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

Article *in* Operations Research Letters · October 2016

DOI: 10.1016/j.orl.2016.10.011

Benjamin Legros
Ecole Centrale Paris

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

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.

Keywords: queueing systems, Markov chains, performance evaluation, waiting time, staffing, queue discipline, priority.

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 that 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. 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 Figure 1.

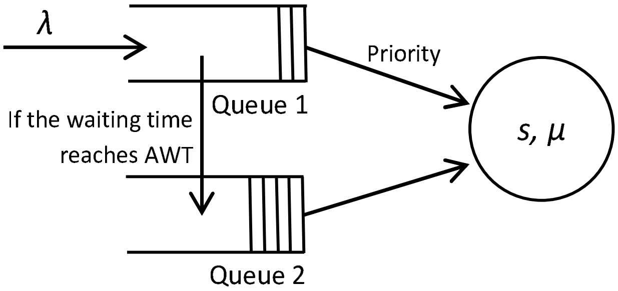

Figure 1: Equivalent Model for the MPW queue

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 $w + t_2 - t_1 >$

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 Figure 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.

3. A service completion with rate $s\mu$ while Queue 1 is empty, Queue 2 is not empty and all servers are busy $(x = 0,\ y \geq 1)$, which changes the state to $(0, y - 1)$. The number of customers in Queue 2 is reduced by 1.

4. A service completion with rate $s\mu q_{x,x-h}$ while Queue 1 is not empty $(0 < x \leq n, y \geq 0)$, which changes the state to $(x - h, y)$, that is, the new FIL is in waiting phase $x - h$.

5. A phase increase with rate γ while Queue 1 is not empty and the FIL is not in waiting phase n $(0 < x < n, y \geq 0)$, which changes the state to $(x+1, y)$. The waiting phase of the FIL is increased by 1.

6. A phase increase with rate $\gamma q_{x,x-h}$ while the FIL in Queue 1 is in waiting phase n $(x = n, y \geq 0)$, which changes the state to $(x-h, y+1)$, that is, the new FIL is in waiting phase $x-h$ and the number of customers in Queue 2 is increased by 1.

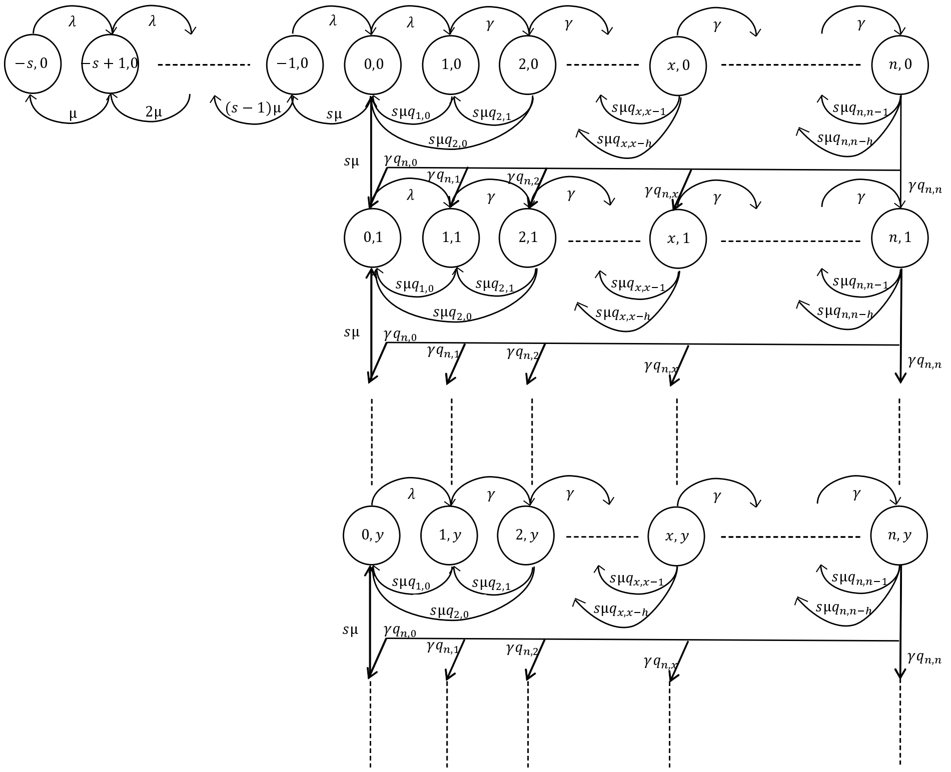

Figure 2: Transition rate diagram

Convergence to the real system. The first approximation is the waiting time in Queue 1. We choose n and γ such that $\frac{n}{\gamma} \triangleq \text{AWT}$. Therefore the maximal time spent in Queue 1 follows an Erlang distribution with n phases and rate n/AWT per phase. This ensures that as n and γ go to infinity, this random variable converges to the deterministic duration, AWT. The second approximation is the transition from Queue 1 to Queue 2. It is assumed in our modeling that after one γ-transition from state $x = n$ only one customer is routed to Queue 2. However, more than one customer could be in phase n (as in any other phase). More precisely, given that one customer is in phase n, this customer is the only one with probability $\frac{\gamma}{\lambda+\gamma}$, or two customers or more are in phase n with probability $\frac{\lambda}{\lambda+\gamma}$. Again, as γ tends to infinity, the probability that only one customer is in one phase is equal to one.

3.1 Stationary probabilities

Let us introduce the notations $a = \frac{\lambda}{\mu}$ and $a_\gamma = s \cdot \frac{a+\gamma/\mu}{s+\gamma/\mu}$. The ratio a represents the traffic intensity of the system and a_γ is a modified version of the traffic intensity. The parameter a_γ is an increasing function of γ which is equal to a for $\gamma = 0$ and equal to s for $\gamma = \infty$. Proposition 1 gives the stationary probability $p_{x,y}$ to be in state (x, y) for $-s \leq x \leq n$ and $y \geq 0$.

Proposition 1

$$\begin{aligned}
p_{-s,0} &= \left[\sum_{x=0}^{s-1} \frac{a^x}{x!} + \frac{a^s}{s!} \frac{\left(1 + \frac{a}{s}\frac{\lambda}{\gamma} - \frac{a}{s}\left(1 + \frac{\lambda}{\gamma}\right)\left(\frac{a_\gamma}{s}\right)^n\right)}{(1-a/s)\left(1 - \frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n\right)} \right]^{-1}, \\
p_{x-s,0} &= \frac{a^x}{x!} \cdot p_{-s,0}, \; \textit{for } 0 \leq x \leq s, \\
p_{x,0} &= \frac{\lambda}{\gamma} \cdot \frac{\left(\frac{s}{a_\gamma}\right)^{n-x} - \frac{a}{s}}{\left(\frac{s}{a_\gamma}\right)^{n} - \frac{a}{s}} \cdot p_{0,0}, \; \textit{for } 1 \leq x \leq n, \\
p_{x,y} &= \frac{\lambda}{\gamma}\left(\frac{a}{s}\right)^y \cdot \frac{1 - \frac{a}{s}}{\left(\frac{s}{a_\gamma}\right)^{n} - \frac{a}{s}} \cdot p_{0,0}, \; \textit{for } 1 \leq x \leq n, y \geq 1, \\
p_{0,y} &= \left(\frac{a}{s}\right)^y \cdot \frac{1 - \frac{a}{s}}{\left(\frac{s}{a_\gamma}\right)^{n} - \frac{a}{s}} \cdot p_{0,0}, \; \textit{for } y \geq 1.
\end{aligned}$$

Proof. We adopt the following approach to derive the stationary probabilities. First, we determine a set of equilibrium equations. Next, using these equilibrium equations we derive a simple explicit expression of the probability that the FIL in Queue 1 is in waiting phase x; $p_x = \sum_{y=0}^{\infty} p_{x,y}$ for $0 \leq x \leq n$. Considering this probability leads to a one-dimensional problem which in turn allows us to compute the probability of an empty system using the normalizing condition. The other stationary probabilities can be derived with a similar approach.

Equilibrium equations. Let S be the state space. Consider the cut between $A_1 = \{(-s,0), \cdots, (x,0)\}$ and $S\backslash A_1$, where $-s \leq x < n$. Observing that $\left(\frac{\gamma}{\lambda+\gamma}\right)^x + \sum_{l=h}^{x-1}\left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^l = \left(\frac{\gamma}{\lambda+\gamma}\right)^h$, we deduce that the cumulative transition rate from state (x, y) to states $(0, y), (1, y) \cdots (x-h, y)$ is $s\mu\left(\frac{\gamma}{\lambda+\gamma}\right)^h$, for $0 \leq h < x < n$ and $y \geq 0$. Therefore, by equating flows across the cut, one may write

$$\lambda p_{x,0} = (s + x + 1)\mu p_{x+1,0}, \text{ for } -s \leq x < 0, \tag{1}$$

$$\lambda p_{0,0} = s\mu p_{0,1} + s\mu \sum_{i=1}^{n} p_{i,0}\left(\frac{\gamma}{\lambda+\gamma}\right)^i, \tag{2}$$

$$\gamma p_{x,0} = s\mu p_{0,1} + s\mu \sum_{i=x+1}^{n} p_{i,0}\left(\frac{\gamma}{\lambda+\gamma}\right)^{i-x}, \text{ for } 0 < x < n. \tag{3}$$

Consider now the cut between $A_2 = \{(x, y') : y' \leq y\}$ and $S \backslash A_2$, where $y \geq 0$. This leads to

$$\gamma p_{n,y} = s\mu p_{0,y+1}, \text{ for } y \geq 0. \tag{4}$$

Finally, from the cut between $A_3 = \{(0, y), (1, y), \cdots (x, y)\}$ and $S \backslash A_3$, where $0 \leq x < n$ and $y \geq 1$, we get

$$(s\mu + \lambda)p_{0,y} = s\mu p_{0,y+1} + s\mu \sum_{i=1}^{n} p_{i,y} \left(\frac{\gamma}{\lambda + \gamma}\right)^{i} + \gamma \left(\frac{\gamma}{\lambda + \gamma}\right)^{n} p_{n,y-1}, \text{ for } y \geq 1, \tag{5}$$

$$\gamma p_{x,y} + s\mu p_{0,y} = s\mu p_{0,y+1} + s\mu \sum_{i=x+1}^{n} p_{i,y} \left(\frac{\gamma}{\lambda + \gamma}\right)^{i-x} + \gamma \left(\frac{\gamma}{\lambda + \gamma}\right)^{n-x} p_{n,y-1}, \text{ for } 0 < x < n \text{ and } y \geq 1, \tag{6}$$

since the cumulative transition rate from state $(n, y-1)$ to states $(0, y), (1, y), \cdots, (n-h, y)$ is $\gamma \left(\frac{\gamma}{\lambda + \gamma}\right)^{h}$, for $0 \leq h < n$ and $y \geq 1$.

Probability of an empty system. Recall that p_x is the probability that the FIL in Queue 1 is in waiting phase x; $p_x = \sum_{y=0}^{\infty} p_{x,y}$ for $0 \leq x \leq n$. One can prove by induction on x that $p_{n-x} = \left(\frac{s+\gamma/\mu}{a+\gamma/\mu}\right)^{x} p_n = \left(\frac{s}{a_\gamma}\right)^{x} p_n$, for $0 \leq x < n$. Using Equation (5), we deduce that $p_0 = \frac{\gamma}{\lambda}\left(\frac{s}{a_\gamma}\right)^{n} p_n$, therefore $p_x = \frac{\lambda}{\gamma}\left(\frac{a_\gamma}{s}\right)^{x} p_0$ for $1 \leq x \leq n$. Moreover, summing up Equations (4) for $y \geq 0$, we get $s\mu(p_0 - p_{0,0}) = \gamma p_n$. Since $p_n = \frac{\lambda}{\gamma}\left(\frac{a_\gamma}{s}\right)^{n} p_0$, $p_0 = \frac{p_{0,0}}{1 - \frac{a}{s}\left(\frac{a_\gamma}{s}\right)^{n}}$. Using now Equation (1), we finally deduce that $p_0 = \frac{\frac{a^s}{s!} p_{-s,0}}{1 - \frac{a}{s}\left(\frac{a_\gamma}{s}\right)^{n}}$. Note that this expression will be used in the proof of Theorem 2 to derive the expected waiting time in Queue 1. Using the fact that the overall sum of the stationary probabilities is equal to one, we obtain the probability of an empty system as in Proposition 1. □

3.2 $P(W <$ AWT), $E(W|W <$ AWT) and $E(W|W >$ AWT)

In Theorem 2, we derive the performance measures under the MPW discipline. In order to relate the performance measures under the MPW discipline and the FCFS discipline we introduce the notation $C(s, a) = P(W > 0)$ (i.e., probability of queueing). This probability is identical in the FCFS discipline and in the MPW discipline due to work-conservation. Recall from [3] page 103 that $C(s, a) = \frac{\frac{a^s}{s!}}{\sum_{x=0}^{s-1} \frac{a^x}{x!} + \frac{a^s}{s!} \frac{1}{1-a/s}} \cdot \frac{1}{1-a/s}$.

Theorem 2 We have

$$SLP = P(W > \text{AWT}) = C(s, a) \cdot \frac{\left(1 - \frac{a}{s}\right) e^{-s\mu(1-a/s)\cdot\text{AWT}}}{1 - \frac{a}{s} e^{-s\mu(1-a/s)\cdot\text{AWT}}},$$

$$E(W|W < \text{AWT}) = \frac{\frac{a^s}{s!}}{s\mu} \cdot \frac{1 - e^{-s\mu(1-a/s)\cdot\text{AWT}}(1 + s\mu(1 - a/s) \cdot \text{AWT})}{(1 - a/s)^2 \left(\left(1 - \frac{a}{s} e^{-s\mu(1-a/s)\cdot\text{AWT}}\right) \sum_{x=0}^{s-1} \frac{a^x}{x!} + \frac{a^s}{s!} \frac{1 - e^{-s\mu(1-a/s)\cdot\text{AWT}}}{1 - a/s}\right)},$$

$$E(W|W > \text{AWT}) = \frac{1 + s\mu \cdot \text{AWT}}{s\mu(1 - a/s)}.$$

Proof. The approach to derive the performance measures first consists of defining the embedded Markov chain at specific instants chosen in order to reach the performance measures at arbitrary instants. Next, by letting γ and n tend to infinity we obtain the results.

The embedded Markov chain. Arriving customers either enter service upon arrival, enter service from Queue 1 after some wait, or are routed to Queue 2. Call the instants when one of these three events occurs Q-instants. Since the events at Q-instants all occur one at a time, in the long-run the system is identical at arrival instants and Q-instants. Since the Poisson arrival process of customers is independent of the system state, the system is identical at arrival instants and arbitrary instants. So, the system is also identical at arbitrary instants and Q-instants. We therefore choose to consider the system at Q-instants to obtain the performance measures (the arrival instants cannot be seen in our Markov chain).

The Q-instants are determined by λ-transitions from state with a vacant server (transition Type 1), $s\mu$-transitions from the other states except in states $(0, y)$ (transition Type 4) and γ-transitions from states (n, y) (transition Type 6), for $y \geq 0$. The overall customer flow at Q-instants is identical to the customer flow at arrival instants and has a rate λ. Therefore, the probability at Q-instants that x servers are busy for $0 \leq x < s$ is $\frac{\lambda}{\lambda} p_{-s+x,0} = p_{-s+x,0}$. The probability that the FIL is in waiting phase x and y customers are in Queue 2 is $\frac{s\mu}{\lambda} p_{x,y}$ for $0 < x < n$, 0 for $x = 0$ and $\frac{s\mu+\gamma}{\lambda} p_{n,y}$ for $x = n$. The stationary probabilities at Q-instants are then completely known. This allows us to derive the performance measures.

Performance measures. The approach to obtain the performance measures is to let γ and n tend to infinity with respect to $\frac{n}{\gamma}$ =AWT.

We first derive the proportion of customers who are routed to Queue 2. A customer moves from Queue 1 to Queue 2 due to a γ-transition from states (n, y) (transition Type 6), $y \geq 0$. The proportion of customers which are moved from Queue 1 to Queue 2 is therefore

$$P(W > \text{AWT}) = \lim_{n,\gamma \to \infty} \frac{\gamma}{\lambda} p_n.$$

Recall from the proof of Proposition 1 that $p_n = \frac{\lambda}{\gamma} \left(\frac{a_\gamma}{s}\right)^n p_0$ and $p_0 = \frac{\frac{a^s}{s!} p_{-s,0}}{1 - \frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n}$. Therefore,

$$\frac{\gamma}{\lambda} p_n = \frac{\frac{a^s}{s!} \left(\frac{a_\gamma}{s}\right)^n p_{-s,0}}{1 - \frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n}. \tag{7}$$

From the expression of $p_{-s,0}$ in Proposition 1 and using $\lim_{n,\gamma\to\infty} \left(\frac{a_\gamma}{s}\right)^{\gamma \cdot AWT} = e^{-s\mu(1-a/s)\cdot \text{AWT}}$, we get the probability of an empty system in an M/M/s queue:

$$\lim_{n,\gamma \to \infty} p_{-s,0} = \left[\sum_{x=0}^{s-1} \frac{a^x}{x!} + \frac{a^s}{s!} \frac{1}{1 - a/s}\right]^{-1}. \tag{8}$$

By applying the last result in Equation (7), we obtain the explicit expression of $P(W > \text{AWT})$.

Consider now the served customers from Queue 1. A served customer from Queue 1 waits x γ-phases with probability $\frac{s\mu}{\lambda}p_x$ for $0 < x \leq n$ and each phase has an expected duration of $1/\gamma$. Therefore,

$$\begin{aligned}
P(W < \text{AWT}) \cdot E(W|W < \text{AWT}) &= \lim_{n,\gamma\to\infty} \frac{s\mu}{\lambda}\sum_{x=1}^{n}\frac{x}{\gamma}p_x \\
&= \lim_{n,\gamma\to\infty} p_0\frac{s\mu}{\gamma^2}\frac{a_\gamma}{s}\frac{-(n+1)\left(1-\frac{a_\gamma}{s}\right)\left(\frac{a_\gamma}{s}\right)^n+1-\left(\frac{a_\gamma}{s}\right)^n}{\left(1-\frac{a_\gamma}{s}\right)^2}.
\end{aligned}$$

In order to compute this limit, we separate the last expression in three parts. First, using the result of Equation (8), we may write

$$\lim_{n,\gamma\to\infty} p_0 = \lim_{n,\gamma\to\infty} \frac{\frac{a^s}{s!}p_{-s,0}}{1-\frac{a}{s}\left(\frac{a_\gamma}{s}\right)^n} = \frac{\frac{a^s}{s!}\left[\sum_{x=0}^{s-1}\frac{a^x}{x!}+\frac{a^s}{s!}\frac{1}{1-a/s}\right]^{-1}}{1-\frac{a}{s}e^{-s\mu(1-a/s)\cdot\text{AWT}}}. \tag{9}$$

Second, we have

$$\lim_{n,\gamma\to\infty} \frac{s\mu}{\gamma^2}\frac{a_\gamma}{s}\frac{1}{\left(1-\frac{a_\gamma}{s}\right)^2} = \lim_{n,\gamma\to\infty} \frac{s\mu}{(s-a)^2}\frac{\left(a+\frac{\gamma}{\mu}\right)\left(s+\frac{\gamma}{\mu}\right)}{\gamma^2} = \frac{1}{s\mu(1-a/s)^2}. \tag{10}$$

Finally, one may write

$$-(n+1)\left(1-\frac{a_\gamma}{s}\right)\left(\frac{a_\gamma}{s}\right)^n+1-\left(\frac{a_\gamma}{s}\right)^n = 1-\left(\frac{a_\gamma}{s}\right)^n-\frac{(n+1)(s-a)}{s+\gamma/\mu}\left(\frac{a_\gamma}{s}\right)^n.$$

Using the assumption $\frac{n}{\gamma} = \text{AWT}$ yields

$$\lim_{n,\gamma\to\infty} -(n+1)\left(1-\frac{a_\gamma}{s}\right)\left(\frac{a_\gamma}{s}\right)^n+1-\left(\frac{a_\gamma}{s}\right)^n = 1-e^{-s\mu(1-a/s)\cdot\text{AWT}}(1+s\mu(1-a/s)\cdot\text{AWT}). \tag{11}$$

Combining Equations (9), (10) and (11) leads to the expression of $P(W < \text{AWT}) \cdot E(W|W < \text{AWT})$ which in turn allows us to derive $E(W|W < \text{AWT})$.

We now consider the expected waiting time of customers who are routed to Queue 2. The probability of having y customers in Queue 2 at Q-instants ($y \geq 0$) is $\sum_{x=1}^{n-1}\frac{s\mu}{\lambda}p_{x,y}+\frac{s\mu+\gamma}{\lambda}p_{n,y}$. From the proof of Proposition 1, we recall that $p_{x,y} = p_{n,y}$ and $p_{n,y} = p_{n,0}\left(\frac{a}{s}\right)^y$ for $1 \leq x \leq n$ and $y \geq 0$. Therefore,

$$\sum_{x=1}^{n-1}\frac{s\mu}{\lambda}p_{x,y}+\frac{s\mu+\gamma}{\lambda}p_{n,y} = p_{n,y}\left[\frac{s\mu}{\lambda}n+\frac{\gamma}{\lambda}\right] = \frac{\gamma}{\lambda}\left(\frac{a}{s}\right)^y p_{n,0}\left[1+\frac{s\mu}{\gamma}n\right].$$

This leads to the expected number in Queue 2. Next, applying Little's Law leads to $E(W|W > \text{AWT})$. Note that since $E(W)$ does not depend on the queue discipline, we could also compute the last performance measure by the relation $P(W < \text{AWT}) \cdot E(W|W < \text{AWT}) + P(W > \text{AWT}) \cdot E(W|W > \text{AWT}) = E(W)$.

This finishes the proof of the Theorem. □

4 Consequences of the MPW discipline

In this section we compare between the MPW and the FCFS discipline in order to evaluate the benefits and unintended consequences of the MPW policy.

Proportion of customers who have waited more than AWT. In Proposition 2, we evaluate the improvement which results from using the MPW instead of the FCFS discipline when considering the SLP as the unique service level measure. As expected, the MPW order achieves a lower SLP. Moreover, the benefits of this discipline are more apparent in congested situations.

Proposition 2 *For $a < s$, the MPW discipline achieves a lower SLP than the FCFS discipline. The difference between the two disciplines increases with the traffic intensity.*

Staffing decisions. We consider now the impact of the MPW order on staffing decisions for the problem of minimizing the number of servers under the constraint $P(W > \text{AWT}) \leq \overline{p}$, for a given $\overline{p} \in [0, 1]$. In Figure 3, we compute the optimal staffing as a function of the arrival rate under the FCFS and the MPW discipline. Under a MPW discipline, we observe that as λ increases the difference in staffing between the two disciplines increases. Moreover, above a threshold on the arrival rate the optimal value for s is simply $s = \lfloor a \rfloor + 1$, where $\lfloor a \rfloor$ is the integer part of a. This can be explained by the asymptotic behavior of $P(W > \text{AWT})$ as the ratio a/s tends to 1. We have

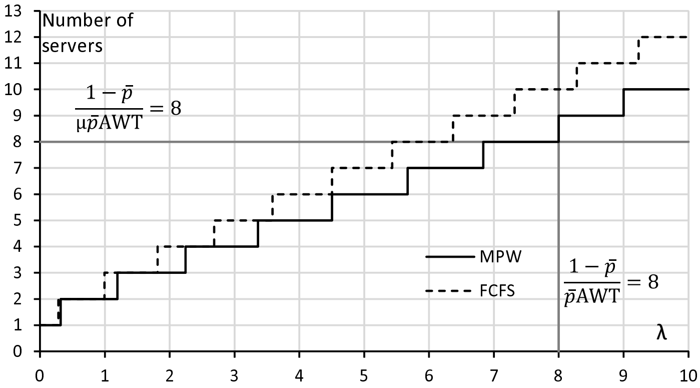

Figure 3: Optimal staffing ($\mu = 1$, $\overline{p} = 20\%$, AWT $= 0.5$)

$$P(W > \text{AWT}) \underset{a/s \to 1}{\sim} \frac{1}{1 + s\mu \cdot \text{AWT}},$$

where $f(x) \underset{x \to x_0}{\sim} g(x)$ means $\lim_{x \to x_0} \frac{f(x)}{g(x)} = 1$ for $x_0 \in \mathbb{R}$. Moreover, since $P(W > \text{AWT})$ is increasing in a, $P(W > \text{AWT}) \leq \frac{1}{1+s\mu \cdot \text{AWT}}$. Thus, with $s \geq \frac{1-\overline{p}}{\mu \overline{p} \cdot \text{AWT}}$ the constraint on $P(W > \text{AWT})$ is met. On the graph this threshold on the number of servers is equal to 8. Yet, the threshold $\frac{1-\overline{p}}{\mu\overline{p}\cdot\text{AWT}}$ does not necessarily ensures the stability of the system. Including the stability condition leads to a necessary condition of $s > \max(\frac{1-\overline{p}}{\mu\overline{p}\cdot\text{AWT}}, a)$. Consequently, if $\lambda \geq \frac{1-\overline{p}}{\overline{p}\cdot\text{AWT}}$, then $s = \lfloor a \rfloor + 1$ is the optimal staffing. This threshold on the arrival rate is also equal to 8 on the graph. Simple staffing solutions can then be derived from this study.

- If $\lambda \leq \frac{1-\overline{p}}{\overline{p}\cdot\text{AWT}}$, then $s = \lfloor \frac{1-\overline{p}}{\mu\overline{p}\cdot\text{AWT}} \rfloor + 1$ achieves the service level constraint and ensures the stability of the system.
- If $\lambda \geq \frac{1-\overline{p}}{\overline{p}\cdot\text{AWT}}$, then $s = \lfloor a \rfloor + 1$ is the optimal solution (i.e., no safety staffing).

From this analysis, we deduce that the staffing level under the MPW discipline can be significantly reduced. Above a threshold on the arrival rate, no safety staffing is even required. This gives another incentive for the manager to use the MPW order instead of the classical FCFS discipline.

Average excess wait. Although the MPW discipline allows the manager to reach more easily the service level defined by SLP, it has an unwanted effect on the waiting time of non-prioritized customers. The average excess (AE) measures this unintended consequence; $\text{AE} = E((W - \text{AWT})^+)$. Under the FCFS discipline, the waiting time W is defined by $P(W = 0) = 1 - C(s, a)$ and by its probability density function; $f_W(t) = s\mu(1-a)C(s,a)e^{-s\mu(1-a)t}$, for $t > 0$. Therefore, we get

$$\begin{aligned}\text{AE} = E((W - \text{AWT})^+) &= \int_{t=0}^{\infty} s\mu(1-a)C(s,a)e^{-s\mu(1-a)t}(t - \text{AWT})^+ \,\mathrm{d}t,\\ &= s\mu(1-a)C(s,a)\int_{t=\text{AWT}}^{\infty} e^{-s\mu(1-a)t}(t - \text{AWT})\,\mathrm{d}t.\end{aligned}$$

By changing the variable t into $u = t - \text{AWT}$, we obtain

$$\text{AE} == s\mu(1-a)C(s,a)e^{-s\mu(1-a)\text{AWT}}\int_{u=0}^{\infty} u \cdot e^{-s\mu(1-a)u}\,\mathrm{d}u = e^{-s\mu(1-a)\text{AWT}} \cdot \frac{C(s,a)}{s\mu(1-a/s)}.$$

In the last expression, we identify the expected waiting time in an M/M/s queue under the FCFS discipline; $\frac{C(s,a)}{s\mu(1-a/s)}$. Under the MPW discipline, the results of Theorem 2 allow us to obtain the AE:

$$\begin{aligned}\text{AE} &= E((W - \text{AWT})^+) = P(W > \text{AWT}) \cdot (E(W|W > \text{AWT}) - \text{AWT})\\ &= (1 + \lambda \cdot \text{AWT}) \cdot \frac{\left(1 - \frac{a}{s}\right) e^{-s\mu(1-a/s)\cdot\text{AWT}}}{1 - \frac{a}{s}e^{-s\mu(1-a/s)\cdot\text{AWT}}} \cdot \frac{C(s,a)}{s\mu(1-a/s)}\end{aligned}$$

In Proposition 3, we prove that the AE is higher under the MPW discipline than under the FCFS discipline.

Proposition 3 *For $a < s$, the MPW discipline achieves a higher AE than the FCFS discipline.*

In Figure 4, we compute the absolute difference between the AE under the MPW order and the FCFS discipline. We observe that:

1. The difference is positive. As expected, the service level for non-prioritized customers is worse under

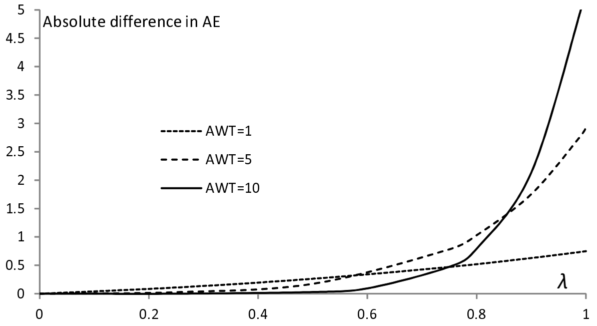

Figure 4: Absolute difference in AE ($s = 1$, $\mu = 1$)

MPW (Proposition 3).

2. The difference increases with the arrival rate. In congested situations, the negative outcome of MPW is important.

3. The influence of AWT depends on the workload. In low workload situations the difference is higher for low AWT. The opposite holds under high workload situations. Two phenomena are in competition. For low AWT, a high proportion of customers becomes non-prioritized. These customers deteriorate the AE. For high AWT, this proportion is lower but the proportion of prioritized customers is higher which in turn induces extreme wait for non-prioritized customers.

Acknowledgments. The author thanks the anonymous referee for very helpful comments and advises.

Remark. Note that short versions of the proofs are provided in the article. The proofs with complete details can be found in the online supplement.

References

[1] E. D. Bailey and T. Sweeney. Considerations in establishing emergency medical services response time goals. *Prehospital Emergency Care*, 7(3):397–399, 2003.

[2] Muhammad El-Taha and Shaler Stidham Jr. *Sample-path analysis of queueing systems*, volume 11. Springer Science & Business Media, 2012.

[3] L. Kleinrock. *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication, 1975.

[4] G. Koole, B.F. Nielson, and T.B. Nielson. First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60(5):1258–1266, 2012.

[5] Ger Koole. Redefining the service level in call centers. *Technical report, Department of Stochastics, Vrije Universiteit, Amsterdam*, 2003.

[6] J. M. Milner and T. L. Olsen. Service-level agreements in call centers: Perils and prescriptions. *Management Science*, 54(2):238–252, 2008.