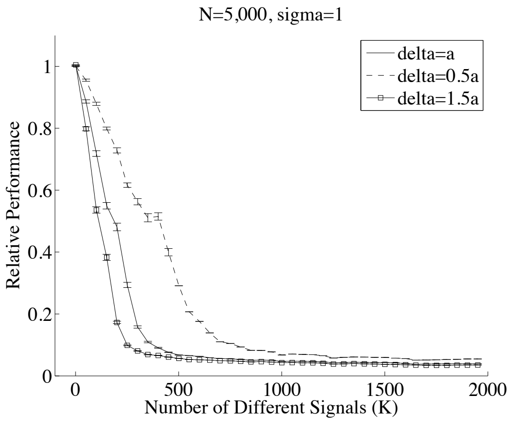

Figure 5: Comparing the relative performances of different ratios between the interval and service time when K varies from 1 to 2,000. When $\delta < a$, customers are arriving faster than customers are leaving so the performance of the system is worse. When $\delta > a$, customers are arriving slower than the rate that customers are being served so the performance is good because not many people are waiting. However, when $\delta > a$, the total time for the signaled system is much larger than the FIFO system. We choose $\delta = a$ because it spaces out the signals enough such that the system performance is good but not too much that the total time is too large.

Figure 10 shows that the smaller the shape parameter β, the better the system performance. When the mean of the beta distribution tends towards one, the system performance increases because the probability that customers listen to their signals increases. The more distribution weight is near 1, the better the signaled queue performs. One way to ensure customers act as intended by the signals is to give incentives based on the difference between their signals and their actual arrival times.

Remark 3 (Incentives). *Rewards may or may not be necessary to ensure that customers listen to their signals. If the operator chooses to use rewards, these do not have to be monetary. For example, the crowd-sourcing traffic app Waze gives customers incentives in the form of points when they report traffic incidents. The underlying human response to incentives is however beyond the scope of this paper.*

APPENDIX

A. PROOFS

Proof of Theorem 1. The proof is broken down into four parts. First, we derive the probability that the FIFO wait time is larger than a constant. Second, we bound the probability that the adjusted wait time is larger than a constant. This is done in three steps, first the probability is bounded by two terms, then, each of these is bounded separately. We then compare the two probabilities for all i.

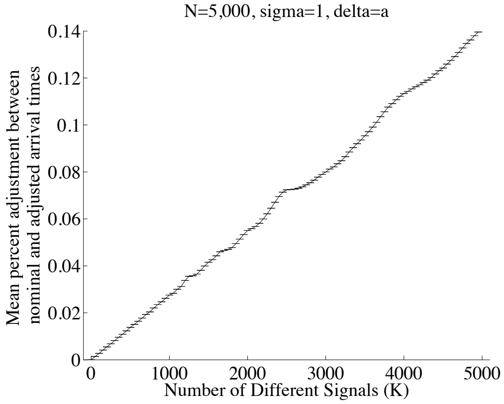

Figure 6: The average relative adjustment made when each customer listens to their signals completely, $\frac{1}{N}\sum_{i=1}^{N}\frac{y_{(i)}-x_i}{x_i}$, as K varies from 1 to 5000. As the number of signals per period, K, increases the average adjustment also increases. We introduced periodic signals to avoid large adjustments such that customers would only have to make reasonable adjustments to arrive at their signals. Observe that K should be chosen small enough such that these adjustments are reasonable.

Part I. The first part of the proof is devoted to

$$\mathbb{P}\left(d_i > \zeta + z \mid d_{i-1} = z\right).$$

First, observe that conditioned on the event $d_{i-1} = z$, we have

$$d_i = \max\{a + x_{i-1} + z - x_i, 0\}.$$

We assume that the nominal arrival times, x_i, are determined by a nonhomogenous Poisson process, where the expected number of arrivals per unit time varies with time. The nominal arrival times are independent of each other and will be Poisson distributed.

We define e_i as the difference between two consecutive nominal arrival times; $x_i = x_{i-1} + e_i$. We assume that $x_1 = 0$ such that $x_i = \sum_{j=1}^{i} e_j$. The e_i are independent of each other and exponentially distributed with the parameter r. We rewrite our wait time using e_i.

$$\begin{aligned}&d_i = \max\{a + x_{i-1} + z - x_i, 0\} =\\ &d_i = \max\{a + z - e_i, 0\}\end{aligned}$$

We aim to find the probability that the FIFO wait time will be greater than a constant. Since $\zeta + z > 0$, we can write

$$\begin{aligned}&\mathbb{P}\left(d_i > \zeta + z \mid d_{i-1} = z\right)\\ &= \mathbb{P}\left(a + z - e_i > \zeta + z \cap a + z - e_i > 0\right)\\ &= \mathbb{P}\left(e_i < a - \zeta \cap e_i < a + z\right)\\ &= \mathbb{P}\left(e_i < a - \zeta\right)\\ &= 1 - e^{-\lambda(a-\zeta)}.\end{aligned}$$

Part II. The second part of the proof is devoted to

$$\mathbb{P}\left(h_i > \zeta + z \mid h_{i-1} = z\right).$$

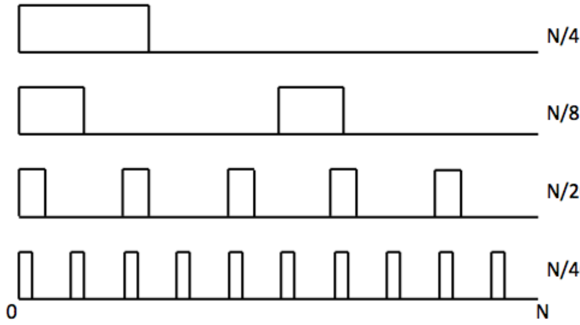

Figure 7: Burst patterns with different numbers R of customers per burst, for $R = N/4, N/8, N/20, N/40$. When R is large, then there are a small number of arrival bursts but the period is very large. When R is small, there there are lots of quick bursts of arrivals.

Observe that like d_i, the new wait time h_i depends on z, the variable $y_{(i)}$, and the service time a:

$$h_i = \max\{a + y_{(i-1)} + z - y_{(i)}, 0\}.$$

Since $\zeta + z > 0$, we can write

$$\begin{aligned}&\mathbb{P}\left(h_i > \zeta + z \mid h_{i-1} = z\right) &(3)\\ &= \mathbb{P}\left(\max\{a + y_{(i-1)} + z - y_{(i)}, 0\} > \zeta + z\right) &(4)\\ &= \mathbb{P}\left(y_{(i-1)} - y_{(i)} > \zeta - a\right). &(5)\end{aligned}$$

Recall that the non-ordered adjusted arrival time is

$$y_i = (1 - \sigma_i)x_i + \sigma_i(\lfloor\frac{x_i}{p}\rfloor + \frac{s_i}{K})p$$

There exists an index that depends on the index of the ordered arrival times, $k(i)$, such that

$$y_{(i)} = (1 - \sigma_{k(i)})x_{k(i)} + \sigma_{k(i)}(\lfloor\frac{x_{k(i)}}{p}\rfloor + \frac{s_{k(i)}}{K})p$$

There exists another index $k(i-1)$ that depends on the index of the ordered arrival times, such that

$$\begin{aligned}y_{(i-1)} &= (1 - \sigma_{k(i-1)})x_{k(i-1)}\\ &+ \sigma_{k(i-1)}(\lfloor\frac{x_{k(i-1)}}{p}\rfloor + \frac{s_{k(i-1)}}{K})p\end{aligned}$$

For ease of notation, we write:

$$\begin{aligned}&\alpha_i = (1 - \sigma_{k(i-1)})x_{k(i-1)} - (1 - \sigma_{k(i)})x_{k(i)}\\ &\beta_i = \sigma_{k(i-1)}(\lfloor\frac{x_{k(i-1)}}{p}\rfloor + \frac{s_{k(i-1)}}{K})p - \sigma_{k(i)}(\lfloor\frac{x_{k(i)}}{p}\rfloor + \frac{s_{k(i)}}{K})p\\ &y_{(i-1)} - y_{(i)} = \alpha_i + \beta_i.\end{aligned}$$

We can rewrite (5) as:

$$\begin{aligned}&\mathbb{P}\left(h_i > \zeta + z \mid h_{i-1} = z\right)\\ &= \mathbb{P}\left(\alpha_i + \beta_i > \zeta - a\right).\end{aligned}$$

By introducing an arbitrary constant γ and conditioning, we obtain:

$$\begin{aligned}&\mathbb{P}\left(\alpha_i + \beta_i > \zeta - a\right)\\ &= \mathbb{P}\left(\beta_i > \zeta - a - \alpha_i \mid \alpha_i < \gamma\right\}\mathbb{P}\{\alpha_i < \gamma)\\ &+ \mathbb{P}\left(\beta_i > \zeta - a - \alpha_i \mid \alpha_i \geq \gamma\right)\mathbb{P}\left(\alpha_i \geq \gamma\right).\end{aligned}$$

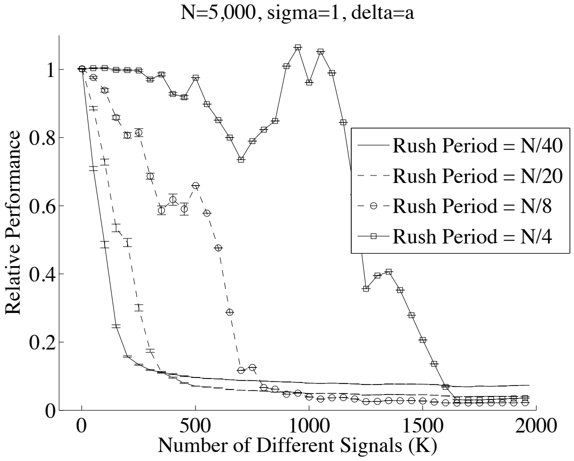

Figure 8: Comparing the relative performances of different burst patterns as K varies from 1 to 2,000. The number of customers arriving during a burst is varied; $R = \frac{N}{4}, \frac{N}{8}, \frac{N}{20}, \frac{N}{40}$. When R is small, then the burst is also small. K should be big enough such that the signal period p is bigger than the burst.

By observing that

$$\mathbb{P}\left(\beta_i > \zeta - a - \alpha_i \mid \alpha_i \geq \gamma\right) \leq 1$$

and that

$$\mathbb{P}\left(\beta_i > \zeta - a - \alpha_i \mid \alpha_i < \gamma\right) \leq \mathbb{P}\left(\beta_i > \zeta - a - \gamma\right),$$

we obtain:

$$\begin{aligned} &\mathbb{P}\left(h_i > \zeta + z \mid h_{i-1} = z\right) && (6)\\ &\leq \mathbb{P}\left(\beta_i > \zeta - a - \gamma\right) + \mathbb{P}\left(\alpha_i \geq \gamma\right). && (7)\end{aligned}$$

Part III. Now, we bound the second term on the right-hand side of (7):

$$\begin{aligned} &\mathbb{P}\left(\alpha_i \geq \gamma\right) && (8)\\ &\leq \mathbb{P}\left((1-\sigma_{k(i-1)})x_{k(i-1)} \geq \gamma\right) && (9)\\ &\leq \mathbb{P}\left(x_{k(i-1)} \geq \gamma\right) && (10)\\ &\leq \mathbb{P}\left(x_N \geq \gamma\right) && (11)\\ &\leq \mathbb{P}\left(x_N - N/\lambda \geq \gamma - N/\lambda\right) && (12)\\ &\leq \mathbb{P}\left(x_N - N/\lambda \geq (\gamma - N/\lambda)\frac{\lambda}{\sqrt{N}}\frac{\sqrt{N}}{\lambda}\right) && (13)\\ &\leq \frac{N\lambda^{-2}}{(\gamma - N/\lambda)^2}, && (14)\end{aligned}$$

where we used the facts that α_i is the difference of two non-negative valued random variable, and that $1-\sigma_{k(i)} \leq 1$, and the last inequality follows by Chebychev's Inequality.

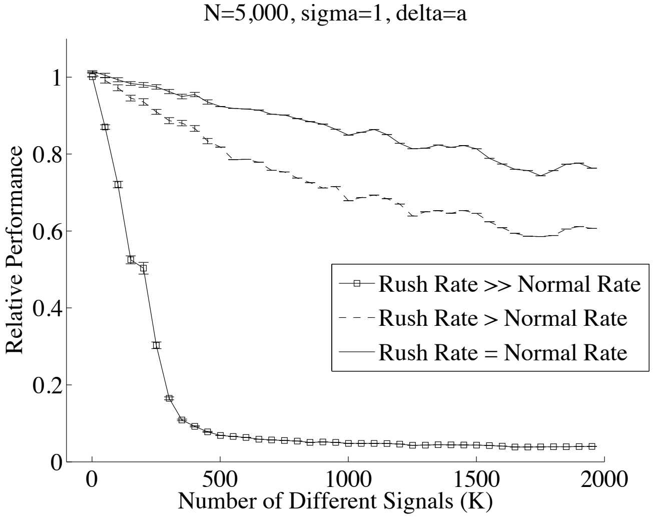

Figure 9: Comparing the relative performances of different arrival rates as K varies from 1 to 2,000. We defined two values for the arrival rate; a large $\bar{\lambda}$, and a small $\underline{\lambda}$. When $\bar{\lambda}$ is much larger than $\underline{\lambda}$, our system performs very well and decreases the wait times by a lot. When $\bar{\lambda}$ is slightly larger than $\underline{\lambda}$, the system performs well. In the extreme case, when $\hat{\lambda} = \underline{\lambda}$, our system does not decrease wait times by very much.

Part IV. Now, we turn our attention to:

$$\begin{aligned} &\mathbb{P}\left(\beta_i > \zeta - a - \gamma\right)\\ &= \mathbb{P}(\sigma_{k(i-1)}(\lfloor\frac{x_{k(i-1)}}{p}\rfloor + \frac{s_{k(i-1)}}{K})p\\ &- \sigma_{k(i)}(\lfloor\frac{x_{k(i)}}{p}\rfloor + \frac{s_{k(i)}}{K})p > \zeta - a - \gamma)\\ &\leq \mathbb{P}\left(\sigma_{k(i-1)}(\lfloor\frac{x_{k(i-1)}}{p}\rfloor + \frac{s_{k(i-1)}}{K})p > \zeta - a - \gamma\right)\\ &\leq \mathbb{P}\left(\sigma_{k(i-1)}\lfloor\frac{x_{k(i-1)}}{p}\rfloor p > \frac{\zeta - a - \gamma}{2}\right)\\ &+ \mathbb{P}\left(\sigma_{k(i-1)}\frac{s_{k(i-1)}}{K}p > \frac{\zeta - a - \gamma}{2}\right).\end{aligned}$$

where we used the Union Bound to obtain the last inequality. Next, we consider the two terms on the right-hand side separately. First, observe that, by similar arguments as Part III,

$$\begin{aligned} &\mathbb{P}\left(\sigma_{k(i-1)}\lfloor\frac{x_{k(i-1)}}{p}\rfloor p > \frac{\zeta - a - \gamma}{2}\right) && (15)\\ &\leq \mathbb{P}\left(x_{k(i-1)} > \frac{\zeta - a - \gamma}{2}\right) && (16)\\ &\leq \frac{N\lambda^{-2}}{((\zeta - a - \gamma)/2 - N/\lambda)^2}. && (17)\end{aligned}$$

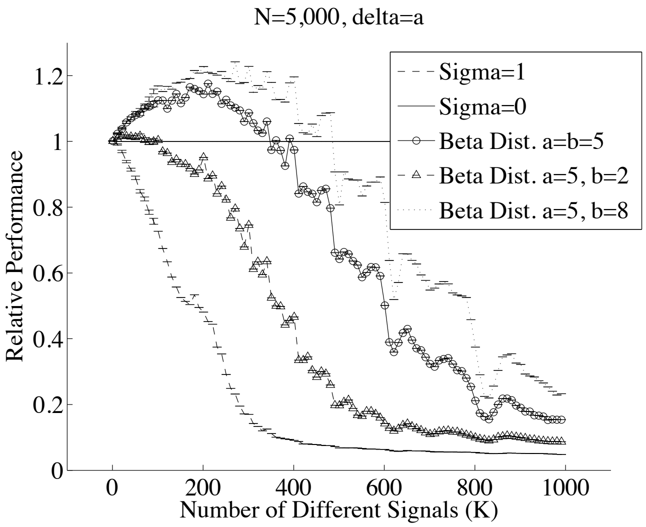

Figure 10: Comparing the relative performances of different distributions of the number of customers listen to their signals, σ, as K varies from 1 to 2,000. The ideal case is when all of the customers listen to their signals, $\sigma = 1$. The worst case is when nobody listens to their signals, $\sigma = 0$. Realistically, not every customer is going to follow their signals exactly. We choose σ to be a beta distribution. We examine the performances of three different beta distributions and all perform better than when $\sigma = 0$. The plots of the beta distributions probability density functions are shown in Figure 11. The system performs best when the mean of the beta distribution is close to one.

Finally, we bound:

$$\mathbb{P}\left(\sigma_{k(i-1)}\frac{s_{k(i-1)}}{K}p > \frac{\zeta-a-\gamma}{2}\right) \tag{18}$$

$$= \mathbb{P}\left(\sigma_{k(i-1)}\frac{s_{k(i-1)}}{K} - \mu_1\mu_2 > \frac{\zeta-a-\gamma}{2p} - \mu_1\mu_2\right) \tag{19}$$

$$\leq \frac{v}{(\frac{\zeta-a-\gamma}{2p} - \mu_1\mu_2)^2}, \tag{20}$$

where we used the independence property and Chebychev's Inequality, and the definitions of v, μ_1, and μ_2. The claim of (2) follows by combining (14), (17), and (20). □

B. REFERENCES

[1] R. Adams. Active queue management: a survey. *Communications Surveys & Tutorials, IEEE*, 15(3):1425–1476, 2013.

[2] C. E. Agnew. Dynamic modeling and control of congestion-prone systems. *Operations research*, 24(3):400–419, 1976.

[3] E. Altman. Application of dynamic games in queues. *Annals of the International Society of Dynamic Games*, pages 309–342, 2005.

[4] F. Antunes, F. Coito, and H. Duarte-Ramos. A linear approach towards modeling human behavior. *Technological Innovation for Sustainability*, pages 305–314, 2011.

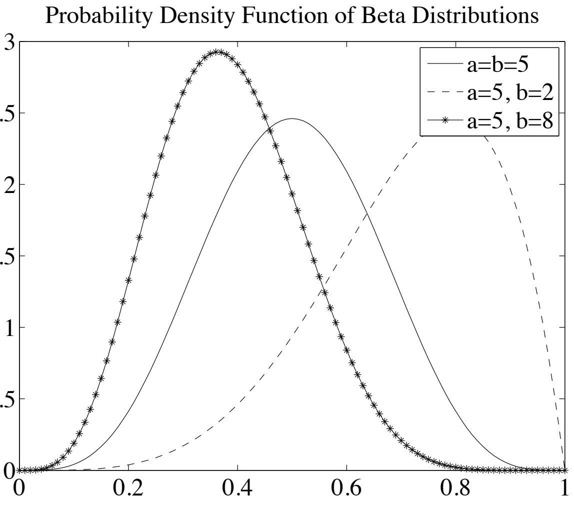

Figure 11: Probability density functions of different beta distributions that are used as the distribution of σ in Figure 10.

[5] M. Armony, N. Shimkin, and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, pages 66–81, 2009.

[6] J. Blanchet, P. Glynn, and S. Meyn. Large deviations for the empirical mean of an M/M/1 queue. To appear in *Queueing Systems*, Special Issue on Simulation of Networks, 2012.

[7] M. Chiang, A. Sutivong, and S. Boyd. Efficient nonlinear optimizations of queuing systems. In *Global Telecommunications Conference*, volume 3, pages 2425–2429, 2002.

[8] W. G. Cochran. *Sampling Techniques*. Wiley, 1977.

[9] R. J. Gibbens and F. P. Kelly. On packet marking at priority queues. *IEEE Transactions on Automatic Control*, 47:1016–1020, 2002.

[10] P. W. Glynn and W. Whitt. A central-limit-theorem version of $l = \lambda w$. *Queueing Systems*, 2:191–215, 1986.

[11] M. Harchol-Balter. *Performance modeling and design of computer systems: queueing theory in action.* Cambridge University Press, 2013.

[12] M. Harchol-Balter, T. Osogami, A. Scheller-Wolf, and A. Wierman. Multi-server queueing systems with multiple priority classes. *Queueing Systems*, 51(3-4):331–360, 2005.

[13] R. Hassin and M. Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Springer, 2003.

[14] S. Kolli and K. Karlapalem. Mama: Multi-agent management of crowds to avoid stampedes in long queues. In *Proceedings of AAMAS*, pages 1203–1204, 2013.

[15] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

[16] M. J. Neely. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks*, 3(1):1–211, 2010.

[17] S. Shakkottai and R. Srikant. *Network Optimization and Control*. Now Publishers, 2008.

[18] S. Stidham Jr and R. Weber. A survey of markov decision models for control of networks of queues. *Queueing Systems*, 13(1-3):291–314, 1993.

[19] W. Vickrey. *Public Economics*. Cambridge University Press, 1994.

[20] U. Yechiali. On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research*, 19(2):349–370, 1971.