Last in Line

Esa Hyytiä*

Department of Computer Science
University of Iceland

Abstract

Queueing is a common praxis in banks, hospitals and transportation, just to name a few. One common performance metric is the mean sojourn time. However, humans experience waiting time in a more complex manner – they dislike being the last in line. We study queueing systems subject to such a cost structure. For the single M/G/1 queue, we derive the corresponding mean costs, value functions and admission costs, which are then applied to route customers to parallel servers.

Keywords: Queueing; last-place aversion; M/G/1; dispatching; MDP

1. Introduction

Many server systems for (human) customers involve queues. Parallel servers are used when the capacity of a single server is insufficient. A typical example is service counters, e.g., at airports or shopping centers, call centers (help desks) and checkout counters. When each server has its own queue and incoming customers are assigned immediately upon arrival to one of them, the corresponding model is known as the dispatching system.

The most common queueing discipline for a single server is the first-come-first-served (FCFS). It is seen as a fair queueing discipline as customers depart in the same order as they arrived – no one is allowed to cut the queue [1]. In addition to fairness, one is often concerned also about the performance, for which the classical metric has been the mean sojourn time. According to Pollazcek-Khinchine's mean value formula, the mean sojourn time with Poisson arrivals and FCFS depends on the second moment of the service time. Consequently, FCFS does a very poor job when service times vary a lot. The optimal scheduling for the mean sojourn time is known as the shortest-remaining-processing-time (SRPT) [2].

So we can say that FCFS promotes fairness and SRPT minimizes the mean sojourn time. However, humans experience waiting systems in a more complex manner especially when they can observe their own position in the queue. It turns out that they simply dislike being the last in line [3]. In this paper, we study queueing systems with such a cost structure in the framework of Markov decision processes. First we analyze the single M/G/1 queue, and then apply the new results to derive efficient (heuristic) dispatching policies for parallel queues that aim to minimize the annoyance the customers experience for being the last in line.

*Corresponding author
Email address: esa@hi.is (Esa Hyytiä)
[1]Postal address: University of Iceland, Dunhagi 5, 107 Reykjavík, Iceland.

Figure 1: The customer being last in line finds the waiting time most unpleasent.

2. Single M/G/1 Queue with Last-in-Line Costs

In this section, we consider the single M/G/1 queue with FCFS scheduling, i.e., customers arrive according to a Poisson process with rate λ, and their service times X_i are i.i.d. with a general distribution $X_i \sim X$. The queue is stable whenever the offered load, $\rho = \lambda\,\mathbb{E}[X]$, is less than one.

We focus on the so-called last-in-line cost structure, where a customer *being last in line* incurs costs at some rate, i.e., until the next customer arrives, or he himself enters the service. An example system is depicted in Figure 1. We let W_L denote the time a customer is last in line, and refer to this time interval as *the last-in-line time*. We consider three cases:

1. Customers have the same unit holding cost rate while being the last in line, and the total cost is thus W_L.
2. The holding cost rates are customer-specific i.i.d. random variables, $H_i \sim H$, that can reflect, e.g., a service class (cf. high priority customers). The total cost is $H \cdot W_L$.
3. Holding cost rate increases in time so that the total cost is $(W_L)^k$ for some integer $k \geq 2$.

In what follows, we give exact expressions for the respective mean costs $\mathbb{E}[W_L]$, $\mathbb{E}[H] \cdot \mathbb{E}[W_L]$ and $\mathbb{E}[(W_L)^k]$, and derive the corresponding value functions and admission costs.

2.1. Linear costs for the time being last in line

First we assume that all customers have equal cost rate, $h = 1$, at which they incur costs while being last in line. Equivalently, each queue incurs costs at unit rate whenever more than one customer is present (i.e., whenever someone is waiting).

Note that this is different from incurring costs at a constant rate whenever a server is busy, which is an elementary model for energy consumption. In fact, for energy consumption, the service order is irrelevant. In contrast, our cost structure is more intricate also in this sense. For example, given two customers, it would be better to serve first the customer with a shorter service time, as then the system will stop incurring costs earlier. In fact, it is easy to show that for every sample path, the optimal scheduling serves the customer with the longest service time last. Other customers can be served in any order. In this sense, the optimal scheduling resembles SRPT, but has less constraints. The same principle can be applied also to systems with multiple servers (without arrivals).

However, in our model customers arrive according to Poisson process and are served in FCFS order, which makes the situation more challenging. First, the current state affects the waiting times of the later arriving customers, and second, the next arriving customer affects the costs incurred by the customer who is currently last in line.

The complete state (size-aware) description of the queue is $(x_1, x_2, \ldots, x_m)$, where x_i denotes the (remaining) service time of customer i with the convention that customer 1 (if any) is currently receiving service, and customer m is last in line. For our cost structure, a sufficient state description is (u, w^*), where $w^* = x_1 + \ldots + x_{m-1}$ is the amount of work in the queue ahead of the customer currently being last in line, and $u = w^* + x_m$ is the total amount of work (backlog).

We define the value function (without discounting) as the limiting expected cost difference between a system starting in state (u, w^*), and the system in steady-state:

$$v(u, w^*) \triangleq \mathbb{E}[R_L] + \mathbb{E}[\sum_{i=1}^{\infty} \left(W_L^{(i)} - \mathbb{E}[W_L]\right)], \tag{1}$$

where R_L denotes the remaining last-in-line time of the customer currently in that position, and $W_L^{(i)}$ denotes the time the i^{th} new customer will be last in line. Clearly, R_L depends only on w^*, whereas the $W_L^{(i)}$ depend only on u.

Proposition 1. *The mean time being last in line in the M/G/1 queue is*

$$\mathbb{E}[W_L] = \bar{c} = \frac{1}{\lambda} - \frac{1-\rho}{\lambda \widetilde{X}(\lambda)}, \tag{2}$$

and the corresponding value function satisfies

$$v(u, w^*) - v(0, 0) = \frac{1 - e^{-\lambda w^*}}{\lambda} + \frac{\lambda u + e^{-\lambda u} - 1}{\lambda \widetilde{X}(\lambda)}, \tag{3}$$

where $\widetilde{X}(s)$ denotes the Laplace-Stieltjes transform (LST) of the service time distribution, $\widetilde{X}(s) = \mathbb{E}[e^{-sX}]$.

Proof: First we introduce a slightly modified cost structure, where arriving customers *"pay a fee"* immediately upon arrival according to their *expected time* to be last in line. In this case, a sufficient state description is u, and the corresponding *"entrance fee"* is

$$c(u) = \mathbb{E}[W_L \mid U = u].$$

With Poisson arrival process, we can compute $c(u)$ exactly,

$$c(u) = \int_0^u t \cdot \lambda e^{-\lambda t}\, dt + e^{-\lambda u} u = \frac{1 - e^{-\lambda u}}{\lambda}. \tag{4}$$

This modification does not change the mean behavior, i.e., on average each customer incurs the same cost,

$$\mathbb{E}[W_L] = \mathbb{E}[\mathbb{E}[W_L \mid U]] = \mathbb{E}[c(U)].$$

However, in each sample path, customers may benefit or suffer from the modification depending on how "lucky" they are. As the customer currently last in line, if any, has already *"paid the entrance fee"*, the value function with the modified costs is

$$\tilde{v}(u) \triangleq \sum_{i=1}^{\infty} \left(\mathbb{E}[W_L^{(i)}] - \mathbb{E}[W_L]\right). \tag{5}$$

and a comparison with (1) shows that $v(u, w^*) = \mathbb{E}[R_L] + \tilde{v}(u)$. Using (4), we obtain an expression for the first term,

$$\mathbb{E}[R_L] = c(w^*) = \frac{1 - e^{-\lambda w^*}}{\lambda}. \tag{6}$$

It turns out that the cost $c(u)$ in (4) is essentially the same as (16) in [4], and substituting $s = \lambda$ into (29) and (30) in [4], yields expressions for the mean cost $\mathbb{E}[c(U)]$ and the value function $\tilde{v}(u)$; $\mathbb{E}[c(U)]$ is (2), and $\tilde{v}(u)$ is the second term in (3). □

For more details, see [4] and the references therein.

A queueing system is considered to be robust if its performance depends only on the mean values. For example, the mean sojourn time with the processor sharing (PS) is insensitive to the service time distribution. Unfortunately, (2) implies that with the last in line metric the situation is the opposite:

Corollary 1. *The mean time being last in line $\mathbb{E}[W_L]$ in the M/G/1 queue is not insensitive to the service time distribution.*

Example 1. *In the M/M/1 queue, $\widetilde{X}(s) = \mu/(\mu+s)$ and the mean cost rate is $\lambda\bar{c} = \rho^2$. This obviously is the same as the probability of having two or more customers in the M/M/1 queue.*

Let $\mathbf{z} = (u, w^*)$ denote the state of the queue, where u is the current backlog and w^* the remaining last-in-line time of the customer currently last in line (if any). The (marginal) admission cost of a customer with service time x is the expected increase in the infinite time-horizon costs,

$$a(\mathbf{z}, x) = v(u + x, u) - v(u, w^*),$$

which, with the last-in-line costs, reduces to

$$a(\mathbf{z}, x) = \frac{e^{-\lambda w^*} - e^{-\lambda u}}{\lambda} + \frac{\lambda x + e^{-\lambda u}(e^{-\lambda x} - 1)}{\lambda \widetilde{X}(\lambda)}. \tag{7}$$

Given no customer is waiting, $w^* = 0$. If the system is empty, then also $u = 0$.

Example 2. *The admission cost to the M/M/1 queue is*

$$a(\mathbf{z}, x) = \frac{e^{-\lambda w^*} - e^{-\lambda u}}{\lambda} + \frac{\lambda x + e^{-\lambda u}(e^{-\lambda x} - 1)}{\lambda}(1 + \rho),$$

whereas in the M/D/1 queue, $w^* = \max(0, u - d)$, and

$$a(u) = de^{\rho} + \frac{e^{-\lambda w^*} - e^{-\lambda(u-d)}}{\lambda}.$$

2.2. Customer-specific holding cost rates

Let us next consider the model with customer-specific holding cost rates H_i, which are i.i.d. random variables, $H_i \sim H$. Similarly as the service times, the holding cost rate becomes known upon arrival. That is, while customer i is waiting last in line, costs are incurred at rate h_i, and the customers $j = i+1, \ldots$ arriving in future have random i.i.d. cost rates $H_j \sim H$.

The mean cost per customer follows immediately from (2),

$$\bar{c} = \mathbb{E}[H] \cdot \mathbb{E}[W_L] = \frac{\mathbb{E}[H]}{\lambda}\left(1 - \frac{1-\rho}{\widetilde{X}(\lambda)}\right).$$

Let $\mathbf{z} = (u, w^*, h^*)$ denote the current state, where the starred quantities refer to the customer currently last in line. Then recall that the first term on the right-hand side of (3) corresponds to the customer currently last in line (if any), and the second term to the customers arriving in the future. Therefore, the value function with the holding costs satisfies

$$v(u, w^*, h^*) - v(0) = \frac{1 - e^{-\lambda w^*}}{\lambda} \cdot h^* + \frac{\lambda u + e^{-\lambda u} - 1}{\lambda \widetilde{X}(\lambda)} \mathbb{E}[H]. \quad (8)$$

The admission cost of a customer with service time x and holding cost rate h in state $\mathbf{z}$ is then

$$a(\mathbf{z}, x, h) = v(u + x, u, h) - v(u, w^*, h^*),$$

yielding

$$a(\mathbf{z}, x, h) = \frac{\left(1 - e^{-\lambda u}\right) h - \left(1 - e^{-\lambda w^*}\right) h^*}{\lambda} + \frac{\lambda x + e^{-\lambda u}(e^{-\lambda x} - 1)}{\lambda \widetilde{X}(\lambda)} \cdot \mathbb{E}[H]. \quad (9)$$

If no one is waiting last in line, then $w^* = h^* = 0$.

Interestingly, with some cost rates the admission cost can be negative. Informally, this means that an unhappy customer currently last in line gets replaced by someone else who does not mind as much for being the last.

2.3. Non-linear costs for being last in line

Next we assume that the cost for the last-in-line time is $(W_L)^k$, where k is some positive integer. Note that $k = 1$ corresponds to the unit cost rate (Section 2.1), whereas with $k \geq 2$, the *cost rate* increases in time to power of $k-1$. Such non-linear costs can be a useful model for customers who slowly become more anxious if they remain last in line for a longer period of time. In this case, the *expected cost* for joining the M/G/1 queue in state u is

$$c_k(u) = \mathbb{E}[(W_L)^k \mid U = u] = \int_0^u t^k \lambda e^{-\lambda t}\, dt + u^k e^{-\lambda u}.$$

Integration by parts yields an explicit expression,

$$c_k(u) = \frac{k!}{\lambda^k}\left(1 - e^{-\lambda u} \sum_{j=0}^{k-1} \frac{(\lambda u)^j}{j!}\right), \quad k = 1, 2, \ldots \quad (10)$$

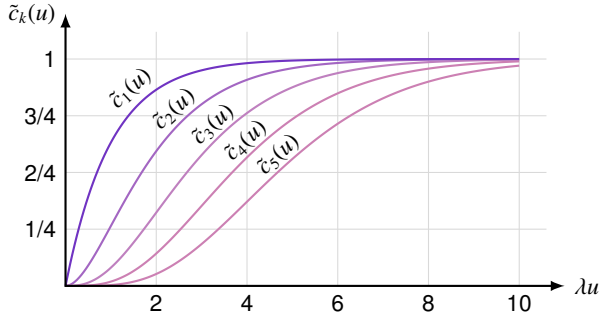

Figure 2: Expected normalized costs $\tilde{c}_k(u)$ for $k = 1, \ldots, 5$.

where λu corresponds to the mean number of customers arriving before the new customer enters service. Note that $c_k(u)$ is a strictly increasing and bounded, $\lim_{u\to\infty} c_k(u) = k!/\lambda^k$. For comparison, with the waiting time metric, $c_W(u) = u$, and the cost function is unbounded, $\lim_{u\to\infty} c_W(u) = \infty$.

For visualization purposes, we define normalized costs as,

$$\tilde{c}_k(u) \triangleq \frac{\lambda^k}{k!} c_k(u) = 1 - e^{-\lambda u} \sum_{j=0}^{k-1} \frac{(\lambda u)^j}{j!}. \quad (11)$$

Figure 2 depicts $\tilde{c}_k(u)$ for $k = 1, \ldots, 5$.

Proposition 2. *The k^{th} moment of the time being last in line in the M/G/1 queue is*

$$\mathbb{E}[(W_L)^k] = \bar{c}_k = \frac{k!}{\lambda^k}\left[1 - \sum_{j=0}^{k-1} \frac{(-\lambda)^j}{j!} \widetilde{W}^{(j)}(\lambda)\right], \quad (12)$$

where $\widetilde{W}(s)$ is the LST of the waiting time W.

Proof: Consider the M/G/1 queue subject to the modified costs $c_k(u)$. Due to PASTA, $\mathbb{E}[(W_L)^k] = \mathbb{E}[c_k(W)]$, and

$$\mathbb{E}[c_k(W)] = \mathbb{E}\left[\frac{k!}{\lambda^k}\left(1 - e^{-\lambda W} \sum_{j=0}^{k-1} \frac{(\lambda W)^j}{j!}\right)\right] = \frac{k!}{\lambda^k}\left[1 - \sum_{j=0}^{k-1} \frac{\lambda^j}{j!} \mathbb{E}[W^j e^{-\lambda W}]\right].$$

Then we recall that

$$\mathbb{E}[X^k e^{-sX}] = (-1)^k \widetilde{X}^{(k)}(s), \quad (13)$$

where $\widetilde{X}^{(k)}(s)$ denotes the k^{th} derivative of the LST of the random variable X. Therefore,

$$\mathbb{E}[(W_L)^k] = \mathbb{E}[c_k(W)] = \frac{k!}{\lambda^k}\left[1 - \sum_{j=0}^{k-1} \frac{(-\lambda)^j}{j!} \widetilde{W}^{(j)}(\lambda)\right].$$

□

Note that (12) reduces to (2) when $k = 1$, as expected.

Example 3. *The k^{th} moment of time being last in line in the M/M/1 queue is $\mathbb{E}[W_L^k] = \rho k!/\mu^k$. If considering the normalized costs* (11)*, then the mean cost is simply $\bar{c}_k = \rho^{k+1}$.*

The value function with the general costs depends also on the time a customer has already spent waiting last in line (age) denoted by Δ. In this case, a sufficient state description is the triple $\mathbf{z} = (u, w^*, \Delta)$, and the corresponding value function (without discounting) is

$$v_k(\mathbf{z}) \triangleq \mathbb{E}[(\Delta + R_L)^k - \Delta^k] + \sum_{j=1}^{\infty} \Big(\mathbb{E}[(W_L^{(j)})^k - \mathbb{E}[(W_L)^k]] \Big).$$

The first term corresponds to the mean cost the customer currently last in line will incur, and it depends on Δ and w^*. The summation corresponds to the mean costs customers arriving in the future will incur, which depends solely on u.

First we consider the modified cost structure where each customer pays according to their *expected costs* upon arrival.

Lemma 1. *The value function for the M/G/1 queue, where the cost incurred upon arrival is $c_k(u) = \mathbb{E}[(W_L)^k \mid U = u]$, satisfies*

$$\tilde{v}_k(u) - \tilde{v}_k(0) = \frac{uk!}{(1-\rho)\lambda^{k-1}} \sum_{j=0}^{k-1} \Big[\frac{(-\lambda)^j}{j!} \Big(\widetilde{W}^{(j)}(\lambda) - \sum_{v=0}^{j} \binom{j}{v} \widetilde{W}^{(j-v)}(\lambda) \widetilde{Y}^{(v)}(\lambda) \Big) \Big], \quad (14)$$

where $\widetilde{W}(s)$ is the LST of the waiting time in the M/G/1 queue and $\widetilde{Y}(s) = (1-e^{-su})/(su)$ is the LST of the uniformly distributed random variable $Y \sim U(0, u)$.

Proof: We utilize a general expression characterizing the value functions of the M/G/1 queue subject to arbitrary cost function $c(u)$ incurred upon arrival [4, Proposition 1],

$$\tilde{v}(u) - \tilde{v}(0) = \frac{\lambda u}{1-\rho} \mathbb{E}[c(W + Y) - c(W)],$$

where W denotes the waiting time in equilibrium, $Y \sim \mathrm{U}(0, u)$, and W and Y are independent. Substituting (10) gives

$$\tilde{v}_k(u) - \tilde{v}_k(0) = \frac{uk!}{(1-\rho)\lambda^{k-1}} \mathbb{E}\Big[e^{-\lambda W} \sum_{j=0}^{k-1} \frac{(\lambda W)^j}{j!} - e^{-\lambda(W+Y)} \sum_{j=0}^{k-1} \frac{(\lambda(W+Y))^j}{j!} \Big],$$

where the expectation can be written as

$$\mathbb{E}\Big[e^{-\lambda W} \sum_{j=0}^{k-1} \frac{\lambda^j}{j!} \left(W^j - e^{-\lambda Y} (W+Y)^j \right) \Big]. \quad (15)$$

After substituting (13) into (15) and utilizing the binomial theorem, the expectation reduces to

$$\sum_{j=0}^{k-1} \frac{(-\lambda)^j}{j!} \left(\widetilde{W}^{(j)}(\lambda) - \sum_{v=0}^{j} \binom{j}{v} \widetilde{W}^{(j-v)}(\lambda) \cdot \widetilde{Y}^{(v)}(\lambda) \right),$$

which completes the proof. □

Explicit expressions for value functions are obtained by substituting $\widetilde{W}^{(j)}(\lambda)$ and $\widetilde{Y}^{(v)}(\lambda)$ into (14). The Pollaczek-Khinchine transform formula for the waiting time,

$$\widetilde{W}(s) = \frac{s(1-\rho)}{s - \lambda(1 - \widetilde{X}(s))},$$

gives the former, and $\widetilde{Y}(s)$ is given in Lemma 1.

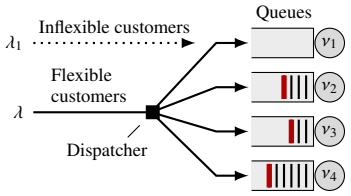

Figure 3: Parallel queues where customers being last in line incur costs.

Proposition 3. *The value function for the non-linear last-in-line costs of form W_L^k in the M/G/1 queue satisfies*

$$v(\mathbf{z}) = \left[\sum_{j=0}^{k} \binom{k}{j} \Delta^{k-j} c_j(w^*) - \Delta^k \right] + \tilde{v}_k(u), \quad (16)$$

where $\mathbf{z} = (u, w^*, \Delta)$ and $\tilde{v}_k(u)$ is given in Lemma 1.

Proof: The mean cost the customer currently last in line will incur is $\mathbb{E}[(\Delta + R_L)^k] - \Delta^k$, which explains the first term in (16). The second term is again equal to the value function with the modified cost structure, given in (14). □

The admission cost of a new customer with service time x is

$$a_k(\mathbf{z}, x) = v_k(\mathbf{z}') - v_k(\mathbf{z}), \quad (17)$$

where $\mathbf{z} = (u, w^*, \Delta)$ is the current state and $\mathbf{z}' = (u + x, u, 0)$ the new state. Similarly as in Section 2.2, it is straightforward to generalize these results with customer-specific coefficients H_i. In fact, even the form of holding cost structure (e.g., linear or quadratic) may depend on the customer or his class.

3. Model for Dispatching System

In this section, we consider a system where the service is provided by a set of parallel servers each having their own queue (cf. queues at supermarkets, road toll stations, etc.). The model is depicted in Figure 3 and is as follows:

- The system comprises K parallel FCFS servers with service rates $\nu_1, \ldots, \nu_K$.
- Customers arrive according to a Poisson process with rate λ, their service demands are i.i.d. with a general distribution, $X_i \sim X$, and they are routed immediately upon arrival to one of the servers. Service time of customer i in server j would be X_i/ν_j.
- Customers incur costs at customer-specific cost rates, $H_i \sim H$, while being last in line.
- Customers' service demands and cost rates (e.g. service class) are observed upon arrival, and the dispatching decision can utilize this information.

In general case, each server j may additionally have their own stream of inflexible customers (see λ_1 in Figure 3). Their service demands and cost rates are also i.i.d., but may obey different distributions than those of flexible customers. Moreover, costs incurred may be non-linear functions of the time spent as the last in line, as discussed in the previous section.

3.1. Last-Place-Aversive Dispatching

Next we utilize the admission costs (7), (9) and (17) to develop efficient, state-aware, dispatching policies that aim to minimize the costs due to being last in line.

First we note that our cost structure is prone to instability issues. To illustrate this, suppose we have two servers, moderately high load, and two types of customers. A large fraction of customers have a very low cost rate h_1, and a small fraction has a very high cost rate h_2. In this case, no matter how long one queue is, it seems intuitively beneficial to still route all *"cheap customers"* to the long queue as the cost rate remains at h_1. The other queue is then dedicated to the *"valuable customers"* who basically experience no queueing, and thus the mean cost rate is about h_1 even though one queue is unstable.

This behavior is an artifact of our simplistic cost structure that only cares about who is last in line. The obvious fix is to include another term to the cost structure, such as the mean sojourn time, to penalize for the excessively long queues. The corresponding value functions can be found, e.g., from [5] and [6], where in latter servers may also have the so-called setup delay when a new busy period starts. The value function for the M/G/1 queue, with customer-specific holding costs H_t for the sojourn time, satisfies

$$v_t(u) - v_t(0) = \frac{\lambda u^2}{2(1-\rho)}\,\mathbb{E}[H_t]. \tag{18}$$

The corresponding admission cost is

$$a_t(u, x, h_t) = h_t(u + x) + \frac{\lambda(2ux + x^2)}{2(1-\rho)}\,\mathbb{E}[H_t], \tag{19}$$

where x and h_t denote the service time and the holding cost of the new customer. The value function for the waiting time (before the service starts) is essentially the same, whereas the admission cost is

$$a_w(u, x, h_w) = h_w u + \frac{\lambda(2ux + x^2)}{2(1-\rho)}\,\mathbb{E}[H_w], \tag{20}$$

where h_w and H_w correspond to holding costs while waiting in the queue. We could also define separate holding cost rates for the three stages: (i) when waiting last in line, (ii) when waiting otherwise, and (iii) when in service. When a cost structure consists of several terms, the corresponding mean costs, value functions and admission costs are simply added together.

The standard MDP procedure to improve any given (static) policy is as follows (see, e.g., [7]):

1. Assume a static basic dispatching policy α_0, where the dispatching decision may depend only on the customer itself (e.g., its class, size or holding cost).
2. With α_0, the system decomposes into K independent parallel M/G/1 queues, and the value function of the whole system is the sum of the queue-specific value functions,

   $$v(\mathbf{z}) = \sum_{i=1}^{K} v^{(i)}(\mathbf{z}_i),$$

   where $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_K)$ defines the state of each server.
3. The first policy iteration step (FPI) yields a new (dynamic) policy, which reduces to choosing the server with the smallest admission cost

   $$\alpha(\mathbf{z}, \mathbf{x}) = \operatorname*{argmin}_i \{a^{(i)}(\mathbf{z}_i, \mathbf{x})\},$$

   where $\mathbf{x}$ defines the service demand x and possible holding costs and other parameters of the customer, and $a^{(i)}(\mathbf{z}_i, \mathbf{x})$ is the admission cost to server i. Ties are resolved, e.g., at random.

4. Numerical Example

In this section, we compare the FPI-based policies to some well-known heuristics. Suppose we have two identical servers, $\nu_1 = \nu_2 = 1$, the service demands are exponentially distributed, $X_i \sim \mathrm{Exp}(\mu)$, and also cost rates are identical, $H = 1$.

The heuristic reference dispatching policies are:

1. *Random split (RND)*, which routes customers uniformly at random to both queues.
2. *Join-the-shortest-queue (JSQ)*, choosing the queue with the least number of customers. JSQ is optimal with respect to the sojourn time if the available information is just the queue length [8]. It is also the individually optimal policy with the given information (for sojourn time).
3. *Least-work-left (LWL)* observes the current backlogs and chooses the queue that minimizes the customer's waiting (and sojourn) time. Hence, LWL is the individually optimal policy with respect to the sojourn time.

The above are compared against two new FPI-based policies that explicitly try to minimize the costs due to being last in line. For simplicity, the basic policy in both cases is RND.

4. *Last-place-aversive (LPA)* policy ignores the queue lengths and routes customers according to (7),

   $$\alpha_{\texttt{LPA}}(\mathbf{z}, x) = \operatorname*{argmin}_i \{a^{(i)}(\mathbf{z}_i, x)\}.$$

5. *Stabilized LPA (sLPA)* policy includes also a term for the sojourn time with a fixed holding cost rate $H_t = \beta$,

   $$\alpha_{\texttt{sLPA}}(\mathbf{z}, x) = \operatorname*{argmin}_i \left\{a^{(i)}(\mathbf{z}_i, x) + a_t^{(i)}(u_i, x, \beta)\right\},$$

   where we have chosen to use $\beta = 0.05$ (same for all customers).

In all cases, possible ties are resolved in favor of the first server.

Figure 4 depicts the simulation results. The left figure shows the absolute performance in terms of the mean cost per customer $\bar{c}$ (i.e., the mean time being last in line, $\mathbb{E}[W_L]$) as a function of the offered load ρ. The middle figure shows the relative performance to RND, $\bar{c}_\alpha/\bar{c}_{\texttt{RND}}$. The figure on the right depicts the mean sojourn time multiplied by $(1-\rho)$.

First, the mean cost with RND is a straight line in accordance with Example 1. Similarly, the scaled sojourn time, $(1-\rho)\mathbb{E}[T]$, is a constant (two M/M/1 queues). The last-in-line performance

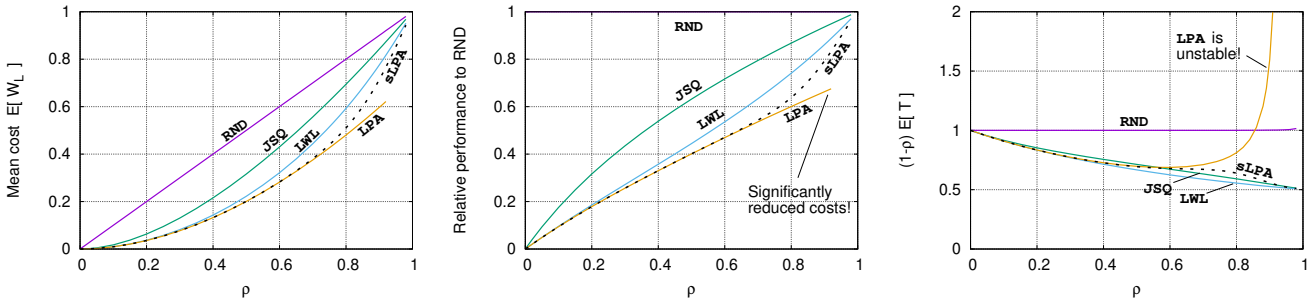

Figure 4: Simulation results with RND, JSQ, LWL and Last-place-aversive FPI policies LPA and sLPA. The left figure shows the absolute performance in terms of the mean cost per customer, the middle figure depicts the relative performance to RND, and the figure on the right depicts the mean sojourn time. Note that LPA becomes unstable, unless the sojourn time (or waiting time) is included to the objective (dotted line, corresponding to sLPA).

with JSQ and LWL is clearly better than with RND when the load is low, but becomes similar under heavy load. In contrast, LPA yields significantly lower costs than any other policy especially when the offered load increases. However, as predicted, it becomes unstable as ρ increases, somewhere about $\rho = 0.9$ in this case. The stabilized variant, sLPA, keeps the mean sojourn time under control while reducing the last-in-line costs at the same time. The exact behavior can be controlled by adjusting the weight parameter β.

5. Conclusions

Humans experience waiting in line in bizarre ways. One peculiarity is that we dislike being the last customer in line. In this paper, we considered the standard M/G/1 queue subject to a cost structure that models this behavior explicitly. In particular, we defined that a customer incurs costs until the next customer joins the queue behind him, or he enters the service. In the general case, the cost rate can be customer specific and increase in time. Thus, even though the queueing discipline is FCFS, the actual costs a customer incurs depend explicitly on what happens in the future. In this sense, the system is similar to the last-come-first-served (LCFS) and processor sharing (PS) systems, where customers arriving in future affect the sojourn time of the present customers.

We analyzed the last-in-line queueing models in the MDP framework and derived exact expressions for the mean costs, value functions and admission costs. These results were then applied to develop the last-place-aversive routing (LPA) policies that aim to minimize the discomfort due to being last in line in the setting of parallel servers. By means of simulations, we showed that a central dispatcher can effectively reduce the discomfort the customers experience due to being the last in line.

Acknowledgements

This work was supported by the Academy of Finland in the FQ4BD project (grant no. 296206) and by the University of Iceland Research Fund in the RL-STAR project.

[1] R. C. Larson, “OR forum – perspectives on queues: Social justice and the psychology of queueing,” *Operations Research*, vol. 35, no. 6, pp. 895–905, 1987.

[2] L. Schrage, “A proof of the optimality of the shortest remaining processing time discipline,” *Operations Research*, vol. 16, no. 3, 1968.

[3] I. Kuziemko, R. W. Buell, T. Reich, and M. I. Norton, ““last-place aversion”: Evidence and redistributive implications,” *The Quarterly Journal of Economics*, vol. 129, no. 1, pp. 105–149, 2014.

[4] E. Hyytiä, R. Righter, J. Virtamo, and L. Viitasaari, “Value (generating) functions for the M^X/G/1 queue,” in *29th International Teletraffic Congress (ITC’29)*, Genoa, Italy, Sep. 2017.

[5] E. Hyytiä, A. Penttinen, and S. Aalto, “Size- and state-aware dispatching problem with queue-specific job sizes,” *European Journal of Operational Research*, vol. 217, no. 2, pp. 357–370, Mar. 2012.

[6] E. Hyytiä, R. Righter, and S. Aalto, “Task assignment in a heterogeneous server farm with switching delays and general energy-aware cost structure,” *Performance Evaluation*, vol. 75–76, no. 0, pp. 17–35, May-June 2014.

[7] P. Whittle, *Optimal Control: Basics and Beyond*. Wiley, 1996.

[8] W. Winston, “Optimality of the shortest line discipline,” *Journal of Applied Probability*, vol. 14, pp. 181–189, 1977.