Creating Work Breaks From Available Idleness

Xu Sun and Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, 10027; {xs2235,ww2040}@columbia.edu

September 9, 2016

Abstract

We show how dynamic priority (DP) rules for assigning available service representatives to arriving customers in customer contact centers can be used to create effective work breaks for the service representatives from naturally available idleness, assuming that the service system is staffed adequately to provide non-negligible idleness. We start by establishing many-server heavy-traffic limits to develop useful approximations for the distributions of server idle times with the customary longest-idle-server-first (LISF) rule and a random-routing (RR) alternative. We show that the pattern of idleness with these rules is totally different but neither produces effective work breaks. We then develop three DP rules and conduct simulation experiments to show that the new DP rules can indeed create effective work breaks from the available idleness. The first DP rule yields unannounced breaks, while the other more refined rules yield announced breaks.

Keywords: customer contact centers, call centers; work breaks; server-assignment rules; many-server queues.

Short Title: Creating Work Breaks

Contact Author: Ward Whitt, ww2040@columbia.edu

1 Introduction

There is now a substantial body of research on customer contact centers developing methods for efficient staffing and operation, as can be seen from Aksin et al. (2007). As these contact centers strive to improve customer experience, a key step in the process may be overlooked: how to enhance call center agent productivity? Without productive agents, it is impossible to provide superior customer support.

As reviewed in §5 of Aksin et al. (2007) on human resource issues, many studies on work-related stress have documented emotional exhaustion and burnout experienced by service representatives. This is attributed to handling high volumes of calls and difficult customers, while being required to meet high performance metrics, e.g., see Sawyerr et al. (2009), Lin et al. (2010). In addition to work overload, service representatives often do the same routine tasks every day and adhere to rigid call scripts, which are found to be monotonous. These negative impacts can result in decreased productivity and job satisfaction.

One way to help improve employee satisfaction and productivity is to provide adequate within-day work breaks. In addition to the common meal breaks, which last about an hour, it may be desirable to include shorter within-day work breaks of about 5 minutes. The importance of work breaks has been studied within the literature on organizational behavior and work psychology, beginning with the classic studies by Taylor (1911) and Mayo (1933), and expanding in recent years, e.g. , Jett and George (2003), Trougakos and Hideg (2009) and Fritz et al. (2013).

In this paper we apply queueing models to investigate if it is possible to create new rules for assigning available (idle) servers to arriving customers in a way that allows the cumulative idleness to be redistributed in order to create effective work breaks for the service representatives. We start by showing that the idle times provided by the customary longest-idle-server-first (LISF) rule are far too short to provide effective work breaks. We also show that the alternative random routing (RR) rule proposed in Mandelbaum et al. (2012) does not yield adequate work breaks either.

Then we introduce three dynamic priority rules that can be used to generate within-day work breaks of about 5 minutes every hour or two in a contact center with about 100 service representatives and 10% idleness, where service times average about 3 minutes. These new rules change the natural pattern of idleness with LISF or RR, replacing the usual idle times by many shorter ones and a few longer ones of the target length. Our first dynamic priority rule achieves unannounced work breaks, preserving work conservation (no server is idle if there is work to do), while the second and third rules achieve announced work breaks, sacrificing work conservation. We develop an optimization framework

to choose good control parameters and we report results from extensive simulation experiments to show that the new dynamic priority rules are effective.

This paper is in the same spirit as other performance analysis studies that recognize and respond to the preferences and concerns of the service representatives. First, Whitt (2006) developed a mathematical model to help analyze the benefit in contact-center performance gained from increasing employee (agent) retention, which is in turn obtained by increasing agent job satisfaction. Sisselman and Whitt (2007) introduced preference-based routing as a means to allow call center agents to help choose what calls they handle; see Biron and Bamberger (2010) for a related industrial psychology study. See §5 of Aksin et al. (2007) for further discussion.

Recent research by Chan et al. (2014) and Mandelbaum et al. (2012) has responded to the concern that server assignment rules should be fair to service representatives as well as customers. This includes a recognition that the service-time distributions of different representatives might not be identical; see Armony and Ward (2010), Atar (2008), Atar et al. (2011).

This paper is organized as follows: In §2 we specify the model, discuss important conservation laws and introduce the base case we shall use in our simulation experiments. In §3, we examine the pattern of idleness with the LISF and RR rules and develop useful approximations for the idle-time distribution. We show that work breaks are not produced naturally by the LISF and RR rules. In §4 we develop and evaluate the three dynamic priority rules designed to convert the available idleness into effective work breaks. Finally, in §5 we draw conclusions. In an online appendix we provide (i) an overview of the notation, (ii) supporting technical details, (iii) more on the simulation methodology and (iv) more results from simulation experiments.

2 The Model, Conservation and the Base Case

We wish to create work breaks out of available idleness. Thus, we start by doing a preliminary analysis of the available idleness. The servers experience alternating idle times and busy times. We let the busy time simply be the customer service time, so that the idle time may be 0. We define the server idle time as the interval between two successive service completions, allowing the possibility of 0 server idle time. (We do not attempt to characterize server busy times composed of two or more successive service times.) We first specify the model, discuss important conservation properties and introduce the base case of the model that we will consider throughout the paper.

2.1 The Model

Throughout this paper we consider the standard $M/GI/s$ multi-server queueing model with s homogeneous servers working in parallel and unlimited waiting space. The service times come from a sequence of independent and identically distributed (i.i.d.) random variables S_i having finite mean and variance. Without loss of generality (by choosing the measuring units for time), we let the mean service time be $E[S] \equiv \mu^{-1} \equiv 1$, where $\equiv$ denotes equality be definition. There is a Poisson arrival process with arrival rate λ that is independent of the service times. Hence, the inter-arrival times U_i are i.i.d random variables with an exponential distribution having mean $EU = 1/\lambda$.

The principal stochastic process is the number of customers in the system, denoted by $N(t)$. We will be especially interested in the number of busy servers $B(t) \equiv N(t) \wedge s$ and the number of idle servers $I(t) \equiv (s - N(t))^+$, where $x \wedge s \equiv \min\{x.s\}$ and $(x)^+ \equiv \max\{x, 0\}$. We assume that the traffic intensity $\rho \equiv \lambda/s\mu = \lambda/s < 1$, so that $(N(t), B(t), I(t)) \Rightarrow (N, B, I)$ as $t \to \infty$, where $\Rightarrow$ denotes convergence in distribution, $E[N] < \infty$, $B = N \wedge s$ and $I = (s - N)^+$.

2.2 Important Conservation

We focus on the long-run steady-state performance throughout the paper. Especially important for understanding allocations of idleness is a conservation law. Given that all arrivals are eventually served and that customer service times are not altered by any of the routing rules, the following (well known) expressions for the mean values are valid:

$$E[B] = \rho s \quad \text{and} \quad E[I] = (1 - \rho)s. \tag{2.1}$$

Formula (2.1) implies that, regardless of the routing rule, each server is idle a proportion $1 - \rho$ of the time. Thus we are concerned with ways to re-allocate the idle time subject to the constraint that (2.1) remains unchanged.

Let D be the duration of a break and let T be the interval between successive breaks (in steady state). Let β (β_t) denote the long-run proportion of time (of the idle time) during which each server is on break. As further conservation relations, we have

$$\beta = \beta_t(1 - \rho) \quad \text{and} \quad \beta = \frac{E[D]}{E[D] + E[T]}. \tag{2.2}$$

From above, we also have the important constraints

$$\beta \leq 1 - \rho \quad \text{and} \quad \frac{E[D]}{E[D] + E[T]} \leq 1 - \rho. \tag{2.3}$$

Assuming that $E[D] \approx \theta$, where θ is the target break duration, we have the approximate relation

$$\beta \approx \frac{\theta}{\theta + E[T]}. \tag{2.4}$$

Combining (2.2) and (2.4), we see that for given load ρ and target θ, we can relate β_t and $E[T]$; if we know one, then we know the other, at least approximately.

2.3 The Base Case

We are thinking of contact centers with 100 or more service representatives. To be specific, throughout the paper we make reference to one concrete system, which we use in all our simulation experiments. Our base case has $s = 100$, $\mu = 1$ and $\lambda = 90$, so that $\rho = 0.90$. We think of a call center in which the mean service times are 3 minutes, so that an hour is a time interval of length 20. The average inter-arrival time is $1/\lambda = 1/90$. Since the load is $\rho = 0.9$, in the long run the service representatives are idle 10% of the time, which is 6 minutes every hour and 12 minutes every two hours. We would like to rearrange that idleness to create a break of duration $\theta \equiv 5$ minutes every hour or two.

Thus, in the time scale with $\mu = 1$, we want breaks of length $5/3$ in each interval of length $20 - 40$. That is, we want breaks roughly equal to 1.6667 service times or about 150 inter-arrival times. Observe that our goal is reasonable in the sense that the target break of $\theta = 5/3$ is less than the available idleness of $2.0 - 4.0$ within a time interval of length $20 - 40$. A break of length $\theta = 5/3$ constitutes 83.3% (41.6%) of the available idleness in an interval of length 20 (40).

3 Idleness with the Basic Routing Rules

In this section we develop approximations for the server idle-time distribution in the $M/GI/s$ multi-server queueing model with the LISF and RR rules and traffic intensity $\rho < 1$. We start by developing approximations for I, the steady-state number of idle servers.

3.1 The Steady-State Number of Idle Servers

From §2.1, steady-state number of idle servers is $I \equiv (s - N)^+$. We now develop exact and approximate expressions for the first two moments of I and the probability $P(I = 0)$. The approximations draw on the many-server heavy-traffic limit theory in Halfin and Whitt (1981). Let Φ and ϕ be the cumulative distribution function (cdf) and probability density function (pdf), respectively, for a standard normal $N(0, 1)$ random variable.

Theorem 3.1 (*steady-state numbers of busy and idle servers*) *In the $M/M/s$ model with traffic intensity $\rho < 1$, the steady-state numbers of busy and idle servers satisfy*

$$\begin{aligned} E[B] &= \rho s, \quad E[I] = (1-\rho)s, \quad P(B=s) = P(I=0) = P(N \geq s) \equiv \alpha, \\ Var(B) &= Var(I) = \rho s(1-\alpha). \end{aligned} \tag{3.1}$$

where α is given in (1.2) *of* Halfin and Whitt (1981). *If $s \to \infty$ with $(1-\rho)\sqrt{s} \to \xi$, $0 < \xi < \infty$, as in Halfin and Whitt (1981), then*

$$\alpha \to [1 + \xi\Phi(\xi)/\phi(\xi)]^{-1}. \tag{3.2}$$

Proof. The formulas in (3.1) follow from §1 of Halfin and Whitt (1981), noting that $I = s - B$, where $B \equiv N \wedge s$ and $E[B] = \alpha s + \sigma_1^{(1)}$ and $E[B^2] = \alpha s^2 + \sigma_1^{(2)}$ for $\sigma_1^{(j)}$ defined in (1.4) of Halfin and Whitt (1981). Specifically, (3.1) follows by algebra from the formulas $\sigma_1^{(1)} = \rho s - \alpha s$ and $\sigma_1^{(2)} = (\rho s)^2 - \alpha s^2 + \rho s(1-\alpha)$ given on the bottom of p. 572 of Halfin and Whitt (1981). Then (3.2) is the MSHT QED limit from Proposition 1 of Halfin and Whitt (1981). ∎

We now want to develop an approximation for the full distribution of I and extend it to the $M/GI/s$ model with a non-exponential service-time distribution having the same mean. For that purpose, we use experience that $P(I=0) = P(B \geq s)$ tends to be approximately independent of the service-distribution; see Whitt (2004). Moreover, the conditional distribution of N given that $N < s$ tends to be distributed approximately the same as the conditional distribution of N in the associated $M/GI/\infty$ model given that $N < s$. Since the $M/GI/\infty$ model has the insensitivity property, i.e., the steady-state distribution of N is independent of the service-time distribution beyond its mean, we assume that property for $M/GI/s$ conditional that $N < s$. Since N has a Poisson distribution in the $M/GI/\infty$ model, we approximate the conditional distribution of I given $I > 0$ by a truncated normal distribution. Thus, for the $M/GI/s$ model, we approximate by

$$B \approx N(m,v) \wedge s \quad \text{and} \quad I \approx (s - N(m,v))^+ \tag{3.3}$$

where $N(m,v)$ is a random variable with a normal distribution having mean m and variance v. The parameters m and v can be obtained by solving the equations

$$\begin{aligned} P(I=0) &\approx P((N(m,v) \geq s) \approx \alpha \\ E[(I)^k] &\approx (1-\alpha)E[(s - N(m,v))^k | N(m,v) < s] \quad \text{for} \quad k = 1,2, \end{aligned} \tag{3.4}$$

with explicit formulas given, e.g., in Proposition 18.3 of Browne and Whitt (1995), which we review in the appendix.

Example 3.1 (*The steady-state number of idle servers in the base case*) For the base case in §2.3 with $s = 100$, $\mu = 1$ and $\rho = 0.9$, $P(I = 0) = 0.215$ by exact calculation. (The approximation in (3.2) is $\alpha \approx 0.223$.) The mean and variance are $E[I] = (1 - \rho)s = 10.0$ and $Var(I) = \rho s(1 - \alpha) = 70.65$.

3.2 Approximations for the Longest-Idle-Server-First (LISF) Rule

We now develop an approximation for the idle-time distribution. We first consider the customary LISF rule for assigning idle servers to new arrivals. As a further approximation, we assume that the number of idle servers found upon service completion is the steady-state number. With that approximating assumption, the steady-state idle time V can be represented approximately as the random sum

$$V \approx \sum_{i=1}^{I} U_i, \tag{3.5}$$

where I is the number of idle servers found by the server upon completing a previous service and the random variables U_i represent the successive inter-arrival times after time t, which is independent of I.

Theorem 3.2 (*moments of the idle time distribution*) *In the $M/M/s$ model with traffic intensity $\rho < 1$, the approximate steady-state idle time distribution with the LISF routing rule, V in (3.5), satisfies*

$$\begin{aligned} E[V] &= (1 - \rho)/\rho, \quad P(V = 0) = P(I = 0) = P(N \geq s) \equiv \alpha, \\ Var(V) &= (1 - \rho\alpha)/(s\rho^2). \end{aligned} \tag{3.6}$$

where α is as in Theorem 3.1.

Proof. From the random sum representation in (3.5), we have $P(V = 0) = P(I = 0)$,

$$E[V] = E[I]E[U_i] = \frac{s(1 - \rho)}{s\rho} = \frac{1 - \rho}{\rho}, \tag{3.7}$$

which agrees with the exact value by Little's law, assuming that the mean service period is $ES = 1$ and that $\rho = ES/(ES + EV)$. We apply the conditional variance formula to compute the variance of V in (3.5). In particular,

$$\begin{aligned} Var(V) &= E[Var(V|I)] + Var(E[V|I]) \\ &= E[I/(s\rho)^2] + Var(I/(s\rho)) \\ &= \frac{s(1 - \rho)}{(s\rho)^2} + \frac{s\rho(1 - \alpha)}{(s\rho)^2} = \frac{1 - \rho\alpha}{s\rho^2}. \quad \blacksquare \end{aligned} \tag{3.8}$$

In the QD MSHT limit, I is asymptotically normal, so that I also is asymptotically normal, e.g., see §7.4 of Whitt (2002). Hence, we suggest the approximation

$$I \approx N(m, v) \vee 0, \tag{3.9}$$

where the mean m and variance v can be obtained by solving the equations

$$\begin{aligned} P(V = 0) &= P(I = 0) \approx \alpha \\ E[V^k] &\approx (1 - \alpha) E[N(m, v))^k | N(m, v) > 0] \quad \text{for} \quad k = 1, 2, \end{aligned} \tag{3.10}$$

just as in §3.1.

Example 3.2 (*The idle-time distribution in the base case for LISF*) For the base case in §2.3 with $s = 100$, $\mu = 1$ and $\rho = 0.9$, we have $P(V = 0) = \alpha = 0.215$, $E[V] = (1 - \rho)/\rho = 0.1111$, $Var(V) = (1 - \rho\alpha)/(s\rho^2) = 0.0100$ and $SD(V) = 0.100$. Recall that an idle time of 5 minutes is 1.6667 in our scale. Since 1.6667 is 15.67 standard deviations above the mean of 0.1111, we judge that it is highly unlikely that an idle time could serve as a satisfactory work break. That is shown by a simulation estimate of the idle time pdf in Figure 1a. We see no mass above 0.6, which is well below 1.6667.

3.3 The Cumulative Idleness in an Interval

We now develop approximations for the cumulative idleness of a single server during in an interval $[0, t]$, which we denote by $C(t)$. We assume that the server just starts service at time 0, so the server's service starts at time 0. Let $\{(S_n, V_n) : n \geq 1\}$ be the sequence of ordered pairs of successive service and idle times.

As an approximation, we assume all idle times are initiated with an independent sample of the steady-state number of idle servers, distributed as $I \approx N((1 - \rho)s, \rho s(1 - \alpha))$ as in (3.1), with this number remaining constant during each idle time. With these approximating assumptions, the sequence $\{(S_n, V_n)\}$ is an alternating renewal process. We emphasize that this structure only holds as an approximation.

Let $X_n \equiv S_n + V_n$ be the n^{th} service cycle. Let $M(t)$ count the number of full service cycles up to time t, Then the cumulative idleness can be represented as

$$C(t) = \sum_{n=1}^{M(t)} V_n + V_c(t), \tag{3.11}$$

where $V_c(t)$ is the completed (or elapsed) idle time in the cycle in progress at time t. In particular, $V_c(t) = 0$ if

$$\sum_{n=1}^{M(t)} V_n + S_{N(t)+1} > t; \tag{3.12}$$

otherwise,

$$V_c(t) = V_{M(t)+1} - [\sum_{n=1}^{M(t)+1} X_n - t]. \tag{3.13}$$

We shall apply the central limit theorem (CLT) to approximate the distribution of $C(t)$ under the assumptions above. In particular, the CLT can be obtained from a functional central limit theorem (FCLT) in Whitt (2000). We apply Theorems 2.1, 6.1 and 6.2 there with the role of the idle and busy times reversed. That yields the limit:

$$t^{-1}[C(t) - (1-\rho)t] \Rightarrow N(0, \sigma_C^2) \quad \text{as} \quad t \to \infty, \tag{3.14}$$

where $\Rightarrow$ denotes convergence in distribution and

$$\sigma_C^2 \equiv \rho^{-1}[(1-\rho)^2\sigma_S^2 + \rho^2\sigma_V^2], \tag{3.15}$$

where $\sigma_S^2 = c_s^2$ is the variance of a mean-1 service time, which is 1 for an exponential service time, and σ_V^2 is the variance of an idle time, which is $(1 = \alpha\rho)/s\rho^2$ as in (3.8) for the $M/M/s$ model with LISF and is given in (3.26) for RR. As a consequence, we have the Gaussian approximation for all t not too small,

$$C(t) \approx N((1-\rho)t, \sigma_C^2 t), \tag{3.16}$$

where σ_C^2 is given in (3.15).

In our base case with $s = 100$, $\mu = 1$, $\rho = 0.9$ and $t = 20$, corresponding to one hour with expected service times equal to 3 minutes, we have

$$\sigma_C^2 = (1/0.9)[(0.01) \times 1) + 0.81(0.01)] = \frac{0.0181}{0.9} = 0.0210 \tag{3.17}$$

and

$$E[C(20)] \approx 2.0 \quad \text{and} \quad SD(C(20)) \approx \sqrt{0.420} = 0.648, \tag{3.18}$$

while

$$E[C(40)] \approx 4.0 \quad \text{and} \quad SD(C(40)) \approx \sqrt{0.840} = 0.917, \tag{3.19}$$

We can apply these approximations to look at the probability that the cumulative idleness over 1 hour or 2 hours is at least 5 minutes.

These are

$$P(C(20) \geq 1.6667) \approx P(N(2.0, 0.4444) \geq 1.667) = P(N(0,1) \geq -0.50) = 0.692 \tag{3.20}$$

and

$$P(C(40) \geq 1.6667) \approx P(N(4.0, 0.8888) \geq 1.6667) = P(N(0,1) \geq -2.47) = 0.9933. \tag{3.21}$$

Thus, with LISF, there is about a one-third chance that the cumulative idleness over an hour will not exceed 5 minutes, but there is less than an 0.67% chance over two hours. As a consequence, it would appear that we should be able to create 5-minute breaks every two hours, without too much difficulty.

Simulation experiments confirm the approximations for the mean $E[C(20)]$ and $E[C(40)]$, but indicate errors in our approximations for the standard deviation. The simulation indicates that 95% confidence intervals for $SD(20)$ and $SD(40)$ with LISF are respectively, $[0.78, 0.81]$ and $(1.09, 112]$. These are about $0.795/0.667 = 1.19$ and $1.105/0.943 = 1.17$ times above our estimated values. That is, our approximations underestimate the true standard deviations.

3.4 Approximations for the Random-Routing (RR) Rule

We now consider the RR rule for assigning idle servers to new arrivals, proposed for an emergency department in Mandelbaum et al. (2012), in which each available server is chosen with equal probability at each customer arrival epoch. This RR rule will make the idle times more variable, so that it is more likely that an idle time could serve as a work break.

Again, as a further approximation, we assume that the number of idle servers found upon service completion is the steady-state number I. Moreover, as assume that number I does not change over the successive arrival times required for the assignment. With these approximating assumptions, the steady-state idle time can be represented approximately as the random sum

$$V \equiv V_{RR} \approx \sum_{i=1}^{A} U_i, \tag{3.22}$$

where $A \equiv A(I)$ is the random number of arrivals required for the assignment, which depends on the number of idle servers found upon service completion, I, and again U_i is a sequence of i.i.d. exponentially distributed random variables, each with mean $EU_i = 1/\lambda = 1/s\rho$, that is independent of $A(I)$.

Conditionally on I, the random variable $A \equiv A(I)$ is a geometrically distributed random variable on the positive integers with parameter $1/I$. Thus I is approximately a random mixture of geometric

distributions. Thus,

$$E[A|I] \approx I \quad \text{and} \quad Var(A|I) = I^2(1 - (1/I)). \tag{3.23}$$

To simplify, we use the approximation

$$Var(A|I) \approx I^2,$$

which is reasonable if I is suitably large. Clearly, this approximation should overestimate $Var(A|I)$.

As a consequence,

$$E[A] = E[I] = s(1 - \rho), \tag{3.24}$$

$$\begin{aligned} Var(A) &= E[Var(A|I)] + Var(E[A|I]) \\ &\approx E[I^2] + Var(I) = 2Var(I) + E[I]^2 \\ &\approx 2s\rho(1 - \alpha) + s^2(1 - \rho)^2, \end{aligned} \tag{3.25}$$

Finally, for RR, $E[V] = (1 - \rho)/\rho$ just as in (3.7), but now the variance is

$$\begin{aligned} Var(V) &= E[Var(V|A)] + Var(E[V|A]) \\ &= E[A/(s\rho)^2] + Var(A/(s\rho)) \\ &\approx \frac{s(1 - \rho)}{(s\rho)^2} + \frac{2s\rho(1 - \alpha) + s^2(1 - \rho)^2}{(s\rho)^2} = \frac{(1 - \rho)^2}{\rho^2} + \frac{1 + \rho - 2\rho\alpha}{s\rho^2}. \end{aligned} \tag{3.26}$$

If the standard deviation $\sqrt{Var(V)}$ is close to the mean $E[V]$, then we can approximate V by an exponential distribution with mean $E[V]$. Otherwise, we might approximate the distribution of V by a mixture of two exponential distributions, matching the mean and variance. But, in contrast to LISF, V_{RR} is not nearly normally distributed.

Example 3.3 (*The idle time distribution in the base case for RR*) For the base case in §2.3 with $s = 100$, $\mu = 1$ and $\rho = 0.9$, the mean EA is 10, just as for $E[I]$, but the variance $Var(A)$ is larger than the variance $Var(I)$, increasing from 70.65 to $Var(A) = 180(1 - 0.215) + 100 = 241.3$. In the illustrative example, again $E[V_{RR}] = 1/9 = 0.1111$, but now we have the larger variance $Var(V_{RR} = 2.513/81 = 0.0310$ with the associated standard deviation $SD(V_{RR}) = 0.176$. We see that the standard deviation is substantially larger than the mean, so that the idle time distribution is more variable than an exponential distribution (for which the standard deviation coincides with the mean).

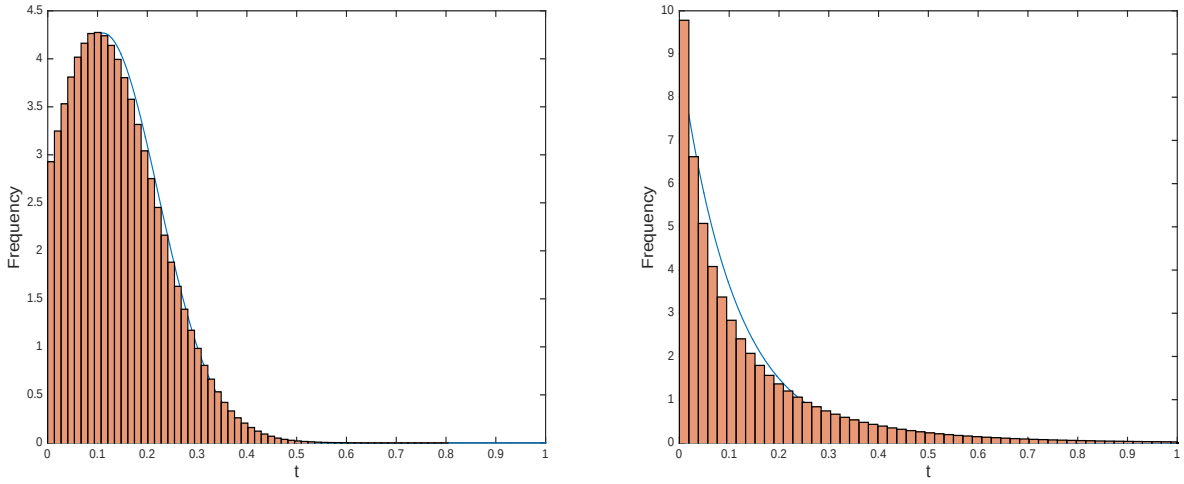
We can use Chebychev's inequality to obtain a crude bound on the probability that an idle time I_{RR} could serve as a work break. In particular,

$$P(V_{RR} \geq 1.6667) \leq P(V_{RR} - E[V_{RR}]| \geq 1.5556) \leq \frac{Var(V_{RR})}{(1.5556)^2} = \frac{0.0310}{2.4199} \approx 0.0128. \tag{3.27}$$

This bound shows that it is highly unlikely that an idle time could serve as a work break of 5 minutes or more.

3.5 Simulation Comparison of the Idle-Time Distributions

Figures 1a and 1b show histograms estimated by simulation of the idle-time pdf with LISF and RR. In each case the atom at time 0 is omitted from the histogram. Consistent with the analysis above, these histograms have the suggested form, i.e., approximately truncated Gaussian for LISF and a mixture of exponentials for RR. The histograms show that there is a significantly greater chance that an idle time could serve as a work break for RR than for LISF, but neither is significant.



(a) LISF (b) RR

Figure 1: Histograms estimated by simulation (with the atom at 0 removed) of the idle-time distribution with LISF (left) and RR (right) for the base case in §2.3

In the simulation, the system was started empty, but the first 100 time units were omitted to allow the system to approach steady state. The 95% confidence intervals for the mean and standard deviation with LISF were $[0.1109, 0.1116]$ and $[0.1004, 0.1012]$, respectively, which strongly supports the formulas 0.1111 and 0.1000 derived in §2. The 95% confidence intervals for the mean and standard deviation with RR were $[0.1106, 0.1116]$ and $[0.1664, 0.1672]$, respectively, which supports the approximations of 0.1111 and 0.176 derived in Example 3.3. The approximate standard deviation of 0.176 is about 5% above the simulation estimate of 0.167. Further discussion of the simulation methodology appears in the appendix.

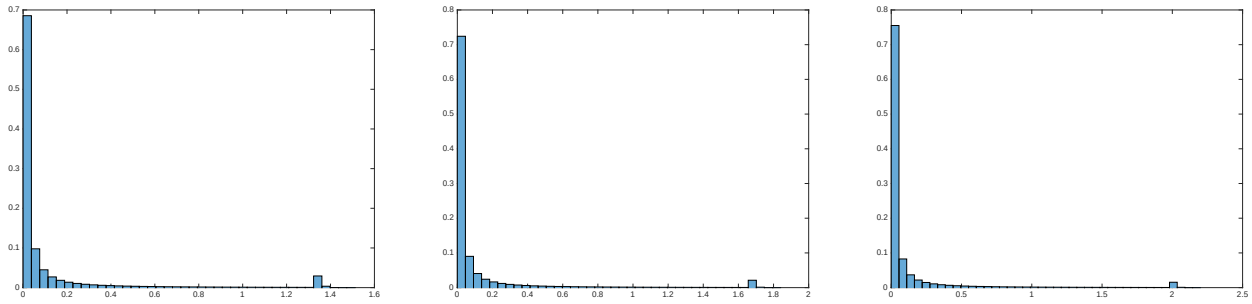
4 Dynamic Priority Rules to Create Work Breaks

We now introduce dynamic priority (DP) rules for assigning idle servers to arriving customers. Our goal is to design routing schemes that can convert server idleness into meaningful work breaks over a day by creating occasional long idle times that serve as a work break with other much shorter idle times in between. In the illustrating example, we might aim to create one work break of at least 5 minutes every one (two) hours, and thus 8 (4) different times over an 8-hour shift.

4.1 The Basic DP Rule: DP1

We start with a basic DP rule, denoted by $DP1 \equiv DP1(\theta)$, which is a function of the target work break duration θ. In our scaling, in which the average service time is equal to 3 minutes, a break of 5 minutes is $5/3 = 1.6667$ time units.

For each server, we maintain the time at which the last work break (idle-period of at least θ) is completed. Given the knowledge of the current time, we can then compute for each server the elapsed time since the last work break ended. Specifically, the elapsed time since the last break is the difference between the current time and the end of the last work break. At each arrival epoch, we first look for idle servers that have been idle for at least θ time units. If there are any of these, then we assign the server with the longest elapsed time since the last work break end time, replace its last work break end time by the current time t and reset the elapsed time to zero. That should prevent work breaks from being much greater than θ. Second, if there are idle servers but all their elapsed idle times are smaller than θ, then we assign the available server with the least elapsed time since the last break, among all available servers. If no server is available, then the entering customer will wait in queue.



(a) $\theta = 4/3$ (b) $\theta = 5/3$ (c) $\theta = 2$

Figure 2: Histograms estimated from simulation of the idle-time distribution with rule $DP1(\theta)$ for three candidate targets θ

Figure 2 displays histograms estimated from simulation of the idle-time distribution with $DP1(\theta)$

for three different values of θ: $\theta = k/3$ for $k = 4, 5, 6$. Panel (b) is for our base model with mean service times of 3 minutes, where the target duration was a 5-minute work break every one or two hours, so that $\theta = 5/3$. Evidently, DP1 is able to create work breaks from idleness. Figure 2 shows that DP1 creates a peak in the distribution at the target θ and the rest of the distribution concentrates near the origin, decaying very rapidly. Overall, we obtain a probability density function which is bimodal.

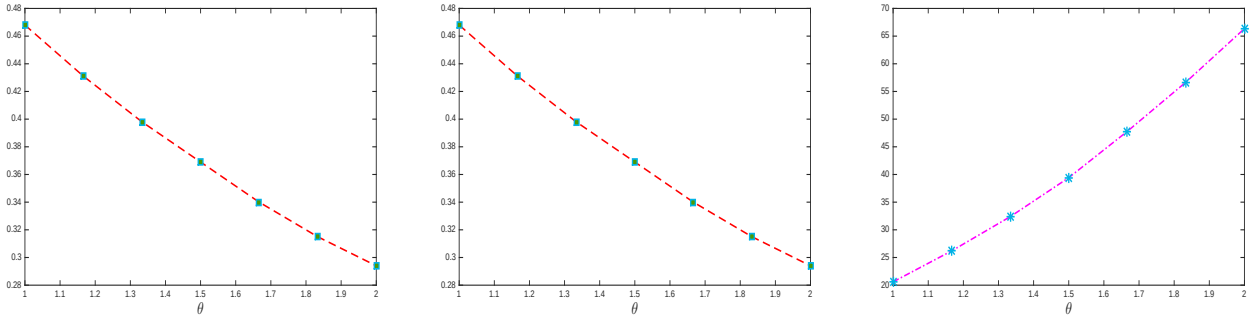
We elaborate in Table 1 by showing the 95% confidence intervals (based on the data collected from the simulation experiments) for the mean and standard deviation of the idle time V and the interval between successive work breaks, which we denote by T. Here we use β_n to represent the long-run proportion of idle times that are work breaks and β_t to represent the long-run proportion of idle time that is made up of work break time.

θ	$E[V]$	$SD(V)$	β_n	β_t	$E[T]$	$SD(T)$
6/6	$0.1112 \pm 5 \times 10^{-4}$	$0.2488 \pm 6 \times 10^{-4}$	$0.0514 \pm 4 \times 10^{-4}$	0.486 ± 0.001	20.60 ± 0.14	9.97 ± 0.10
7/6	$0.1112 \pm 5 \times 10^{-4}$	$0.2653 \pm 7 \times 10^{-4}$	$0.0406 \pm 3 \times 10^{-4}$	0.431 ± 0.001	26.19 ± 0.20	11.81 ± 0.13
8/6	$0.1113 \pm 4 \times 10^{-4}$	$0.2800 \pm 7 \times 10^{-4}$	$0.0330 \pm 2 \times 10^{-4}$	0.398 ± 0.001	32.37 ± 0.23	13.69 ± 0.19
9/6	$0.1113 \pm 4 \times 10^{-4}$	$0.2928 \pm 7 \times 10^{-4}$	$0.0271 \pm 2 \times 10^{-4}$	0.369 ± 0.001	39.42 ± 0.28	15.78 ± 0.24
10/6	$0.1108 \pm 5 \times 10^{-4}$	$0.3035 \pm 9 \times 10^{-4}$	$0.0225 \pm 2 \times 10^{-4}$	0.340 ± 0.001	47.79 ± 0.42	18.58 ± 0.29
11/6	$0.1108 \pm 5 \times 10^{-4}$	$0.3138 \pm 1 \times 10^{-3}$	$0.0190 \pm 2 \times 10^{-4}$	0.315 ± 0.002	56.68 ± 0.52	21.18 ± 0.34
12/6	$0.1112 \pm 5 \times 10^{-4}$	$0.3236 \pm 1 \times 10^{-3}$	$0.0163 \pm 2 \times 10^{-4}$	0.294 ± 0.002	66.35 ± 0.64	23.63 ± 0.43

Table 1: Estimated performance measures of DP1(θ) as a function of θ

Consistent with the conservation laws in §2.2, $E[V]$ approximately equals $1/9 = 0.1111$ for all θ, because the long-run proportion of idleness is solely determined by the traffic intensity ρ, independent of the server-assignment rule (provided that it is work-conserving). In addition, $SD(V)$ increases as θ grows, but not significantly. Table 1 and Figure 3 also show that As θ increases from 1 to 2, the proportion of idle times that are work breaks shrinks by two thirds, changing very rapidly from 0.0514 to 0.0163. The proportion of idle time occupied by work breaks also decreases, but at a lower speed, changing from 0.486 to 0.294. The mean inter-break time $E[T]$ more than triples from approximately 20.60 to around 66.35. Finally, we note that the parameters $\rho, \beta, \beta_t, \theta, E[T]$ are related via (2.1)-(2.4) in §2.2. Hence, for given ρ and θ, knowing one of β_t or $E[T]$ provides the other.

It is also useful to look at the pattern of successive idle times over a long horizon. Figure 4 shows that we succeed in getting occasional long idle times with many very short ones in between. That is done by displaying the sample-path of successive idle-times for a set of randomly selected servers in an $M/M/s$ queue operating under DP1 where $\theta = 5/3$ over a time interval of length 300. As usual, we choose $\lambda = 90$, $s = 100$ and $\mu = 1$. Here the vertical axis measures the length of an idle-time and the



(a) Estimate of β_n (b) Estimate of β_t (c) Estimate of $E[T]$

Figure 3: Simulation estimates of relevant performance metrics with rule DP1 for $\theta = 5/3$.

horizontal axis indexes the successive idle times. Based on the plots, we conclude that DP1 generates occasional long idle times with many tiny ones in between. Over a long horizon, these work breaks occur fairly regularly.

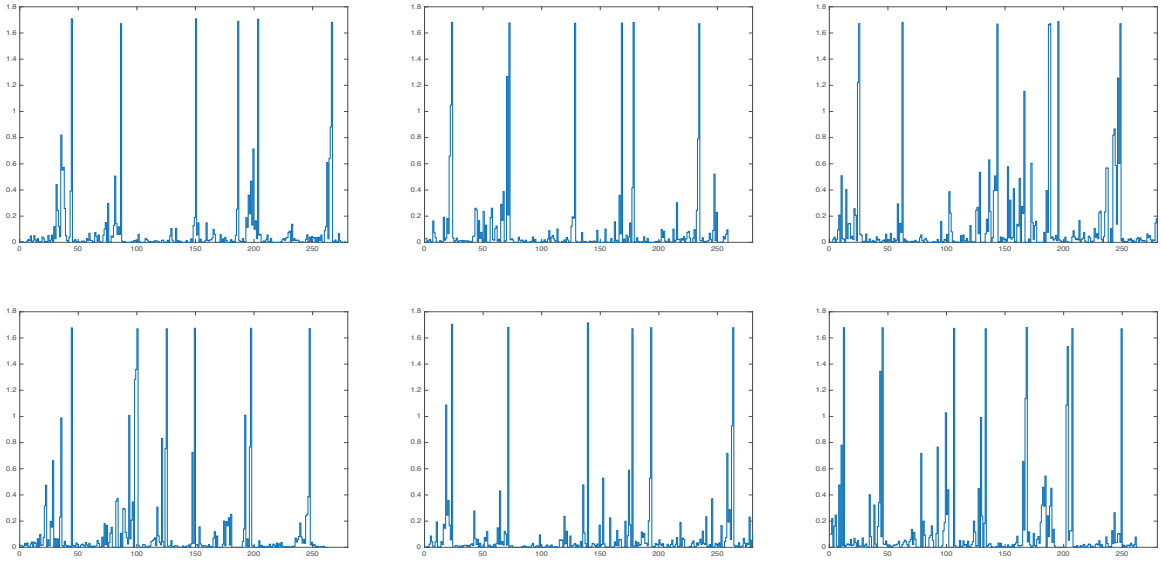
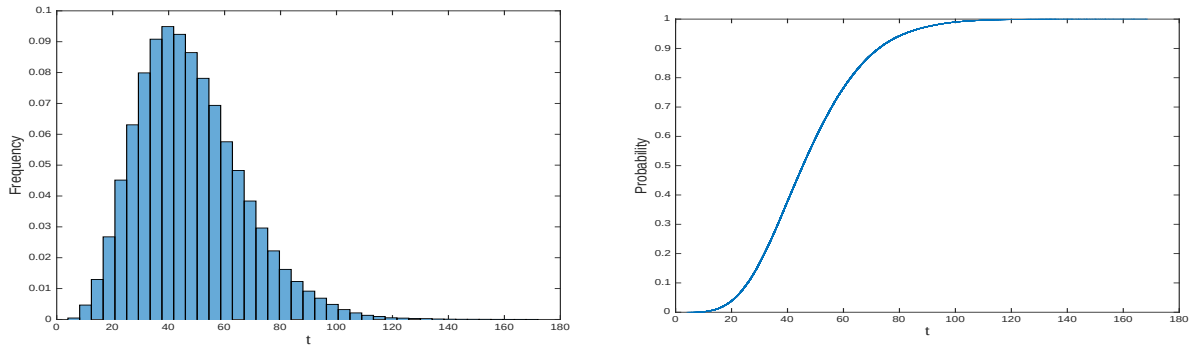


Figure 4: Six sample paths of successive idle times for DP1 with $\theta = 5/3$ in an interval of length 300

A great appeal of DP1 is its simplicity. To implement this rule, it suffices to choose a target break duration θ and then make routing decisions based on the elapsed time since the last break end time. In addition, because servers are always available if not serving customers, this dynamic priority rule is work-conserving, i.e., there are no idle servers if there are customers in queue. As a consequence, the performance for customers is the same as the standard s-server queue using LISF.

Nevertheless, there are three serious drawbacks in DP1. First, Figure 4 shows that that there tend

to be long idle periods that occur right before many of the work breaks. We regard this as undesirable, because we want all long idle periods to be work breaks. Second, closely rated to the first drawback, the interval between successive breaks tends to be too long, often being above the interval $[20, 40]$. Indeed, Table 1 shows that the mean is 47.8 for $\theta = 5/3$. The full distribution is shown in Figure 5, with a histogram on the left and the empirical cumulative distribution function (ecdf) on the right.



(a) Histogram

(b) ECDF

Figure 5: The histogram (left) and ecdf (right) estimated from simulation of the distribution of T, the time between breaks, with rule DP1 for $\theta = 5/3$

Finally, we want to be able to tell a server that a work break is beginning when it starts, so that the server can leave and be away for the allotted time; i.e., we want to be able to announce the work breaks. That is evidently not possible with DP1.

4.2 Dynamic Routing Rules with Announced Work Breaks

In order to create announced work breaks, we introduce a threshold control parameter τ. For each server, we keep track of the elapsed time since the last work break (an idle time of length at least θ). For each server, we let the first idle time after time τ has elapsed since the last work break be an announced work break. At this idle time, the break is announced and the server is not available to provide service for the duration θ after that time. With this new $DP2 \equiv DP2(\theta, \tau)$ dynamic priority rule, there can still be unannounced work breaks, because a work break might occur before time τ has elapsed since the last break.

It remains to determine what values of τ are desirable. Given that we want to achieve a break of 5/3 every 20-40 time units, it is natural to expect that we should have $20 \leq \tau \leq 40$. To help decide, let $p_B \equiv p_B(\theta, \tau)$ be the long-run proportion of all work breaks that are announced. Clearly, p_B is a decreasing function of both θ and τ, while the mean time between successive breaks, $E[T]$, is an

increasing function of τ. To get more announced breaks for given θ, we would make τ small.

On the other hand, many announced work breaks can degrade the performance experienced by customers. Thus, when considering what is a good value for τ, we need to consider the tradeoff between more announced work breaks, as measured by high values of p_B and low values of $E[T]$, and performance degradation for the customers. To measure that performance degradation, let p_D be the steady-state delay probability, and let $E[Q]$ and $SD(Q)$ be the mean and standard deviation of the steady-state queue length.

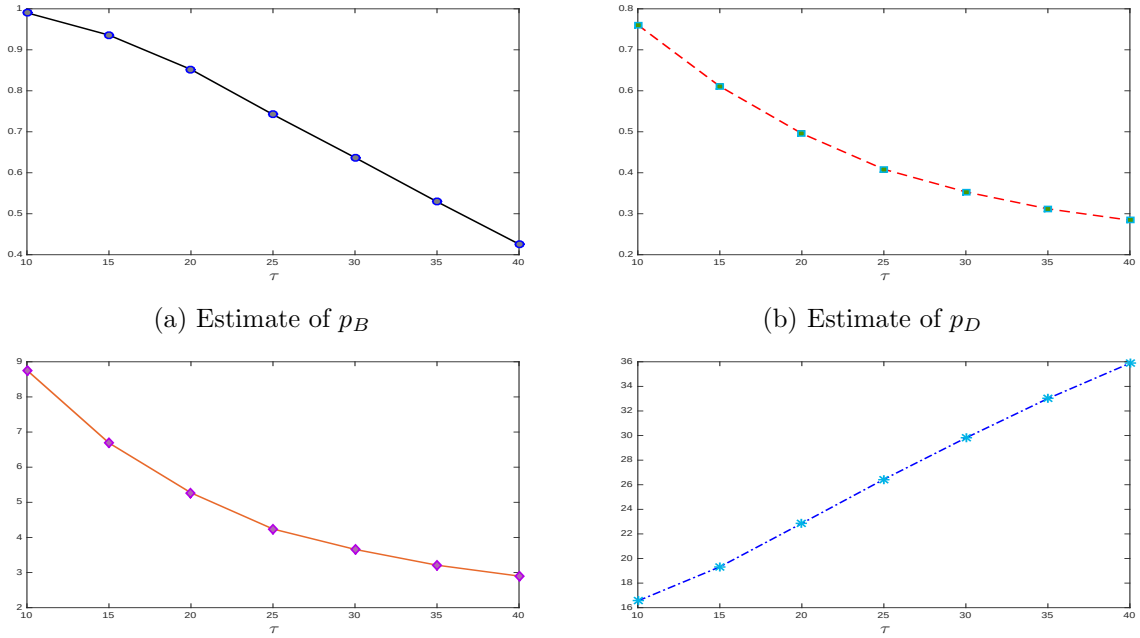
For the target break duration $\theta = 5/3$, the tradeoff is shown in Table 2 and Figure 6. Table 2 shows the performance measures of $DP2$ as a function of τ for $\theta = 5/3$ and for $10 \leq \tau \leq 40$. For $\tau = 10$, we see that virtually all work breaks are announced ($p_B > 0.99$) and the interval between breaks is quite short ($E[T] = 16.6$), but there is severe performance degradation for the customers. For example, the probability of delay is $p_D > 0.76$, which is far greater than the value 0.215 with $DP1$. In balance, the relevant range would seem to be $20 \leq \tau \leq 30$, but the desire to further reduce the performance degradation for customers leads us to consider the next rule $DP3$.

	p_B	p_D	$E[Q]$	$SD(Q)$	$E[T]$	$SD(T)$
$\tau = 10$	$0.9903 \pm 5.8 \times 10^{-4}$	$0.7607 \pm 2.3 \times 10^{-3}$	8.76 ± 0.08	9.84 ± 0.11	16.55 ± 0.07	7.57 ± 0.10
$\tau = 15$	$0.9359 \pm 2.2 \times 10^{-3}$	$0.6109 \pm 4.3 \times 10^{-3}$	6.69 ± 0.09	9.20 ± 0.12	19.32 ± 0.07	5.85 ± 0.10
$\tau = 20$	$0.8521 \pm 3.9 \times 10^{-3}$	$0.4958 \pm 4.7 \times 10^{-3}$	5.27 ± 0.08	8.47 ± 0.10	22.86 ± 0.07	5.17 ± 0.09
$\tau = 25$	$0.7421 \pm 4.3 \times 10^{-3}$	$0.4085 \pm 4.3 \times 10^{-3}$	4.24 ± 0.07	7.81 ± 0.12	26.44 ± 0.06	5.21 ± 0.08
$\tau = 30$	$0.6371 \pm 5.7 \times 10^{-3}$	$0.3533 \pm 3.6 \times 10^{-3}$	3.66 ± 0.07	7.53 ± 0.13	29.81 ± 0.08	5.92 ± 0.09
$\tau = 35$	$0.5294 \pm 6.7 \times 10^{-3}$	$0.3120 \pm 3.8 \times 10^{-3}$	3.21 ± 0.07	7.14 ± 0.12	33.01 ± 0.12	7.09 ± 0.09
$\tau = 40$	$0.4259 \pm 7.0 \times 10^{-3}$	$0.2842 \pm 3.6 \times 10^{-3}$	2.90 ± 0.06	6.85 ± 0.14	35.88 ± 0.15	8.33 ± 0.09

Table 2: Estimated performance measures of DP2(θ, τ) as a function of τ for $\theta = 5/3$

We can also expose the impact of the threshold parameter τ by looking at appropriate sample paths. For greater insight, we let $I_d(t) \equiv s - N(t)$ be the number of idle servers at time t, allowing it to be negative as well as positive. Thus $-I_d(t) = Q(t)$, the queue length, when $I_d(t) < 0$, and $I(t) = I_d(t)^+$. We let $S_b(t)$ be the number of servers on break at time t. Figure 7 displays sample paths of the number of servers on break, $S_b(t)$, and the number of idle servers, $I_d(t)$, for six different values of τ when $\theta = 5/3$. Panel (a) with $\tau = 15$ shows extreme performance degradation for customers because we see many places with $S_b(t)$ well above $I_d(t)$. Moreover, after periods when many servers are on break (e.g., when $S_b(t) > 10$, we often see a severe drop in $I_d(t)$, which indicates a buildup of a large queue. Consistent with Table 2 and Figure 6, we see fewer large gaps as τ increases.

To provide another perspective of these sample paths, Figure 8 displays sample paths of the gap



(a) Estimate of p_B

(b) Estimate of p_D

(c) Estimate of $E[Q]$

(d) Estimate of $E[T]$

Figure 6: Simulation estimates of key performance metrics with rule DP2(θ, τ) as a function of τ for $\theta = 5/3$

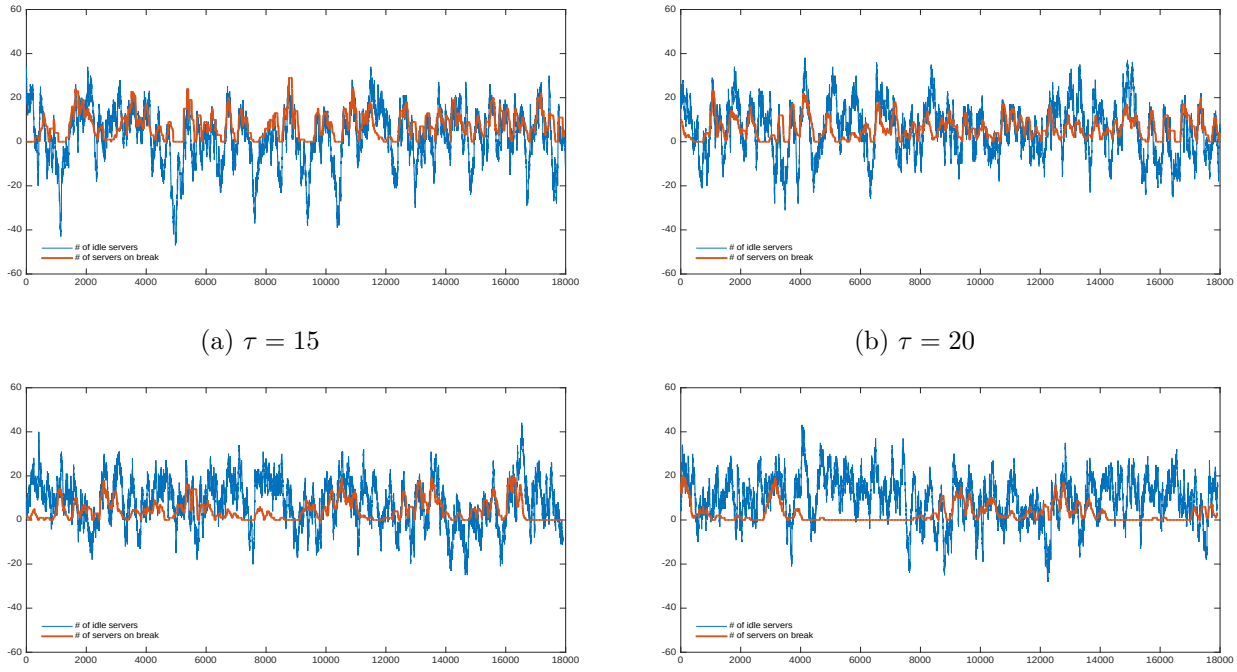
$G(t) \equiv S_b(t) - I(t)$ for six different values of τ when $\theta = 5/3$. We emphasize the positive values of $G(t)$, which when servers on break would be used; negative values show idleness. Figure 8 shows that $G(t)$ tends to decrease as τ increases.

To obtain a quantitative measure of this performance degradation caused by the servers on break, we introduce the average performance gap, defined by

$$\gamma \equiv \lim_{t\to\infty} \frac{1}{t} \int_0^t G^+(s)\, ds \tag{4.1}$$

Table 3 shows the quantitative measures of the number of servers on break with DP2(θ, τ) as a function of τ for $\theta = 5/3$.

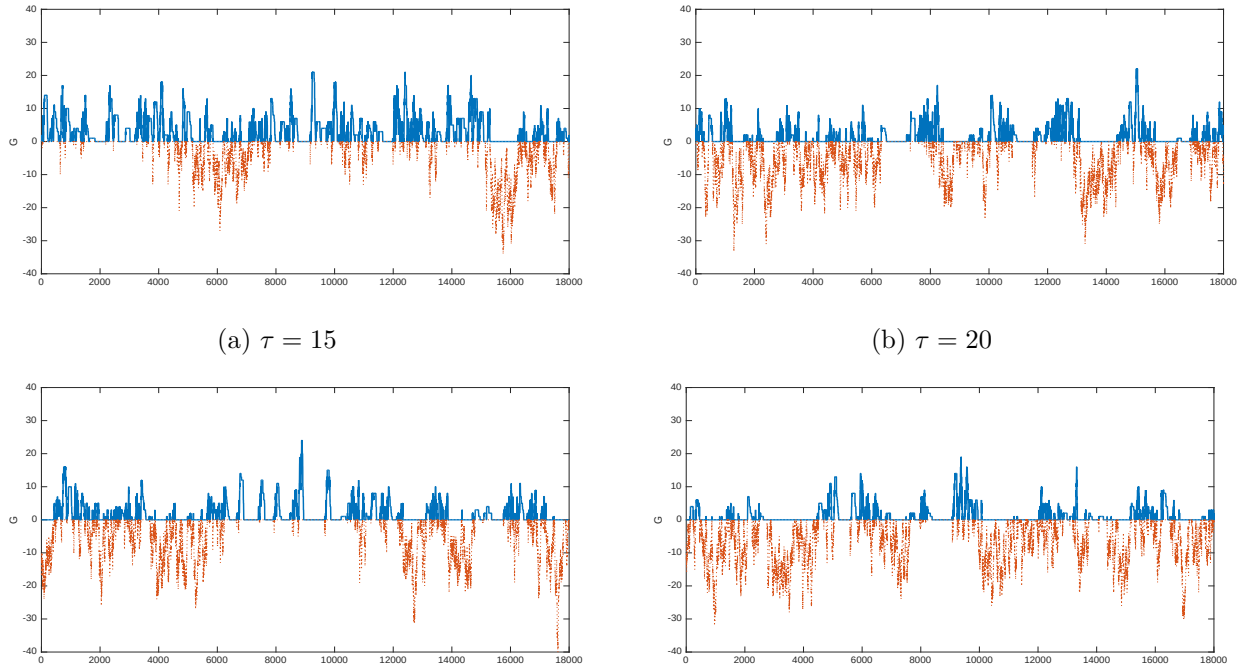
We infer from Tables 2 and 3 and Figures 6, 7 and 8 that it would be desirable to take further measures to reduce the performance degradation for customers caused by announced work breaks. That leads us to our third DP rule: DP3.



(a) $\tau = 15$ (b) $\tau = 20$

(c) $\tau = 25$ (d) $\tau = 30$

Figure 7: Sample paths of the number of servers on break, $S_b(t)$, and the number of idle servers, $I_d(t) \equiv s - N(t)$, with rule DP2 as a function of τ for $\theta = 5/3$



(a) $\tau = 15$ (b) $\tau = 20$

(c) $\tau = 25$ (d) $\tau = 30$

Figure 8: Sample paths of $G(t) \equiv S_b(t) - I(t)$, with rule DP2 as a function of τ for $\theta = 5/3$

	$E[S_b]$	$SD(S_b)$	γ
$\tau = 10$	9.0577 ± 0.0368	7.1299 ± 0.0144	3.9801 ± 0.0133
$\tau = 15$	7.4482 ± 0.0189	6.0842 ± 0.0223	2.7828 ± 0.0177
$\tau = 20$	5.7874 ± 0.0207	5.3364 ± 0.0263	1.9215 ± 0.0166
$\tau = 25$	4.3975 ± 0.0191	4.7304 ± 0.0193	1.3393 ± 0.0151
$\tau = 30$	3.3604 ± 0.0368	4.2193 ± 0.0144	0.9692 ± 0.0126
$\tau = 35$	2.5439 ± 0.0244	3.7685 ± 0.0215	0.7162 ± 0.0109
$\tau = 40$	1.9336 ± 0.0249	3.3291 ± 0.0213	0.5284 ± 0.0101

Table 3: Quantitative measures of the number of servers on break with DP2(θ, τ) as a function of τ for $\theta = 5/3$

4.3 Bounding the Number of Servers on Break: $DP3$

In order to reduce the performance degradation for customers, we now modify $DP2$ by placing a bound η on the number of servers that can be on break at any time. For given θ and τ, we thus have rule $DP3 \equiv DP(\theta, \tau, \eta)$.

Here is how rule $DP3$ works: We identify three possible states for each server: busy, idle or on break. We call servers active if they are either busy or idle. For each active server, we keep track of the elapsed time since the last break. We say that an idle server is "due for a break" if its elapsed time exceeds τ. But we will make break announcement only if the number of servers on break is less than a predetermined threshold η. In doing so, we ensure that the number of servers on break never exceeds η. If we let an idle server is told to go on break, the server's status switches to on-break and we act as if the server is no longer in the system until after the work-break is over (In the simulation, this is achieved by removing its elapsed time from the list of elapsed-times). If a server is due for a break but the number of servers on break reaches η, we do not make break announcement to that server. But we will assign a server that is "due for a break" a higher priority level (for a work break) and keep track of the elapsed time since the first high priority designation has been assigned to that server. Once a server finishes a work-break, we perform two tasks: (i) We reset its elapsed time since the end of the last break to zero and we will either assign to it a customer in queue if the queue is not empty (hence the status switches from on-break to busy) or let it stay idle if no customer is waiting (thus its status switches from on-break to idle). (ii) We look to see if there are any high priority servers that are idle. We choose from these servers the one with the longest elapsed time since it received this high priority level. Finally, In the case where there are multiple servers available when a customer enters, this customer will be assigned to the server with the shortest elapsed time.

We now study the impact of the control parameters τ and η for given θ, again focusing on the base case in which $\theta = 5/3$. Tables 4, 5 and 6 and Figure 9 display the basic performance measures P_B, p_D, $E[Q]$, $SD(Q)$, $E[T]$ and $SD(T)$ as a function of τ and η for three values of τ (20, 25 and 30) and for ten values of η, $1 \leq \eta \leq 10$. In particular, estimates of the 95% confidence intervals are shown in Tables 4, 5 and 6.

Just as for the parameter τ alone, we see that there is a strong tradeoff in the choice of η, for given τ, between the effectiveness of the breaks for the servers and the performance experienced by customers. That tradeoff is dramatically shown in Table 4. For $\tau = 20$ and $\eta = 1$, there is very little performance degradation for customers; e.g., the delay probability is about 0.27 instead of the LISF

	$\tau = 20$		$\tau = 25$		$\tau = 30$	
η	p_B	p_D	p_B	p_D	p_B	p_D
1	[0.253, 0.258]	[0.225, 0.231]	[0.249, 0.253]	[0.223, 0.228]	[0.245, 0.248]	[0.223, 0.227]
2	[0.468, 0.474]	[0.258, 0.264]	[0.457, 0.465]	[0.256, 0.260]	[0.446, 0.453]	[0.255, 0.260]
3	[0.641, 0.649]	[0.295, 0.301]	[0.626, 0.632]	[0.291, 0.296]	[0.603, 0.611]	[0.284, 0.289]
4	[0.770, 0.774]	[0.330, 0.338]	[0.755, 0.762]	[0.327, 0.333]	[0.718, 0.727]	[0.314, 0.319]
5	[0.858, 0.863]	[0.367, 0.374]	[0.841, 0.847]	[0.357, 0.363]	**[0.760, 0.772]**	[0.332, 0.338]
6	[0.917, 0.921]	[0.398, 0.406]	**[0.865, 0.873]**	[0.377, 0.382]	[0.725, 0.735]	[0.335, 0.342]
7	**[0.941, 0.946]**	[0.429, 0.437]	[0.832, 0.839]	[0.384, 0.389]	[0.697, 0.705]	[0.338, 0.345]
8	[0.926, 0.935]	[0.446, 0.454]	[0.805, 0.814]	[0.387, 0.393]	[0.672, 0.680]	[0.340, 0.346]
9	[0.904, 0.912]	[0.454, 0.461]	[0.787, 0.794]	[0.390, 0.396]	[0.664, 0.675]	[0.341, 0.348]
10	[0.891, 0.897]	[0.461, 0.467]	[0.773, 0.781]	[0.392, 0.398]	[0.655, 0.665]	[0.344, 0.351]

Table 4: 95% confidence intervals for performance measures of DP3(θ, τ, η) as a function of τ and η for $\theta = 5/3$. The bold entries are where the maximum for p_B is attained.

	$\tau = 20$		$\tau = 25$		$\tau = 30$	
η	$E[Q]$	$SD(Q)$	$E[Q]$	$SD(Q)$	$E[Q]$	$SD(Q)$
1	[2.22, 2.36]	[5.96, 6.31]	[2.19, 2.28]	[5.95, 6.15]	[2.21, 2.29]	[6.01, 6.23]
2	[2.55, 2.66]	[6.33, 6.56]	[2.55, 2.66]	[6.41, 6.69]	[2.55, 2.65]	[6.39, 6.70]
3	[2.97, 3.09]	[6.81, 7.14]	[2.92, 3.02]	[6.76, 6.97]	[2.85, 2.94]	[6.62, 6.82]
4	[3.36, 3.52]	[7.18, 7.56]	[3.31, 3.45]	[7.15, 7.48]	[3.13, 3.25]	[6.96, 7.23]
5	[3.73, 3.90]	[7.44, 7.79]	[3.61, 3.74]	[7.33, 7.59]	[3.32, 3.43]	[7.13, 7.35]
6	[4.08, 4.27]	[7.71, 8.04]	[3.83, 3.94]	[7.52, 7.72]	[3.38, 3.50]	[7.12, 7.34]
7	[4.37, 4.55]	[7.86, 8.16]	[3.91, 4.00]	[7.58, 7.75]	[3.41, 3.55]	[7.22, 7.48]
8	[4.60, 4.75]	[8.06, 8.30]	[3.95, 4.08]	[7.42, 7.76]	[3.46, 3.57]	[7.25, 7.48]
9	[4.80, 4.81]	[8.10, 8.32]	[3.97, 4.10]	[7.56, 7.76]	[3.48, 3.61]	[7.27, 7.52]
10	[4.77, 4.90]	[8.17, 8.39]	[4.01, 4.13]	[7.66, 7.86]	[3.49, 3.61]	[7.25, 7.46]

Table 5: 95% confidence intervals for the mean and standard deviation of the steady-state queue-length with DP3(θ, τ, η) as a function of τ and η for $\theta = 5/3$

value of 0.22, but the algorithm for announced breaks is ineffective; e.g. only 25% of the work breaks are announced. On the other hand, for $\tau = 20$ and $\eta = 10$, the algorithm for announced breaks is effective; e.g.. only 89% of the work breaks are announced, but there is severe performance degradation for customers, e.g., now p_D has more than doubled, reaching 0.46.

In particular, note that the probability a break is announced, p_B, is not monotone in the bound η. For $\tau = 20, 25$ and 30, the largest value of p_B is attained for $\eta = 7, 6$ and 5, respectively. These values are highlighted in Table 4. The numerical results demonstrate that the third parameter η helps, and that it should not be chosen above this bound.

As a supplement to Table 6 and panel (d) of Figure 9, Figure 10 displays estimates of the histograms of the distribution of T as a function of η for $\tau = 20$. (Corresponding ecdf's appear in the appendix.)

	$\tau = 20$		$\tau = 25$		$\tau = 30$	
η	$E[T]$	$SD(T)$	$E[T]$	$SD(T)$	$E[T]$	$SD(T)$
1	[44.61, 45.70]	[18.77, 19.43]	[44.26, 45.07]	[18.69, 19.12]	[44.51, 45.12]	[18.43, 18.86]
2	[41.85, 42.52]	[18.02, 18.50]	[41.32, 42.00]	[17.61, 17.91]	[41.48, 42.08]	[17.218, 17.568]
3	[38.69, 39.45]	[16.14, 16.38]	[38.43, 38.66]	[15.53, 15.76]	[38.43, 38.88]	[14.85, 15.06]
4	[35.69, 36.09]	[13.86, 14.12]	[35.69, 36.06]	[13.10, 13.30]	[36.20, 36.52]	[12.23, 12.42]
5	[32.54, 32.93]	[11.58, 11.89]	[33.10, 33.37]	[10.52, 10.77]	[34.22, 34.48]	[9.46, 9.64]
6	[29.87, 30.14]	[9.64, 9.89]	[30.83, 31.02]	[8.27, 8.43]	[32.63, 32.84]	[7.79, 7.97]
7	[27.57, 27.79]	[8.05, 8.23]	[29.17, 29.29]	[6.99, 7.15]	[31.72, 31.89]	[6.92, 7.11]
8	[25.87, 26.08]	[7.07, 7.24]	[28.23, 28.37]	[6.26, 6.44]	[31.18, 31.32]	[6.58, 6.74]
9	[24.77, 24.92]	[6.32, 6.50]	[27.24, 27.36]	[5.88, 6.04]	[30.76, 30.94]	[6.31, 6.45]
10	[24.19, 24.31]	[5.95, 6.14]	[27.66, 27.79]	[5.59, 5.72]	[30.49, 30.65]	[6.17, 6.28]

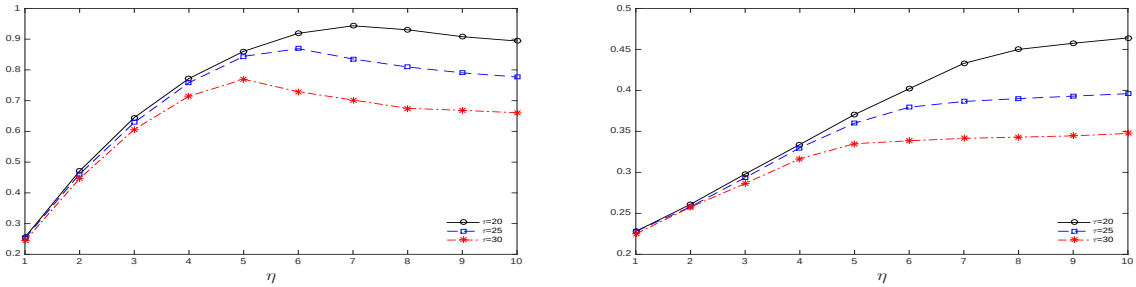
Table 6: 95% confidence intervals for the mean and standard deviation of T, the interval between successive work breaks, with DP3(θ, τ, η) as a function of τ and η for $\theta = 5/3$.

Consistent with the bold entries in Table 4, Figure 10 shows dramatic improvement as η increases up from 3 to $5-7$ and then some degradation at 8 and above.

Paralleling Figure 7, Figure 11 displays sample paths of the number of servers on break, $S_b(t)$, and the number of idle servers, $I_d(t) \equiv s - N(t)$, for six different values of η when $\tau = 20$ and $\theta = 5/3$. Panel (a) with $\eta = 9$ shows a severe performance degradation for customers because we see many places with $S_b(t)$ well above $I_d(t)$. In addition, after periods when the number of servers reaches a high level, we often observe a big downward spike in $I_d(t)$, which suggests a buildup of large queue. This is consistent with what we saw in Tables 4 and 5.

To provide an analog of Figure 8, we show in the appendix the sample paths of the gap $G(t) \equiv S_b(t) - I(t)$ for six different values of η when $\theta = 5/3$ and $\tau = 20$. Again, the positive part of $G(t)$ measures the degree of performance degradation caused by the servers on break. It shows that $G(t)$ tends to increase as η increases.

We can also expose the impact of the the threshold parameters τ and η on S_b by looking at the mean and the standard deviation of S_b estimated from simulation experiments. Based on Table 7, the number of servers on break tends to increase as τ increases. However, how the parameter η will affect S_b may depend on the specific value of τ. For example, when $\tau = 20, 25$, $E[S_b]$ increases in η; when $\tau = 30$, $E[S_b]$ first increases and then slightly decreases as η grows.



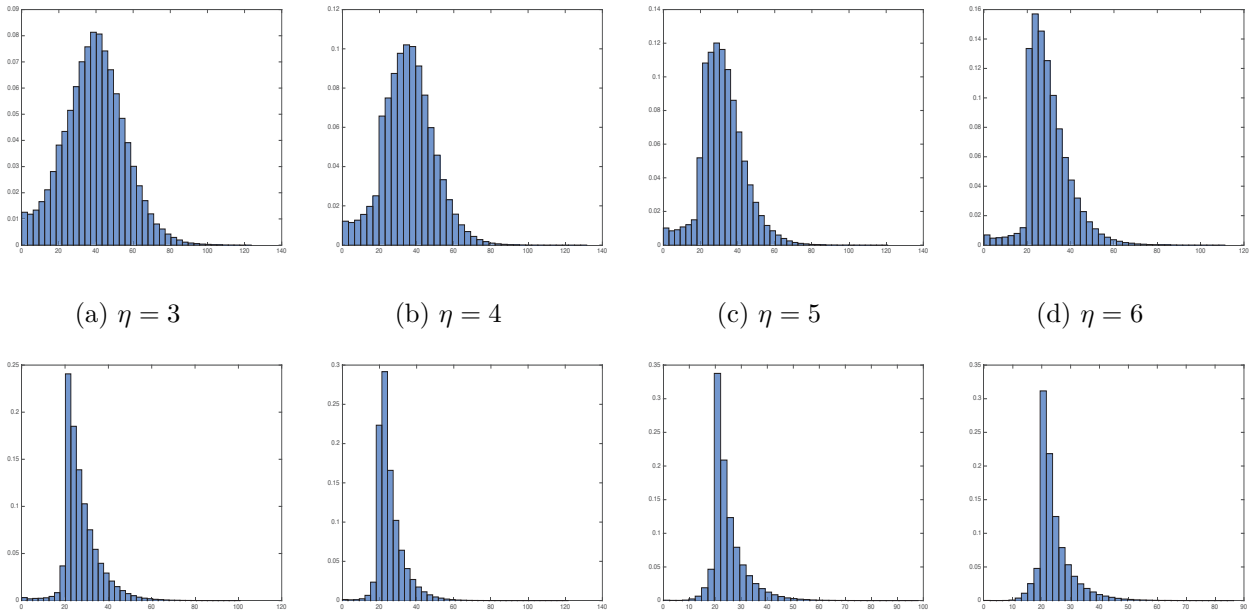
(a) Estimate of p_B

(b) Estimate of p_D

(c) Estimate of $E[Q]$

(d) Estimate of $E[T]$

Figure 9: Key performance metrics for DP3(θ, τ, η) as a function of τ and η for $\theta = 5/3$.



(a) $\eta = 3$

(b) $\eta = 4$

(c) $\eta = 5$

(d) $\eta = 6$

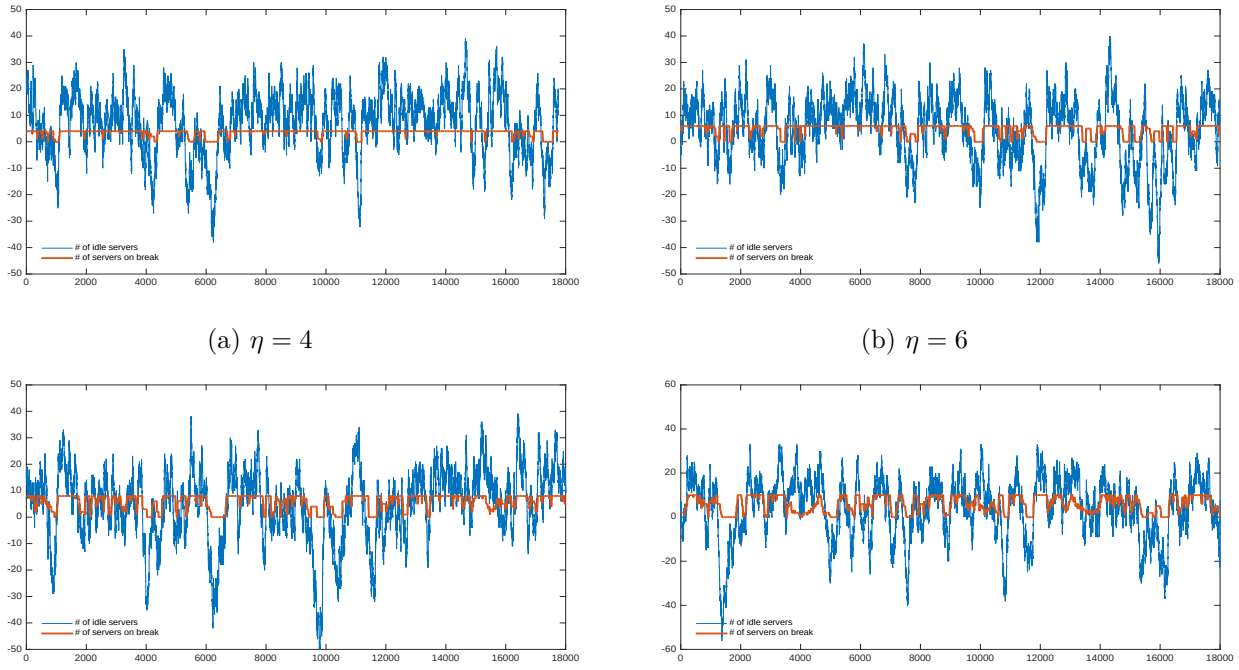
(e) $\eta = 7$

(f) $\eta = 8$

(g) $\eta = 9$

(h) $\eta = 10$

Figure 10: Histograms estimated from simulation of the inter-break-time distribution with rule DP3 as a function of η for $\theta = 5/3$ and $\tau = 20$



(a) $\eta = 4$ (b) $\eta = 6$

(c) $\eta = 8$ (d) $\eta = 10$

Figure 11: Sample paths of the number of servers on break, $S_b(t)$, and the number of idle servers, $I_d(t) \equiv s - N(t)$, with rule DP3 as a function of η for $\theta = 5/3$ and $\tau = 20$

	$\tau = 20$		$\tau = 25$		$\tau = 30$	
η	$E[S_b]$	$SD(S_b)$	$E[S_b]$	$SD(S_b)$	$E[S_b]$	$SD(S_b)$
1	[0.906, 0.911]	[0.285, 0.292]	[0.901, 0.905]	[0.294, 0.298]	[0.881, 0.885]	[0.316, 0.324]
2	[1.785, 1.793]	[0.585, 0.596]	[1.768, 1.775]	[0.602, 0.611]	[1.720, 1.728]	[0.648, 0.657]
3	[2.622, 0.639]	[0.913, 0.934]	[2.593, 2.604]	[0.940, 0.952]	[2.500, 2.513]	[1.012, 1.024]
4	[3.410, 3.435]	[1.279, 1.303]	[3.357, 3.373]	[1.315, 1.331]	[3.157, 3.278]	[1.441. 1.459]
5	[4.151, 4.181]	[1.657, 1.688]	[4.021, 4.041]	[1.733, 1.750]	[3.525, 3.562]	[1.964, 1.982]
6	[4.817, 4.852]	[2.066, 2.097]	[4.429, 4.456]	[2.229, 2.247]	[3.517, 3.548]	[2.483, 2.494]
7	[5.341, 5.376]	[2.502, 2.532]	[4.492, 4.522]	[2.734, 2.752]	[3.473, 3.504]	[2.855, 2.866]
8	[5.592, 5.628]	[2.971, 2.994]	[4.485, 4.520]	[3.140, 3.152]	[3.431, 3.463]	[3.135, 3.146]
9	[5.686, 5.730]	[3.381, 3.408]	[4.465, 4.499]	[3.457, 3.471]	[3.390, 3.428]	[3.355, 3.371]
10	[5.730, 5.764]	[3.740, 3.759]	[4.452, 4.485]	[3.708, 3.722]	[3.387, 3.427]	[3.541, 3.556]

Table 7: 95% confidence intervals for the mean and standard deviation of S_b, the number of servers on break, with DP3(θ, τ, η) as a function of τ and η for $\theta = 5/3$.

4.4 An Optimization Formulation

In order to choose the parameters τ and η, we now formulate an optimization. In particular, we suggest performing a simple optimization with a cost function that is a convex weighted sum of $1 - p_B$ and $p_D - p_D^*$, where p_D^* is the LISF value, which equals 0.215 in the base case; i.e.,

$$C \equiv C(p_B, p_D) \equiv w(1 - p_B) + (1 - w)(p_D - p_D^*), \quad 0 \le w \le 1, \tag{4.2}$$

where the weight w reflects the relative cost we wish to attribute to p_B versus p_D. (We choose this structure because we want p_B as close as possible to the maximum possible value, 1, and we want p_D as close as possible to the minimum possible value, p_D^*.) Because of the lack of monotonicity in Table 4, we see that it suffices to restrict attention to $\eta \le 7$ for $\tau = 20$.

Figure 12 shows the cost C as a function of η for $\theta = 5/3$ and $\tau = 20, 25$ and 30 and the three weights $w = 0.3, 0.5, 0.7$. Panel 12a shows that for $\tau = 20$ the optimum η^* is attained at $\eta = 5, 6$ and 7, respectively, when $w = 0.3, 0.5$ and 0.7. Panel 12c shows that for $\tau = 30$ the optimum η^* is attained at $\eta = 5$ for all these w. Panel 12d supports the use of $\tau = 20$ and $\eta = 7$ for all $w \ge 5$, which emphasizes the breaks.

4.5 Comparison with the Standard $M/M/(s-b)$ Model

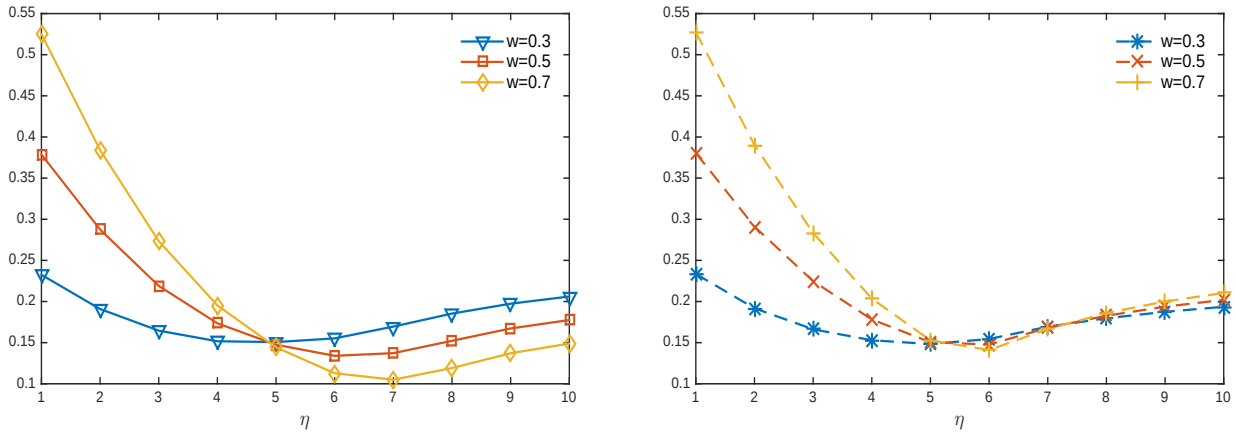
An alternative way to obtain work breaks is to place a constant number of servers on break. If we put b servers on break at all times, then we obtain an $M/M/(s-b)$ model with the customary LISF server assignment rule. It is useful to compare $DP3(\theta, \tau, \eta)$ to an $M/M/(s-b)$ LISF model by considering a range of b from $\lfloor E[S_b] \rfloor$, the greatest integer less than or equal to $E[S_b]$, to η.

In order to make performance comparisons, Table 8 displays the key performance measures for the base case with $s = 100$, $\mu = 1$ and $\rho = 0.90$ related to Tables 4 and 5.

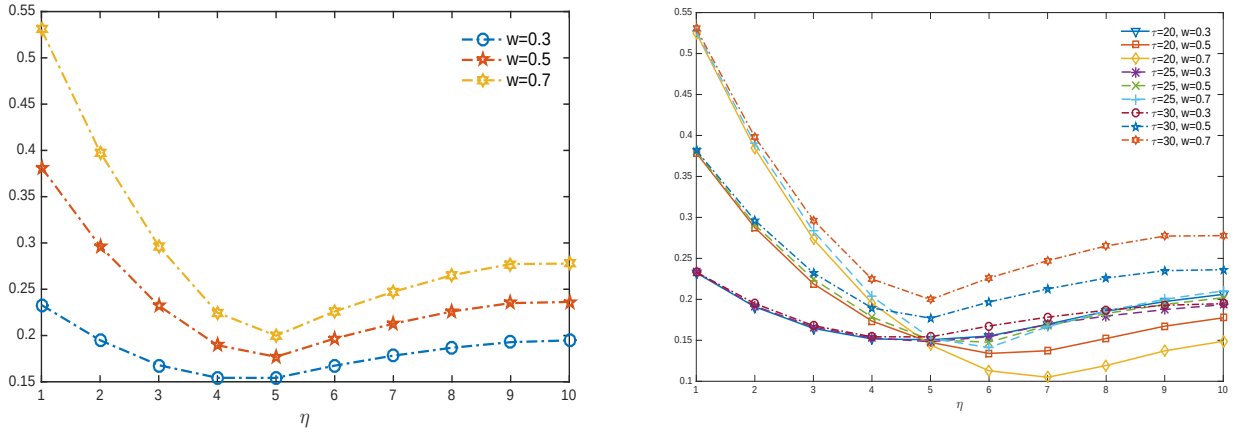
b	0	1	2	3	4	5	6	7	8
p_D	0.216	0.257	0.304	0.358	0.420	0.488	0.564	0.648	0.737
$E[Q]$	1.90	2.51	3.33	4.45	6.00	8.23	11.54	16.68	25.19
$SD(Q)$	5.62	6.69	8.01	9.65	11.75	14.49	18.20	23.48	31.00

Table 8: Performance measures for the standard $M/M/(100-b)$ queue with $\lambda = 90$ and $\mu = 1$

We find that $DP3(\theta, \tau, \eta)$ outperforms $M/M/(s-b)$ LISF for $\lfloor E[S_b] \rfloor \le b \le \eta$. For example, when $\theta = 5/3$, $\tau = 20$ and $\eta = 7$, Table 7 shows that $E[S_b] = 5.35$. Tables 4 and 5 show that $p_D = 0.43$,



(a) $\tau = 20$ (b) $\tau = 25$



(c) $\tau = 25$ (d) $\tau = 20, 25, 30$

Figure 12: Cost function for DP3(θ, τ, η) in (4.2) as a function of η and τ for $\theta = 5/3$ and $w = 0.3, 0.5, 0.7$

$E[Q] = 4.5$ and $SD(Q) = 8.0$, which are all less than the values for $b = 5$ in Table 8, and far less than the value for $b = \eta = 7$.

5 Conclusion

In this paper we developed new rules for assigning idle servers to arriving customers in a contact center in order to create effective work breaks from available idleness. The first one-parameter dynamic priority rule DP1 yields unannounced breaks while maintaining work conservation, while the final three-parameter refined rule DP3 yield announced breaks by sacrificing work-conservation.

After specifying the model, discussing important conservation laws and introducing our base case in §2, in §3 we started by developing convenient exact and approximate formulas for (i) the steady-state number of idle servers, (ii) the idle-time distribution and (iii) the cumulative idleness over an interval

[0,t] for the classic longest-idle-server-first (LISF) rule for assigning idle servers to new customers. This analysis showed that the servers usually would not experience an adequate work break during a day. We showed that the random routing (RR) rule proposed in Mandelbaum et al. (2012) produces quite different idle times, but also would seldom produce a work break during a day.

Then we introduced three new dynamic priority rules designed to create work breaks out of available idleness. Each succeeding rule builds on the one before. The first DP rule, $DP1(\theta)$ depends only on the target break duration θ; it assigns idle servers to the new arrival in the order of elapsed time since their last break end time. We showed by simulation experiments that long idle times of duration θ can be created by $DP1$. We found that by varying the control parameter θ, we can adjust the frequency at which the breaks occur. There were three shortcomings of $DP1$: First, long idle times less than the target often occur just before the break; second, the intervals between successive breaks tend to be too long; and, third, it is not possible to announce when the idle time will serve as a work break.

The other DP rules are designed to address those shortcomings. First, $DP2 \equiv DP(\theta, \tau)$ makes work-break announcement whenever a server becomes idle and its elapsed time since the last work break end time exceeds a threshold time τ. We found that, with DP2, we are able to generate announced work breaks, but with the loss of work conservation, there could be quite severe degradation in performance.

We therefore proposed $DP3 \equiv DP3(\theta, \tau, \eta)$. Roughly speaking, DP3 acts just like DP2 except that we place a restriction η on the maximum possible number of servers allowed to be on break at any given time. Servers that are "due for a break" are placed in a "queue for breaks", see §4.3 for more details. We have demonstrated by performing simulation experiments that DP3 (with properly chosen control parameters) is remarkably effective for both generating announced work breaks reasonably well without sacrificing much operational performance. In §4.4 we created an optimization framework to choose the parameters τ and η for any given θ. In §4.5 we showed that the performance of DP3 is better than the LISF with a larger fixed number of servers on break.

Much work remains to be done in the future. While we have shown that it is possible to create within-day work breaks from available idleness, it remains to investigate whether or not these rules would improve the satisfaction of service representatives. Second, we applied simulation to describe the performance of the DP rules, e.g., to find the idle-time distribution. It remains to develop supporting analytical formulas, either exact or asymptotic, and supporting optimality theory.

Acknowledgement

Research support from NSF (CMMI 1265070 and 1634133) is gratefully acknowledged.

References

Aksin OZ, Armony M, Mehrotra V (2007) The modern call center: a multi-disciplinary perspective on operations management research. *Production Oper. Management* 16:665–688.

Armony M, Ward A (2010) Fair dynamic routing policies in large-scale service systems with heterogeneous servers. *Oper. Res.* 58(3):624–637.

Atar R (2008) Central limit theorem for a many-server queue with random service times. *Ann. Appl. Prob* 18(4):1548–1568.

Atar R, Shaki YY, Shwartz A (2011) A blind policy for equalizing cumulative idleness. *Queueing Systems* 67(4):275–293.

Biron M, Bamberger P (2010) The impact of structural empowerment on individual well-being and performance: Taking agent preferences, self-efficacy and operational constraints into account. *Human Relations* .

Browne S, Whitt W (1995) Piecewise-linear diffusion processes. Dshalalow J, ed., *Advances in Queueing*, 463–480 (Boca Raton, FL: CRC Press).

Chan W, Koole G, L'Ecuyer P (2014) Dynamic call center routing policies using call waiting and agent idle times. *Management Science* 16(4):544–560.

Fritz C, Ellis AM, Demsky CA, Lin BC, Guros F (2013) Embracing work breaks. *Organizational Dynamics* 4(42):274–280.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.

Jett QR, George JM (2003) Work interrupted: A closer look at the role of interruptions in organizational life. *Academy of Management Review* 28(3):494–507.

Lin YH, Chen CY, Hongand WH, Y-CLin (2010) Perceived job stress and health complaints at a bank call center: comparison between inbound and outbound services. *Industrial health* 48(3):349–356.

Mandelbaum A, Momcilovic P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* 58(7):1273–1291.

Mayo E (1933) *The Human Problems of an Industrial Civilization* (Glenville, IL: Scott Foresman).

Sawyerr OO, Srinivas S, Wang S (2009) Call center employee personality factors and service performance. *Journal of Services Marketing* 23(5):301–317.

Sisselman MJ, Whitt W (2007) Value-based routing and preference-based routing in customer contact centers. *Production Oper. Management* 16(3):277–291.

Taylor FW (1911) *Principles of Scientific Management* (New York: Harper and Brothers).

Trougakos JF, Hideg I (2009) Momentary work recovery: The role of within-day work breaks. Sonnentag S, Perrewe PL, Ganster DC, eds., *Research in Occupational Stress and Well Being* (Emerald Group, Bingley, UK).

Whitt W (2000) Limits for cumulative input processes to queues. *Probability in the Engineering and Informational Sciences* 14:123–150.

Whitt W (2002) *Stochastic-Process Limits* (New York: Springer).

Whitt W (2004) A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research* 52(6):922–941.

Whitt W (2006) The impact of increased employee retention upon performance in a customer contact center. *Manufacturing and Service Oper. Management* 81(3):221–234.