Operations Research Letters 44 (2016) 839–845

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Unintended consequences of optimizing a queue discipline for a service level defined by a percentile of the waiting time

Benjamin Legros

PSB Paris School of Business, Department of Economics, 59 rue Nationale, 75013 Paris, France

ARTICLE INFO

Article history:
Received 23 June 2016
Received in revised form
20 October 2016
Accepted 21 October 2016
Available online 27 October 2016

Keywords:
Queueing systems
Waiting time
Staffing
Queue discipline
Priority

ABSTRACT

In service systems, the service level is often represented by a percentile of the waiting time. This creates an incentive to optimize the queue discipline. For this purpose, in an M/M/s queue setting, we prove that the optimal discipline gives priority to the oldest customer who has waited less than the acceptable waiting time. Next, we derive explicitly the performance measures. Finally, we show that although this discipline may reduce staffing costs, it leads to excessive wait for non-prioritized customers.

1. Introduction

Context and Motivation. In numerous service system, the management is interested in representing the information on waiting times by a single number to facilitate comparisons. A percentile of the waiting time is the typically chosen in this purpose. This metric is often preferred to the average speed of answer (ASA) because the former was perceived to be more informative; see [1]. In particular, the ASA does not take into account the variability of the waiting time.

However, measuring the service level by the percentage of customers that has to wait longer than a specified amount of time (SLP) has also disadvantages. First, this metric gives no information on how long customers who have exceeded the acceptable waiting time (AWT) still have to wait. Second, it provides an incentive to managers to give priority to customers who have not yet reached the AWT, thereby increasing even more the waiting time of customers that have waited longer than the AWT.

The fact that system operators may attempt to optimize wait time percentiles is, in many situations, an unintended consequence that was not anticipated by those who proposed using the percentiles as performance measures. It is thus interesting to study policies that optimize the SLP in order to better understand the consequences of such a "rational" decision. Already, [5] illustrates numerically that optimizing the SLP is a bad choice for most of the other service level measures in a setting where customers who have waited more than AWT are dropped. [6], in a call center context with contract and non-contract customers, also show that the delay percentile used in practice results in long delays and high coefficients of variations compared to what they might have achieved under a first-come-first-served policy. We aim in this paper to further investigate the consequences of this managerial decision.

A very simple way to minimize the SLP is to optimize the queueing discipline. Changing the queueing discipline is attractive since it has no impact on the ASA, and does not force radical rejection decisions which could be badly perceived. However, as pointed out by [5], this may have bad consequences. The aim of this paper is to quantify and evaluate these consequences. It is interesting to determine (i) how much the SLP can be improved when changing the queue discipline, (ii) how the staffing decisions may be impacted and (iii) how bad can be the service level deterioration for non-prioritized customers.

Contributions. We propose in this paper to reconsider the M/M/s queue for which we optimize the queueing discipline to an objective of minimizing the SLP. We prove in Section 2 that the optimal policy gives a priority to the oldest customer who has waited less than the AWT. The proposed discipline is intuitive and has already be mentioned by [5] but, to the best of our knowledge, it has not been evaluated in the queueing literature.

The main objective is to determine the proportion of customers who have waited less than AWT time units, $P(W < \text{AWT})$ and compare this metric to what can be found with a FCFS discipline.

E-mail address: benjamin.legros@centraliens.net.

http://dx.doi.org/10.1016/j.orl.2016.10.011
0167-6377/© 2016 Elsevier B.V. All rights reserved.

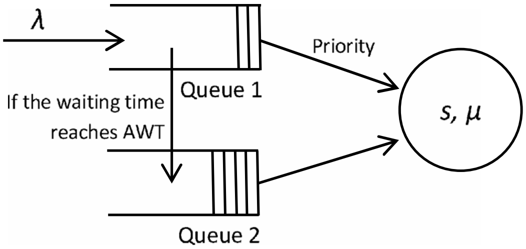

Fig. 1. Equivalent model for the MPW queue.

In order to differentiate between prioritized and non-prioritized customers, we are also interested in the conditional expected waiting times $E(W|W < \text{AWT})$ and $E(W|W > \text{AWT})$ and by the average excess waiting time $E((W - \text{AWT})^+)$. Closed-form expressions of these performance measures are derived in Section 3. The difficulty to compute these metrics is that the decision to change a high priority customer into a low priority one does not depend on a classical state definition like the number of high priority customers but on the experienced waiting time of a given customer. The solution to overcome this difficulty is to use the discretized waiting time of the first high priority customer in line as a state definition.

In Section 4, we evaluate the consequences of giving a priority to customers who have waited less than the AWT. The expected consequence is that this new policy improves SLP especially in congested situations. It may also lead to cheap staffing solutions. We show in particular that above a threshold on the traffic intensity, no safety staffing is required. Yet, this discipline strongly deteriorates the waiting time of non-prioritized customers. This unwanted consequence is also significant in congested situations.

2. Optimal discipline and setting

We consider a multi-server single queue with s identical, parallel servers. The arrival process of customers is Poisson with rate λ. Service times are independent and exponentially distributed with rate μ. To ensure stability, we assume $\lambda < s\mu$. Our queuing model only differs from the classical M/M/s queue by the queue discipline. The chosen queue discipline minimizes the SLP among all non-preemptive, work-conserving policies. It is defined as follows.

- A strict non-preemptive priority is given to customers who have waited less than AWT.
- The discipline for prioritized customers is FCFS.
- The discipline for non-prioritized customers is arbitrary.

The optimality of this policy is proven in Theorem 1 using sample path arguments.

We name this discipline the MPW discipline (Minimized Percentile of the Waiting time). The M/M/s queue under MPW discipline is equivalent to a particular V-queueing model with two queues; Queue 1 and Queue 2, where customers in Queue 1 have a non-preemptive priority over customers in Queue 2. The arrival process in Queue 1 is Poisson with parameter λ and the arrival process in Queue 2 is generated by customers in Queue 1 who have waited exactly AWT time units without being served. This equivalent queueing model is depicted in Fig. 1.

Theorem 1. *In order to minimize SLP, it is optimal to give priority to the first customer in line in Queue* 1.

Proof. We prove this result by considering a fixed sample path of the stochastic process. This sample path is determined by arrival instants, departure instants, and service initiation instants. Since customers in Queue 1 and in Queue 2 have the same service time distribution, we can assume that the service times are only determined by the order of service initiations. In the long-run, this is equivalent to considering that the service times are determined by customers; e.g., see [2]. Therefore an interchange for the order of service of two customers does not affect the event epochs.

Consider an arbitrary policy π. Suppose that at time t_1, under policy π, a server becomes free and selects a Type 2 customer as the next one to serve, even though there is a Type 1 customer in Queue 1 who has waited w time units so far. Due to work-conservation, there will be a later time instant, say t_2, where the initially considered Type 1 customer will be scheduled in service. At this instant t_2 either this initial Type 1 customer is still a Type 1 customer if $w + t_2 - t_1 \leq \text{AWT}$ or this initial Type 1 customer has changed into a Type 2 customer.

Now consider the policy π' which follows all actions of π except that it schedules a Type 1 customer at t_1 and a Type 2 customer at t_2. The total number of customers who enter service before t_2 is equal under both policies. However, the number of Type 2 customers who enter service before t_2 is higher (if $w + t_2 - t_1 > \text{AWT}$) or equal under π. This proves that a priority should be given for Queue 1 customers. To prove that FCFS in Queue 1 is optimal, the same approach can be applied. □

3. Performance analysis

We use a non-traditional approach for the modeling of Queue 1, as proposed in [4]. The idea is to discretize the waiting time of the first customer in line (FIL) by a succession of exponential phases with rate γ per phase instead of using the traditional definition of the number of customers in the queue. The maximal number of possible waiting phases in Queue 1 is denoted by n. After leaving this last waiting phase a customer – if not served – is routed to Queue 2. This modeling is an approximation of the real system.

State definition. The system is modeled using a two dimensional continuous-time Markov chain. We denote by (x, y) a state of the system for $-s \leq x \leq n$ and $y \geq 0$, where x represents the servers state or the waiting time in Queue 1 and y represents the number of customers in Queue 2. More precisely, states with $-s \leq x \leq 0$ correspond to an empty Queue 1 and $s + x$ busy agents. States with $0 < x \leq n$ correspond to the phase at which the FIL in Queue 1 is waiting and all agents are busy. Lumping together the states representing free servers and the waiting time of the FIL in Queue 1 in one dimension can be done as servers cannot be free while customers are waiting. Note that the number of customers in Queue 1 is not used in the state definition. Yet, the method proposed by [4] allows us to obtain the distribution of the queue length using the waiting phase of the FIL.

Transitions. The transition rate diagram is depicted in Fig. 2. We next describe the 6 possible transitions in the Markov chain. When the FIL changes, because of a service completion (see transition Type 4) or because of the current FIL moving to Queue 2 (see transition Type 6), the waiting time phase changes from $x > 0$ to $x - h$ with probability $q_{x,x-h}$, where $q_{x,x-h} = \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h$ for $0 \leq h < x$ and $q_{x,0} = \left(\frac{\gamma}{\lambda+\gamma}\right)^x$, see [4].

1. An arrival with rate λ while Queue 1 is empty ($-s \leq x \leq 0, y \geq 0$), which changes the state to $(x + 1, y)$. If $-s \leq x < 0$ and $y = 0$, then the number of busy servers is increased by 1. If $x = 0$ and $y \geq 0$, then the FIL entity is created.
2. A service completion with rate $(s + x)\mu$ while queues 1 and 2 are empty ($-s < x \leq 0, y = 0$), which changes the state to $(x - 1, y)$. The number of busy servers is reduced by 1.