# 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

June 30, 2017

**Abstract**

We develop new rules for assigning available service representatives to customers in customer contact centers and other large-scale service systems in order to create effective work breaks for the service representatives from naturally available idleness. These are unplanned breaks occurring randomly over time. We consider both announced breaks as well as unannounced breaks. Our goal is to make the mean and variance of the interval between successive breaks suitably small. Given a target break duration, we propose assigning idle servers based on the elapsed time since their last break. We show that our proposed server-assignment rules are optimal for the many-server heavy-traffic (MSHT) fluid model. Extensive simulation experiments support the proposed server-assignment rules in practical cases and confirm the MSHT approximation formulas when the number of servers is very large.

*Keywords:* work breaks; server-assignment rules; customer contact centers, large-scale service systems; many-server heavy-traffic limits; fluid models.

# 1 Introduction

In this paper we apply queueing models to investigate new rules for assigning available (idle) servers to customers that redistribute the cumulative idleness to create effective work breaks for the service representatives. In doing so, we identify two different kinds of unplanned work breaks, unlike the conventional planned breaks that can be part of a daily schedule posted in advance: (i) random *announced breaks*, and (ii) random *unannounced breaks*. For announced breaks, the server is told they will be on break when the break is announced, so that they are "off duty" during the break; for unannounced breaks, the servers are not told, so that they are always "on call" if needed to meet customer demand.

We were motivated by customer contact centers (call centers), but concern about the server experience also arises more widely, e.g., in the evolving sharing economy, such as ad-hoc taxi services. For customer contact centers, there is now a substantial body of research 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) and 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 can be monotonous. This negative impact can decrease 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).

## 1.1 Our Objectives

We first consider unannounced breaks and then afterwards announced breaks. Servers would naturally prefer announced breaks, but unannounced breaks are attractive because, unlike announced breaks, they can be work-conserving (non-idling); i.e., no customer waits in queue if there is an available server, so that customers experience no performance degradation.

Our broad goal is to determine if it is possible to redistribute idleness to create effective work breaks and, if so, how to do so. For that purpose, we assume that we have a standard $M/GI/n$ queueing model with $n$ homogeneous servers working in parallel and unlimited waiting space. We assume that there is a target break duration $\theta$; we call any idle time exceeding $\theta$ a break.

Motivated by call centers, for our simulation examples we focus on a *base case*, which is the $M/M/n$ model with $n = 100$ servers, traffic intensity $\rho = 0.9$, mean service time $E[S] = 1$ and $\theta = 5/3$. We are thinking of mean service times of 3 minutes, so we measure time in units of 3 minutes. Roughly, we would like to obtain a 5-minute break every $1 - 2$ hours. That goal translates to a break of length 5/3 every time interval of $20 - 40$. That goal is feasible for $\rho = 0.9$ because each server is idle $(1 - \rho) \times 100\% = 10\%$ of the time, which is 6 minutes every hour or 12 minutes every two hours.

We first study unannounced breaks. To evaluate them, we introduce a specific criterion. Let $T \equiv T(\theta)$ be the steady-state interval between successive breaks, i.e., the elapsed time from the end of one break to the end of the next. Our main goal is to minimize $E[T]$.

However, we also want to control the variability of $T$, which we represent by the standard deviation $SD(T)$. We want both $E[T]$ and $SD(T)$ to be suitably small. We will consider a strong form of optimality involving lexicographical order in which we first minimize $E[T]$ and then, from the set of optimal policies, minimize the standard deviation $SD(T)$.

## 1.2 Our Main Contributions

(i) We show that the standard longest-idle-server-first (LISF) server-assignment rule and the natural alternative random routing (RR) rule, which generate unannounced breaks, generate the breaks too infrequently.

(ii) We introduce server-assignment rules that assign idle servers according to the elapsed time since their last break ended, which we call "the age." We first assign idle servers who have completed a break (are experiencing an idle time greater than or equal to $\theta$), assigning the idle server with the largest elapsed idle time first. After all those servers are assigned, we assign the idle servers not currently on break (with current idle times less than $\theta$), assigning the server with the least age first. Thus we always assign the idle server least due a break. We call this first server-assignment rule $D_1 \equiv D_1(\theta)$, using $D$ for "*d*ynamic priority" and "*d*ue for a break."

(iii) We show that important insight into this server-assignment problem can be gained by considering the deterministic fluid model that arises in the many-server heavy-traffic (MSHT) fluid limit in which the arrival rate and number of servers are allowed to grow, while the service-time distribution is held fixed. In particular, we show that the $D_1$ rule and the variant introduced for announced breaks are both optimal for the fluid model, first minimizing $E[T]$ and then minimizing $SD(T)$. Explicit formulas for the steady-state performance show that (i) the distribution of the random interval between breaks, $T$, is insensitive to the arrival process beyond its rate, (ii) the mean $E[T]$ is also insensitive to the service-time distribution beyond its mean, but (iii) the standard deviation $S(T)$ increases with increasing service-time variability.

(iv) We also consider announced work breaks, for which we necessarily lose the non-idling property (servers on break remain idle even if customers wait in queue). We propose a modification of the rule $D_1(\theta)$ for announced breaks: With $D_2 \equiv D_2(\theta, \tau, \eta)$ we announce a work break whenever the age exceeds a threshold $\tau$. (For a busy server, the break begins upon service completion; for an idle server, the break begins immediately.) During the break, the server is then off duty, and so unavailable to serve new demand until the break is over. In addition, we impose an upper bound $\eta$ on the number of servers that can be on break at any time. If a server cannot be given a break, it is given high priority for a future break.

(v) Finally, we report results of extensive simulation experiments. These simulation experiments show for the base case with $n = 100$ that the new server-assignment rules are effective. For large $n$, the simulations confirm the MSHT fluid formulas.

### 1.3 Related Literature and Organization

Other studies have recognized and responded to the preferences and concerns of the service representatives. First, Whitt (2006b) developed a mathematical model to help analyze the benefit in contact-center performance gained from increasing employee 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).

There is a large literature on MSHT limits and approximations. The MSHT fluid model for the steady-state performance in §3 is a variant of the standard MSHT fluid model with the first-come first-served (FCFS) service discipline and, if considered, the LISF server-assignment rule, in Whitt (2006a), Liu and Whitt (2012a) and Kaspi and Ramanan (2011), but here we consider the underloaded quality-driven (QD) regime. Convergence to steady-state for that standard fluid model is considered in §5 of Liu and Whitt (2011) and in Theorem 3.9 and §6 of Kaspi and Ramanan (2011). For the standard model, MSHT limits are established in Kaspi and Ramanan (2011) and Liu and Whitt (2012b, 2014). Since we are considering the QD MSHT regime, the standard MSHT limit is the same as for the infinite-server system in Theorem 3.1 of Pang and Whitt (2010).

This paper is organized as follows: In §2 we formalize the work-conserving server-assignment rules and introduce a general Markov process that describes the evolution of the system state for the $D_1$ rule. We also discuss important conservation laws and show that breaks occur too infrequently with the LISF and RR rules. In §3 we establish our results for the MSHT fluid model. We report results of simulation experiments for the $D_1$ rule yielding unannounced breaks in §4 and for the $D_2$ rule yielding announced breaks in §5. Finally, in §6 we draw conclusions. We present additional supporting material in the online supplement. In particular, we describe how we implemented the server-assignment rules $D_1$, SISF and $D_2$ in our simulations; we present distribution and renewal process details for the case of hyperexponential service times; and we present additional simulation results.

## 2 The Stochastic Model for Server-Assignment Rules

We consider the standard $M/GI/n$ multi-server queueing model with $n$ homogeneous servers working in parallel and unlimited waiting space with customers assigned to service in a first-come first-served (FCFS) order. The service times come from a sequence of independent and identically distributed (i.i.d.) random variables $S_i$ having finite mean and variance and cumulative distribution function (cdf) $F$ having a probability density function (pdf) $f$, with $F(t) = \int_0^t f(s)\, ds$, $t > 0$. 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. Then the variance coincides with the squared coefficient of variation (scv, variance divided by the square of the mean), which we denote by $c_s^2$. There is a Poisson arrival process with arrival rate $\lambda \equiv \rho < 1$ 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 $EU = 1/\rho$. We also assume that there is a specified target break duration $\theta$. We call any idle time of length $\theta$ or longer a (work) break.

### 2.1 The Server-State Stochastic Process Accounting for Breaks

For our server-assignment rules, we maintain the state of each server, including the elapsed time since the last break. Let $S_k(t)$ be the state of server $k$, $1 \leq k \leq n$, for some designated order of the servers. Let the possible values of $S_k(t)$ be vectors of real numbers $(b, a, c)$ in the set $\Sigma \equiv \{0, 1\} \times [0, \infty)^2$, where $b$ is an indicator variable with $b = 1$ if the server is *busy* serving a customer and $b = 0$ if the server is idle, $a$ is the *age*, i.e., the elapsed time since the last break, and $c$ is the elapsed time of the *current busy period* if the server is busy or of the *current idle period* if the server is idle. Thus the state of all servers at time $t$ is given by the vector $S(t) \equiv (S_1(t), \ldots, S_n(t))$ taking values in the set $\Sigma^n$. The state of the full system at time $t$ is then $(Q(t), S(t))$, where $Q(t)$ is the number of customers in the system. The overall state space is thus $\mathcal{S} \equiv \mathcal{N} \times \Sigma^n$, where $\mathcal{N}$ is the set of nonnegative integers.

The stochastic process $(Q, S) \equiv \{(Q(t), S(t)) : t \geq 0\}$ evolves over time as a consequence of arrivals, service completions and server assignments. Arrivals are generated exogenously by the Poisson arrival process with rate $\rho$, while service completions occur an independent random service time with cdf $F$ after the server has been assigned to the customer. (There are no service interruptions.) Hence, to understand the full evolution of the system, it only remains to specify how the servers are assigned to customers.

### 2.2 Work-Conserving Server-Assignment Rules

Server-assignment rules can be classified into two types: work-conserving or non-work-conserving. Work-conserving (or non-idling) policies immediately assign one of the idle servers to a customer whenever there is a customer in need of service (in queue or upon arrival) and there is an idle server. Non-work-conserving policies might let the customer wait in queue until a later time. These notions are important for us because announced work breaks require policies that are in general non-work-conserving, whereas unannounced work breaks do not. To quickly see why announced breaks require non-work-conserving policies, note that a server could be on a break of duration $\theta$ when a customer arrives; that customer will wait in queue if there are no other servers available.

The problem of choosing a good work-conserving server-assignment policy can be formulated as a stochastic decision process. We can formulate a discrete-time general-state Markov decision problem as in Puterman (2005) if we let the discrete times be the successive arrival epochs and the service completion times, but we will look at the policy $D_1$ and other work-conserving policies directly in continuous-time.

The server-assignment policies operate only when a server assignment is needed and at least one idle server is available; we call that a server-assignment time. We will consider only stationary Markov service-assignment rules, which at any server-assignment time $t$ depend only on the state $S(t)$ at time $t$ and are otherwise independent of $t$. A deterministic server-assignment rule is thus a map $\pi : \Sigma^n \rightarrow \{1, \ldots, n\}$ taking the server state $S(t)$ at time $t$ into the index of the server to be assigned at time $t$; we thus write $\pi(t) = \pi(S(t))$. A randomized server-assignment rule is a map $\pi : \Sigma^n \rightarrow \mathcal{P}(\{1, \ldots, n\})$, where $\mathcal{P}(\{1, \ldots, n\})$ is the space of probability distributions on the set $\{1, \ldots, n\}$. In this case, $\pi(S(t))$ maps the state $S(t)$ at time $t$ into a probability distribution on the indices of the server to be assigned at time $t$.

To formalize the work-conserving server-assignment policies we consider, let $\mathcal{I}(t)$, $\mathcal{E}(t)$ and $\mathcal{N}(t)$ be the sets of servers that, at time $t$, are idle, idle and currently experiencing a break, and idle but not experiencing a break, respectively; i.e., $\mathcal{I}(t) \equiv \{k : S_{k,1}(t) = 0, 1 \leq k \leq n\}$, $\mathcal{E}(t) \equiv \{k : S_{k,3}(t) \geq \theta, k \in \mathcal{I}(t)\}$ and $\mathcal{N}(t) \equiv \{k : S_{k,3}(t) < \theta, k \in \mathcal{I}(t)\}$. First the *Longest-Idle-Server-First* (LISF) policy assigns the idle server that has been idle the longest, i.e.,

$$\pi_{LISF}(t) \equiv \arg\max \{S_{k,3}(t),\ k \in \mathcal{I}(t)\}. \tag{2.1}$$

The RR rule is a randomized rule that assigns each server in $\mathcal{I}(t)$ with equal probability.

The new $D_1$ rule first assigns the server in $\mathcal{E}(t)$ *that has experienced the longest break*, but if no server has completed a break, then $D_1$ assigns the server in $\mathcal{N}(t)$ *least due a break*; i.e.,

$$\begin{aligned} \pi_{D_1}(t) &\equiv \arg\max \{S_{k,3}(t) : k \in \mathcal{E}(t)\} \quad \text{if} \quad \mathcal{E}(t) \neq \phi, \quad \text{and} \\ &\equiv \arg\min \{S_{k,2}(t); k \in \mathcal{N}(t)\} \quad \text{if} \quad \mathcal{E}(t) = \phi \quad \text{and} \quad \mathcal{N}(t) \neq \phi. \end{aligned} \tag{2.2}$$

We also consider a myopic modification of $D_1$ which we call the *shortest-(least)-idle-server-first* (SISF) rule, which first looks for servers experiencing a break, just like $D_1$, but if there are none, then assigns the server whose current idle time is least, i.e.,

$$\begin{aligned} \pi_{SISF}(t) &\equiv \arg\max \{S_{k,3}(t); k \in \mathcal{E}(t)\} \quad \text{if} \quad \mathcal{E}(t) \neq \phi, \quad \text{and} \\ &\equiv \arg\min \{S_{k,3}(t); k \in \mathcal{N}(t)\} \quad \text{if} \quad \mathcal{E}(t) = \phi \quad \text{and} \quad \mathcal{N}(t) \neq \phi. \end{aligned} \tag{2.3}$$

Note that the age plays no role for SISF.

## 2.3 A Function-Valued Continuous-Time Markov Process

To understand the approximating deterministic fluid model for the policy $D_1$ and other work-conserving policies introduced in §3, it is convenient to consider an alternative continuous-time representation.

Given that we have no special interest in individual servers, we can focus on associated counting processes. In particular, now using the subscript $n$ to denote the stochastic model with $n$ servers, let

$$\begin{aligned} B_n(t,x,y) &\equiv \sum_{k=1}^{n} 1_{\{S_{k,1}(t)=1, S_{k,2}(t)\le x, S_{k,3}(t)\le y\}} \quad \text{and} \\ I_n(t,x,y) &\equiv \sum_{k=1}^{n} 1_{\{S_{k,1}(t)=0, S_{k,2}(t)\le x, S_{k,3}(t)\le y\}}, \end{aligned} \tag{2.4}$$

where $1_A$ is the indicator function of the set $A$; i.e., $1_A = 1$ on $A$ and $1_A = 0$ otherwise, so that $B_n(t,x,y)$ is the number of busy servers at time $t$ with age at most $x$ and elapsed current service time at most $y$, while $I_n(t,x,y)$ is the number of servers that are idle at time $t$ with age at most $x$ and elapsed idle time (since their last service completion) at most $y$. (Necessarily, $x \ge y$ for $I_n(t,x,y)$.)

Thus, $B_n \equiv \{B_n(t,\cdot,\cdot) : t \ge 0\}$ and $I_n \equiv \{B_n(t,\cdot,\cdot) : t \ge 0\}$ can each be regarded as a stochastic process with values in $\mathcal{D}^2$, where $\mathcal{D}$ is the function space of all right-continuous real-valued functions with left limits, as in Whitt (2002), while $\mathcal{D}^2 \equiv \mathcal{D} \times \mathcal{D}$ is the usual two-fold product space. Aside from customer identity, the stochastic process $(Q_n, B_n, I_n) \equiv \{(Q_n(t), B_n(t,\cdot,\cdot), I_n(t,\cdot,\cdot)) : t \ge 0\}$, where $Q_n(t)$ is again the number in system at time $t$, is equivalent to the stochastic process $(Q, S)$ in §2.1 (with subscript $n$ added now). Let $B_n(t) \equiv B_n(t,\infty,\infty)$ be the number of busy servers at time $t$; and let $I_n(t) \equiv I_n(t,\infty,\infty)$ be the number of idle servers at time $t$. We clearly have $B_n(t) = \min\{Q_n(t), n\}$ and $I_n(t) = \max\{n - Q_n(t), 0\}$.

For the $M/GI/n$ model with $\rho < 1$ and the $D_1$ server-assignment rule, it is evident that the stochastic process

$$(Q_n, B_n, I_n)_t \equiv (Q_n(t), B_n(t,\cdot,\cdot), I_n(t,\cdot,\cdot)) \equiv \{\{(Q_n(t), B_n(t,x,y), I_n(t,x,y) : x \ge 0, y \ge 0\} : t \ge 0\} \tag{2.5}$$

as a function of $t$ is a Markov process with general state space. We will be interested in the steady-state behavior, which we assume is well defined. In particular, with $\Rightarrow$ denoting convergence in distribution, we assume that there exists a random element $(Q_n, B_n, I_n)_\infty$ such that, for any initial state $(Q_n, B_n, I_n)_0$, $(Q_n, B_n, I_n)_t \Rightarrow (Q_n, B_n, I_n)_\infty$ as $t \to \infty$, and if the initial state $(Q_n, B_n, I_n)_0$ is the limit $(Q_n, B_n, I_n)_\infty$, then $(Q_n, B_n, I_n)_t$ becomes a stationary stochastic process, distributed as $(Q_n, B_n, I_n)_\infty$ for all $t$. (We conjecture that this conclusion can be proved as a theorem, but it does not follow immediately from standard Markov process theory because the state space is uncountably infinite.) When we refer to the steady-state quantities, we omit the index $t$.

**Remark 2.1** (*stochastic process for any work-conserving rule*) It is significant that the stochastic process $\{\{Q_n(t), B_n(t,\infty,y), I_n(t,\infty,\infty) : y \ge 0\} : t \ge 0\}$ is the same for any work-conserving server-assignment rule. The (work-conserving) server-assignment rule only alters the server ages and current idle times, which are excluded from the general form in (2.5) by the arguments assigned the value $\infty$ in this representation.

## 2.4 Conservation Laws

In this section we consider general server-assignment rules, both work-conserving and not, subject to the regularity conditions that (i) all arrivals are eventually served, (ii) customer service times are not altered by any of the server-assignment rules and (iii) there is a well defined steady state (so we now omit $t$). We have just formulated the $D_1$ rule and assumed that it satisfies condition (iii). In this general setting, conservation laws are important for understanding allocations of idleness.

First, the following (well known) expressions for the steady-state mean values follow from Little's law, e.g., see Whitt (1991):

$$E[B_n] = \rho n \quad \text{and} \quad E[I_n] = (1-\rho)n, \tag{2.6}$$

where $B_n \equiv B_n(\infty, \infty)$ and $I_n \equiv I_n(\infty, \infty)$. Formula (2.6) implies that, regardless of the server-assignment rule, on average 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 (2.6) remains unchanged. Henceforth, we omit the subscript $n$ except for $(Q_n, B_n, I_n)$.

Let $V$ denote the steady-state interval between successive service times (now omitting the subscript $n$ even though the distribution of $V$ depends on $n$), with $V$ taking on the value 0 when the server is immediately reassigned. Given that each server experiences alternating service times with $E[S] = 1$ and idle times, we have the relations

$$1 - \rho = \frac{E[V]}{E[V]+1}, \quad \text{so that} \quad E[V] = \frac{1-\rho}{\rho} \quad \text{for all} \quad n. \tag{2.7}$$

From (2.7), we see that, for given $\rho$, the number of servers and the server-assignment rule cannot alter $E[V]$.

Let $D$ be the duration of a break (an idle time of at least $\theta$) and let $T$ be the interval between successive breaks (end-to-end, in steady state). Let $\beta$ be the rate breaks occur, let $\pi_\beta$ ($\pi_{\beta,I}$) be the long-run proportion of time (of the idle time) during which each server is on break. As further conservation relations, we have

$$\beta = \frac{1}{E[T]}, \quad \pi_\beta = \frac{E[D]}{E[T]} \quad \text{and} \quad \pi_{\beta,I} = \frac{\pi_\beta}{1-\rho}. \tag{2.8}$$

We now apply these relations to characterize the rate at which breaks occur. Consistent with intuition, the maximum possible rate at which breaks could occur is when all idle times are either $\theta$ or 0.

**Theorem 2.1** (*the rate breaks occur*) *Given $\rho$ and $\theta$, the rate at which breaks occur is a function of the distribution of the idle time $V$, in particular,*

$$\beta = \frac{(1-\rho)P(V \geq \theta)}{E[V]} = \rho P(V \geq \theta), \tag{2.9}$$

*so that*

$$\beta \leq \beta^* \equiv \frac{1-\rho}{\theta}. \tag{2.10}$$

*The upper bound $\beta^*$ in* (2.10) *is attained if a proportion $p \equiv E[V]/\theta = (1-\rho)/(\rho\theta)$ of the idle times are $\theta$ and the rest are* 0.

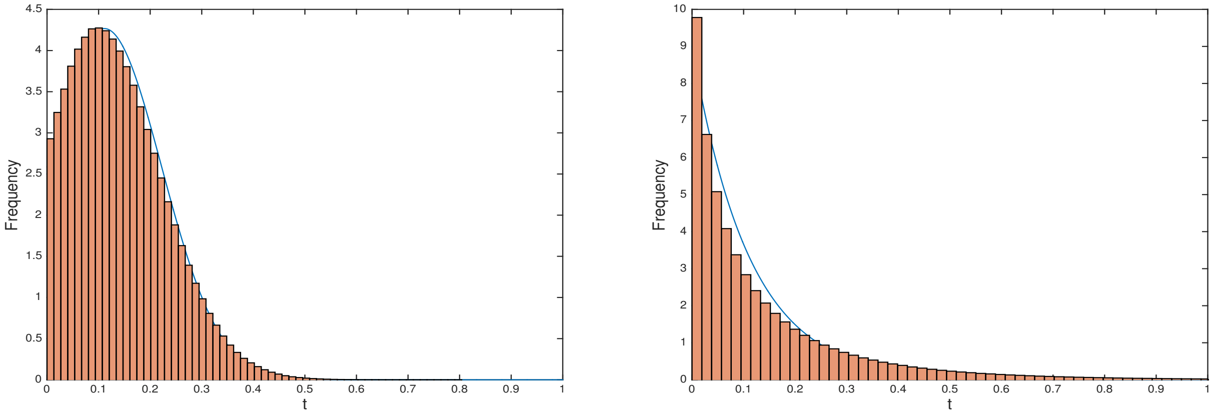
**Proof.** First, we can combine (2.7) and (2.8) to obtain (2.9). Then we can apply Markov's inequality with (2.9) and (2.7) to obtain (2.10). Finally, it is easy to check that this bound is attained by the two-point distribution concentrating on $\{0, \theta\}$. ■

**Remark 2.2** (*attaining and approaching the bound*) For the $M/GI/n$ model, it is evident that the bound $\beta^*$ on the rate breaks occur cannot be attained by any work-conserving server-assignment rule, because we cannot force all idle times to be either 0 or $\theta$. However, we conjecture that $D_1$ attains this upper bound asymptotically in the MSHT limit as $n \to \infty$. In §3 we provide strong support for that conjecture by showing that this upper bound is attained in the deterministic fluid model that should arise in the MSHT limit.

### 2.5 LISF and RR in the Base Case

We started our research by studying the idleness in the $M/M/n$ model with the LISF and RR server-assignment rules. For the $M/M/n$ base case with $n = 100$, $\rho = 0.9$,. $E[S] = 1$ and target break $\theta = 5/3$ to represent 5 minutes, (2.7) implies that the (expected) cumulative idleness over over $[0, 40]$ (or 2 hours), is 4 (or 12 minutes), which is evidently sufficient to produce effective work breaks, but the LISF and RR rules do not generate them frequently enough. To illustrate, Figure 1 shows histograms estimated by simulation of the steady-state idle-time pdf with LISF and RR for the base case. (The atom at time 0 is omitted from the histogram.) 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 sufficient, because there is neglible mass above 1.0, but we need at least 1.67 to get a break.

Consistent with Figure 1, Our analysis indicates that, in the base case, LISF produces a steady-state idle time $V$ that has a distribution that is approximately a truncated Gaussian distribution having



(a) LISF

(b) RR

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

$P(V = 0) = 0.215$, $E[V] = (1 - \rho)/\rho = 0.1111$ and $SD(V) = 0.100$. Since $\theta = 5/3$ is 15.7 standard deviations above the mean, it is highly unlikely that an idle time will be a break.

In contrast, with RR, our analysis indicates that $V$ has a distribution that is approximately a mixture of exponential distributions, having $E[V] = (1 - \rho)/\rho = 0.1111$ and $SD(V) = 0.176$. the standard deviation is larger than for LISF but still the target $\theta$ is more than 9 standard deviations above the mean.

# 3 The MSHT Fluid Model for the $D_1$ Server-Assignment Rule

In this section we present our main theoretical result, concluding that the new $D_1$ server-assignment rule achieves the maximum possible rate of breaks in Theorem 2.1 for the fluid model. In §3.1 we introduce the fluid model. In Remark 3.3 we briefly discuss the MSHT limit (not proved here) that supports using the fluid model as an approximation for the stochastic model. In §3.2 we review properties from renewal theory that we use. In §3.3 we state and prove our main result. Finally, in §3.4 we discuss the SISF server-assignment rule, which yields the same mean time between breaks, but a higher variance.

## 3.1 The Deterministic MSHT Fluid Model for $D_1(\theta)$

We now consider the deterministic fluid model that approximates the $M/GI/n$ model with the $D_1$ server-assignment rule. In this model we replace discrete customers and discrete servers that experience random service times by continuous divisible deterministic fluid. For treating server idleness, it is convenient to consider both customer fluid and server fluid. Customer fluid arrives exogenously over time at rate $\rho < 1$. We let the service capacity be 1, so that we are considering an underloaded fluid model. In this underloaded deterministic model, starting empty or in steady state, there is never any customer fluid waiting in queue.

Somewhat informally, the individual atoms of customer fluid arrive to be served and enter service immediately upon arrival, where each atom of customer fluid is matched with an atom of server fluid from the pool of idle service capacity to provide that service. Thus, both customer fluid and server fluid arrive at the service facility at rate $\rho$, where these are joined to provide service. Consistent with (2.6), in steady state the quantity of customer fluid in service (together with server fluid) at each time is $\rho$, while the quantity of idle server fluid is $1-\rho$. The total service capacity is the sum: $\rho + (1-\rho) = 1$.

**Remark 3.1** (*the role of proportions*) While it is natural to think of the experience of individual atoms of fluid as following stochastic processes, as in the paragraph above, but that can be formalized using proportions. For example, a major component of the stochastic model is a sequence of random service times. This is a sequence of i.i.d. random variables each distributed as a random variable $S$ with cdf $F$. In the fluid model, these distributions should be interpreted as proportions. For the fluid model, we understand that $F(x)$ is the proportion of fluid that is served within time $x$ after it started service. Stochastic properties such as independence are also captured in the natural way. The proportion of server fluid that experiences two consecutive service completions by time $x$ is $P(S_1 + S_2 \leq x)$, where $S_1$ and $S_2$ are i.i.d. random variables, with the usual convolution distribution.

What we have said so far applies to any work-conserving server-assignment rule. Indeed, it corresponds to the easy underloaded special case of the fluid model in Whitt (2006a) and Liu and Whitt (2012a), where this part of the fluid model is carefully formalized. The policy $D_1$ plays a role when we keep track of the ages of each atom of server fluid, i.e., the time since its last break ended, and need to determine *which* server fluid is sent to the service facility to provide the required service. Consistent with (2.2), the rule $D_1$ first assigns fluid from the idle server fluid with elapsed current idle time at least $\theta$, giving priority to the larger elapsed current idle times. At that instant, when the atom of server fluid is assigned, the age is reset to 0. If more server assignment is needed, then $D_1$ assigns the fluid with elapsed current idle time less than $\theta$, giving priority to the smaller ages. Aside from the instants at which breaks end, as time advances the age of all server fluid increases at unit rate.

The deterministic fluid process that describes the evolution of the fluid model can be the natural analog of the stochastic process $(B_n, I_n)$ in (2.4). For the fluid model, $B(t, x, y)$ is the amount of busy server fluid at time $t$ with age at most $x$ and elapsed current service time at most $y$, while $I(t, x, y)$

is the amount of idle server fluid at time $t$ with age at most $x$ and elapsed idle time (since their last service completion) at most $y$. (As before, $x \geq y$ for $I(t,x,y)$.) We are interested in the steady-state of this fluid process $(B, I)$, which we denote by omitting the $t$.

Just as in Remark 2.1, part of the steady-state is already known. By Theorem 3.1 (a) of Whitt (2006a), for the busy fluid we can write

$$B(\infty, y) \equiv \int_0^y b_c(\infty, u)\, du \quad \text{for} \quad b_c(\infty, y) = \rho F^c(y), \quad y \geq 0, \tag{3.1}$$

and for the idle fluid we can write $I(\infty, \infty) = 1 - \rho$.

It remains to determine the full steady-state of the fluid process, given by $\{(B(x,y), I(x,y)) : x \geq 0,\ y \geq 0\}$. Just as Jackson (1957) originally found the steady-state distribution of a Jackson queueing network, we will obtain the steady-state of the $D_1$ fluid model by direct construction, i.e., by guessing the answer and verifying that it works. For $D_1$, our idea is that, in steady state, there ought to be a critical threshold $\tau$ such that, at all times, all fluid completing service after time $\tau$ is given a break, and so is assigned to service exactly $\theta$ time units later, whereas all remaining fluid is reassigned to service immediately. It is evident that this policy is consistent with $D_1$. The key is to show that there exists a unique threshold $\tau^*$ such that the above policy is consistent with the fluid model. (We elaborate on the critical threshold at the end of the online supplement.)

**Remark 3.2** (*conservation laws in the fluid model*) The conservation laws in §2.4 have natural analogs for the associated deterministic fluid model considered here. They are identical, except we remove the $n$ in (2.6).

**Remark 3.3** (*many-server heavy-traffic (MSHT) limits*) Important insight into the deterministic $D_1$ fluid model we have developed can be gained by seeing that it should serve as the limit in a many-server heavy-traffic (MSHT) functional weak law of large numbers (FWLLN) for an appropriately-scaled sequence of the $M/GI/n$ models we introduced in §2. We let the models be indexed by $n$, where in model $n$ the number of servers is $n$ and the arrival rate is $\lambda_n = \rho n$ for $0 < \rho < 1$, while the service-time distribution is held fixed. (For these asymptotic results, we can extend the arrival process from $M$ to $G$; we only require that the arrival process satisfy a FWLLN.) Since we have $\rho < 1$, the MSHT limit is in the underloaded quality-driven (QD) MSHT regime. The QD regime is required for the idleness of each server to be non-negligible in the limit, as required for non-negligible breaks.

In fact, we do not prove the full FWLLN here, but seeing it can help understanding. The conjectured MSHT FWLLN states that

$$(\bar{Q}_n, \bar{B}_n, \bar{I}_n)_t \Rightarrow (Q, B, I)_t \quad \text{as} \quad n \to \infty \tag{3.2}$$

using the topology of uniform convergence for $t$ over bounded intervals, where the limit $(Q, B, I)_t$ is the fluid process and we average for each $n$; i.e.,

$$(\bar{Q}_n, \bar{B}_n, \bar{I}_n)_t \equiv n^{-1}(Q_n, B_n, I_n)_t \quad \text{for all} \quad t \quad \text{and} \quad n, \tag{3.3}$$

with $(Q_n, B_n, I_n)_t$ defined in (2.5) above for model $n$. It is significant that this MSHT FWLLN has been established for the special case in Remark 2.1. That special case can be regarded as a consequence of Theorem 3.1 of Pang and Whitt (2010) or Liu and Whitt (2012b).

### 3.2 Relevant Renewal Theory

For non-exponential service-time distributions, the critical threshold $\tau^*$ for the $D_1$ fluid model depends on the renewal function associated with the service times. Indeed, it is natural that renewal theory should play a role, because we are considering immediately reassigning fluid upon service completion. Renewal theory naturally arises when we consider the number of times that an atom of server fluid is assigned before the age reaches $\tau^*$ and the server is assigned a break. Thus we need to review some properties of renewal processes.

Let $N \equiv \{N(t) : t \geq 0\}$ be the renewal counting process associated with successive i.i.d. service times $S_k$, i.e.,

$$N(t) \equiv \max\{k \geq 0 : S_0 + S_1 + \cdots + S_k \leq t\}, \quad t \geq 0, \tag{3.4}$$

where $S_0 \equiv 0$. We will exploit the mean of the renewal process, called the *renewal function*,

$$m(t) \equiv E[N(t)], \quad t \geq 0, \tag{3.5}$$

and the associated *renewal excess* (after time $t$),

$$R(t) \equiv S_{N(t)+1} - t, \quad t \geq 0. \tag{3.6}$$

As in §3.3 of Ross (1996), we apply Wald's equation to express the expected value as

$$E[R(t)] = E[S](E[N(t)] + 1) - t = E[N(t)] + 1 - t \quad \text{for all} \quad t \geq 0. \tag{3.7}$$

### 3.3 The Steady-State of the $D_1$ Fluid Model

Recall that we consider the $G/GI$ fluid model with: (i) service capacity 1, (ii) arrival rate $\rho < 1$, (iii) service-time proportions with cdf $F(x) \equiv P(S \leq x)$ having pdf $f$ with mean 1 and finite scv $c_s^2$, (iv) the $D_1$ server-assignment rule with target work breaks of length $\theta$, where $E[V] \equiv (1-\rho)/\rho < \theta$ and (v) in steady-state. As a regularity condition, we assume that $m(t)$ is continuous and strictly increasing with $m(0) = 0$, so that $m(t)$ has a unique inverse; it suffices for the service-time pdf $f$ to be continuous and positive in a neighborhood of the origin (but not necessarily $f(0) > 0$); see §XI.3 of Feller (1971). Let $\stackrel{\mathrm{d}}{=}$ denote equality in distribution.

**Theorem 3.1** (*the steady-state of the MSHT $G/GI$ fluid model with rule $D_1(\theta)$*) *Under the conditions above,* (*a*) *there exists a unique time $\tau^* \equiv \tau^*(\rho, \theta, F)$, $0 < \tau^* < \infty$, such that all fluid completing service with age at least $\tau^*$ is given a break of length $\theta$, and thus is assigned exactly $\theta$ time units later, while all fluid completing service with age less than $\tau^*$ is reassigned instantaneously and so experiences* 0 *idle time. The critical time $\tau^*$ is the unique root of the equation*

$$m(\tau^*) = \frac{1}{p} - 1 > 0, \tag{3.8}$$

*where $p \equiv (1 - \rho)/(\rho\theta) < 1$ and $m(t)$ is the renewal function associated with the service-time cdf $F$ in* (3.5)*. As a consequence, work breaks* (*idle times of length at least $\theta$*) *occur at the upper bound rate from Theorem 2.1,*

$$\beta^* = \frac{1 - \rho}{\theta} = p\rho, \tag{3.9}$$

*independent of the service cdf $F$ beyond its mean.*

(*b*) *The proportion of fluid that experiences time less than or equal to $x$ between breaks is $P(T^* \le x)$, where $T^* \equiv T(\tau^*)$ is a nondegenerate random variable with*

$$T^* \stackrel{\mathrm{d}}{=} \tau^* + R(\tau^*) + \theta = N(\tau^*) + 1 + \theta, \tag{3.10}$$

*where $N(t)$ is the renewal counting process associated with the cdf $F$ and $R(t)$ is the renewal excess, so that*

$$E[T^*] = m(\tau^*) + 1 + \theta = \frac{1}{\beta^*} \quad \textit{and} \quad Var(T^*) = Var(R(\tau^*)). \tag{3.11}$$

(*c*) *The steady-state densities of the server fluid content in service with age $x$, $b(x)$, and idle server fluid content with age $x$, $g(x)$, satisfy*

$$b(x) = \beta^* 1_{\{0 \le x < \tau^*\}} + \beta^* P(R(\tau^*) \ge x - \tau^*) 1_{\{\tau^* \le x < \infty\}} \tag{3.12}$$

*and*

$$\begin{aligned} g(x) = 0 \cdot 1_{\{0 \le x < \tau^*\}} &+ \beta^* P(R(\tau^*) \le x - \tau^*) 1_{\{\tau^* \le x < \tau^* + \theta\}} \\ &+ \beta^* (P(x - \tau^* - \theta \le R(\tau^*) \le x - \tau^*) 1_{\{\tau^* + \theta \le x < \infty\}} \end{aligned} \tag{3.13}$$

*for $\beta^*$ in* (3.9)*, $\tau^*$ the solution of equation* (3.8) *and $R(t)$ the renewal excess in* (3.6)*. As a consequence, the associated cumulative functions satisfy*

$$0 = I(\tau^*, \infty) < I(x, \infty) < I(\infty, \infty) \equiv I = 1 - \rho, \quad \tau^* < x < \infty, \tag{3.14}$$

*and*

$$B(\tau^*, \infty) = \beta^* \tau^* < B(x, \infty) < B(\infty, \infty) \equiv B = \rho, \quad \tau^* < x < \infty. \tag{3.15}$$

(*d*) *As a consequence, $D_1$ is lexicographically optimal for the fluid model, first minimizing $E[T]$ and then minimizing $Var(T)$ among all policies that yield the minimal $E[T]$.*

**Proof.** It is immediately evident that the claimed performance is consistent with the $D_1$ rule, because all idle server fluid content that has been idle for exactly $\theta$ experiences a break and is then immediately assigned to service. On the other hand, all the rest of the fluid (the fluid with age less than $\tau^*$) is immediately reassigned upon service completion. Moreover, by Theorem 2.1 and Remark 3.2, the rate of breaks is the maximum possible. However, it remains to show that a unique policy of this form can be realized and what its performance consequences are.

The key to a short proof is converting the present model into the model in Whitt (2006a) and Liu and Whitt (2012a) by creating new "macro service-times," which combine the consecutive service times experienced between breaks. Given $\tau^*$, the new combined service-time is $\tilde{S} \equiv \tau^* + R(\tau^*)$ with cdf $\tilde{F}$ and pdf $\tilde{f}$. Thus, in the underloaded $D_1$ fluid model, each atom of fluid experiences alternating breaks of length $\theta$, which we think of as interarrival times, and service times with cdf $\tilde{F}$. The steady-state performance of this $D_1$ model coincides with the previous $G/GI$ fluid model, as in Whitt (2006a) and Liu and Whitt (2012a), if we consider the service-time cdf $\tilde{F}$ and a fluid arrival process with rate $\beta^* E[\tilde{S}]$. The lower arrival rate resulting from the higher mean of cdf $\tilde{F}$ is balanced by the longer service time; i.e.,

$$b(x) = (\beta^* E[\tilde{S}])\tilde{f}_e(x) = (\beta^* E[\tilde{S}])(\tilde{F}^c(x)/E[\tilde{S}]) = \beta^* \tilde{F}^c(x), \tag{3.16}$$

which coincides with (3.12). The density $b$ in (3.16) then coincides with (3.2) in Theorem 3.1 (a) of Whitt (2006a). The density $g$ in (3.13) follows from observing that all idle fluid remains exactly for time $\theta$ after it arrived.

It remains to show that there exists a unique pair $(\tau^*, \beta^*)$ satisfying (3.8) and (3.9). To start, the renewal function has a unique inverse, because we have made assumptions that ensure it is continuous and strictly increasing. Thus, (3.8) necessarily has a unique solution.

On the other hand, given the form of the busy-server density $b(x)$ in (3.12), and the total busy server content $B = \rho$, we have $\rho = \beta^*\tau^* + \beta^* E[R(\tau^*)] = \beta^*(m(\tau^*) + 1)$, where $\beta^*$ is the rate breaks occur. Hence,

$$\beta^* = \rho/(m(\tau^*) + 1). \tag{3.17}$$

Given the $D_1$ policy, For $T^*$ in (3.10), we also have $T^* \stackrel{\mathrm{d}}{=} \tau^* + R(\tau^*) + \theta$, where $R(\tau^*)$ is the residual service time beyond $\tau^*$, so that

$$\beta^* = \frac{1}{E[T^*]} = \frac{1}{m(\tau^*) + 1 + \theta}. \tag{3.18}$$

Combining (3.17) and (3.18), we obtain the unique solution with $\tau^*$ in (3.8) and $\beta^*$ in (3.9). We remark that, as an alternative argument, we could also apply (2.7) and Remark 3.2: On average, each server experiences, $m(\tau^*)$ idle times of length 0 followed by one of length $\theta$. Hence,

$$E[V] = \frac{\theta}{m(\tau^*) + 1} = \frac{1-\rho}{\rho}, \tag{3.19}$$

from which we also obtain (3.8). Because there is a unique solution to equation (3.8), there is a unique fluid performance associated with $D_1$.

Finally, it remains to establish the lexicographical optimality. The analysis above shows that minimizing the mean $E[T]$ requires the two-point idle-time distribution, which is tantamount to immediately assigning all fluid with age less than $\tau^*$ the instant it completes service. At first glance, it might appear that $D_1$ is the only server-assignment rule minimizing $E[T]$ (and maximizing the rate of breaks) for the fluid model, but that is not the case. We can obtain alternative rules with the same $E[T]$, but higher variance $Var(T)$, by changing which fluid is immediately reassigned after completing service. The only remaining freedom if we fix the mean $E[T]$ at the optimal value is *which* fluid we assign immediately upon completing service. The only alternatives involve randomizing over the age while holding the mean $E[T]$ fixed, but that additional randomization necessarily increases the variance, by virtue of convex stochastic order, as in §9.5 of Ross (1996). An example is the SISF rule discussed in the next section. ■

**Remark 3.4** (*exponential service*) The solution in Theorem 3.1 simplifies if the service time $S$ is a mean-1 exponential random variable $M(1)$, because then $m(\tau^*) = \tau^*$ and $R(x^*) \stackrel{\mathrm{d}}{=} M(1)$, so that $\tau^* = (1/p) - 1$ and $T^* \stackrel{\mathrm{d}}{=} \tau^* + \theta + M(1)$. Then $b(x) = p\rho(1_{\{0 \le x < \tau^*\}} + e^{-(x-\tau^*)}1_{\{\tau^* \le x\}})$ and

$$g(x) = p\rho((1 - e^{-(x-\tau^*)})1_{\{\tau^* \le x < \tau^* + \theta\}} + (e^{-(x-\tau^*-\theta)} - e^{-(x-\tau^*)})1_{\{x \ge \tau^* + \theta\}}).$$

**Remark 3.5** (*approximating or calculating the renewal function and the mean excess*) Because the service distribution has a density (and thus is nonlattice) with $\sigma^2 < \infty$, see Proposition 3.4.8 of Ross (1996),

$$R(t) \Rightarrow S_e \quad \text{as} \quad t \to \infty \tag{3.20}$$

and

$$E[R(t)] \to E[S_e] = \frac{E[S^2]}{2E[S]} = \frac{E[S](c_s^2+1)}{2} \quad \text{as} \quad t \to \infty, \tag{3.21}$$

where $S_e$ is a random variable with the equilibrium-excess cdf $F_e$ associated with the service time cdf $F(t) \equiv (S \le t)$, i.e.,

$$F_e(t) \equiv P(S_e \le t) \equiv \frac{1}{E[S]} \int_0^t P(S > u)\, du, \quad t \ge 0. \tag{3.22}$$

By equation (2) of Eick et al. (1993),

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}, \tag{3.23}$$

so that for our case in which $E[S] = 1$, we have (3.21) and

$$Var(S_e) = E[S_e^2] - (E[S_e])^2 = \frac{E[S^3]}{3} - \left(\frac{E[S^2]}{2}\right)^2. \tag{3.24}$$

For applications, provided that $t$ is not too small, we thus might use the approximations

$$R(t) \approx S_e \quad \text{and} \quad E[R(t)] \approx E[S_e]. \tag{3.25}$$

For special distributions, $S_e$ can serve as an upper bound for $R(t)$. In particular, if $F$ has the increasing mean residual life (IMRL) or decreasing failure rate (DFR) property, then the distribution of $R(t)$ is increasing in $t$ in the sense of increasing convex order or stochastic order, respectively; see Brown (1980, 1981). The $H_2$ example we consider in §4.3 has the DFR property.

Alternatively, we can explicit numerical results by computing $m(t) \equiv E[N(t)]$ and $E[R(t)]$ numerically, e.g., by numerical transform inversion, as discussed in §13 of Abate and Whitt (1992).

**Remark 3.6** (*the experience of individual servers*) Individual servers (atoms of fluid) experience alternating busy periods distributed as $T_B \overset{\mathrm{d}}{=} \tau^* + R(\tau^*)$ and idle periods of length $T_I \equiv \theta$, which form an alternating renewal process with i.i.d. busy cycles distributed as $T^* = T_B + T_I$, as in §3.4.1 of Ross (1996). The form of the age densities in (3.12) and (3.13) can be explained by this alternating renewal process structure; e.g., by Theorem 4.8.4 of Ross (1996), $b(x) = P(T_B > x)/E[T^*] = \beta^* P(\tau^* + R(\tau^*) > x)$.

With simulation data, it is natural to observe the steady-state age of busy and idle fluid. Thus, we naturally observe densities of random variables $A_B$ and $A_I$ having the conditional age distribution for fluid in service (or idle) in steady state, conditional on it being busy (or idle). Clearly, $A_B$ and $A_I$ have densities $b(x)/\rho$ and $g(x)/(1-\rho)$, respectively. What we see at an arbitrary time in steady state can be understood from the renewal structure.

### 3.4 Other Rules Maximizing the Rate of Breaks: SISF

We now expand upon part (d) of Theorem 3.1 by illustrating an alternative server-assignment rule with the optimal mean $E[T]$, but higher variance $Var(T)$. The alternative rule is the shortest-idle-server-first (SISF) rule, which assigns the fluid with current idle time greater than or equal to $\theta$ first, just like $D_1$, but then assigns the fluid with the least (shortest) *current* idle time first. In fact, it is more evident that the SISF rule should produce the extremal two-point steady-state idle-time distribution, because it focuses directly on the current idle time.

The steady-state idle fluid content in the SISF fluid model can be represented by $I(y) = \int_0^y g(u)\,du$, $t \geq 0$, which represents the idle server content that has been idle for time $y$. The SISF rule dictates that we first assign fluid with idle time $\theta$ (or above, if present) and then assign idle fluid with age 0 (or above, if necessary). If SISF can achieve routing from the two end points only, then the density $g$ will be uniform over the interval $[0,\theta]$.

To see what is possible, we start with the fluid flow rates. Let $\lambda$, $\delta$ and $\alpha$ be the steady-state arrival rate of customer fluid, the departure rate of customer fluid (also the arrival rate of newly idle server fluid), and the assignment rate of idle server content. These have the obvious steady-state values $\lambda = \delta = \alpha = \rho$. Let $\alpha_0$ and $\alpha_\theta$ be the rate of assignment of fluid that has been idle for time 0 and $\theta$, respectively. If feasible, then we have $\alpha = \alpha_0 + \alpha_\theta$. By Theorem 2.1, the maximum possible value of breaks is $\alpha_\theta = \beta^* = (1-p)/\theta = p\rho$, where $p$ can be interpreted as the proportion of idle fluid on break, leaving $\alpha_0 = (1-p)\rho$ for immediate reassignment. Thus, SISF does assign fluid from the two end points only. SISF first assigns all fluid that has been idle for time $\theta$ and then immediately re-assigns a proportion $1-p$ of the newly idle server content. That makes $g(y) = (1-\rho)/\theta$, $0 < y < \theta$, and $\alpha_\theta = g(\theta-)$ (the left limit at $\theta$), where $g(\theta-) = (1-\rho)/\theta = [(1-\rho)/(\rho\theta)]\rho = p\rho$. That routing occurs at each successive service completion time. Thus, the proportion of time between successive breaks with SISF can be represented by the random sum

$$T \approx \theta + \sum_{i=1}^{N(p)} S_i, \tag{3.26}$$

where $N(p)$ is a random variable with the geometric distribution on the positive integers having mean $1/p$ for $p \equiv E[V]/\theta = (1-\rho)/(\rho\theta)$ and $S_i$ are i.i.d. mean-1 service-time random variables with cdf $F$

and variance $\sigma^2$ that are independent of $N(p)$, so that

$$E[T] = \theta + \frac{1}{p} = \theta + \frac{\theta}{E[V]} = \frac{\theta}{1-\rho} = \frac{1}{\beta^*}, \tag{3.27}$$

as it should, and

$$Var(T) = Var(S)E[N(p)] + E[S]^2 Var(N(p)) = \frac{\sigma^2}{p} + \frac{1-p}{p^2} = \frac{p\sigma^2 + 1 - p}{p^2} = \left(\frac{\rho\theta}{(1-\rho)}\right)^2. \tag{3.28}$$

which equals $1/p^2 = ((\rho\theta/(1-\rho))^2$ when $\sigma^2 = 1$.

We can easily compare SISF to $D_1$ for $M$ service: For $D_1$, $Var(T) = Var(R(\tau^*)) = Var(M(1)) = 1$, which is less than $1/p^2$, typically much less. For the base case, $1/p = 15.0$, so that $Var(T) = 225$ for SISF. We will show that these fluid formulas are consistent with simulation for large $n$.

# 4 Simulation Experiments for Unannounced Breaks: $D_1$ and $SISF$

In §2.1 and §2.2 of the online supplement we indicate how we implement the $D_1$ and $SISF$ server-assignment rules in the simulation. In §4.1 we discuss how we execute the simulation and perform the statistical estimates. In §4.2 we report simulation results for the $M/M/n$ model in the base case. In §4.3 we report additional results for the $D_1$ rule with a hyperexponential ($H_2$) service-time distribution. (We present background for the $H_2$ distribution in §3.3 of the online supplement.)

## 4.1 Statistical Estimation

Our simulations used $r = 20 - 50$ i.i.d. replications of an $M/G/n$ system observed over a time interval of length between $2000 - 40,000$ depending on the value of $n$ after a warmup period of length $50 - 100$ to allow the system that started empty to approach steady state. (We remark that the appropriate choices depend on $n$, largely because the sample size is proportional to both $n$ and $t$; see Srikant and Whitt (1996),Whitt (1989) and Ni and Henderson (2015).) Idle times and periods between successive breaks are collected from all $n$ servers.

To estimate the probability of an event, we first compute the sampling frequency within each replication. Then the overall estimate is the sample average of the $r$ values, which should be approximately Gaussian distributed with unknown variance. Hence, the 95%-confidence interval (CI) is constructed using the Student-$t$ distribution with $t_{0.025}(r-1)$; e.g., see §8 of Walpole et al. (1993). For a random variable $X$, the first two moments $m_k \equiv E[X^k]$, $k = 1, 2$, are estimated by the sample averages $\bar{m}_1$ and $\bar{m}_2$ within each replication. Then the overall estimates $\bar{m}_1$ and $\bar{m}_2$ are taken to be the sample averages of the $r$ values, which again should be Gaussian; e.g., see p. 2 of Ni and Henderson (2015). Hence, again the 95% CI's can be constructed in the same way with $t_{0.025}(r-1)$.

Within each replication, the variance formula is $\sigma^2 = m_2 - m_1^2$. We therefore estimate the standard deviation (std) within each replication by $\bar{\sigma} = \sqrt{\bar{m}_2 - \bar{m}_1^2}$. We then obtain $r$ estimates of the std, one of each replication. We estimate the overall std as the sample average of these. The way to construct CI for the std is less straightforward, because $\bar{\sigma}$ is not normally distributed due to the fact that $m_1^2$ is no longer Gaussian. To circumvent this difficulty, we use sample quantiles to construct the CI.

### 4.2 Simulation Results

We now report simulation results for $D_1$ and $SISF$. (More results appear in the appendix.) We primarily focus on the base $M/M/n$ case with $\rho = 0.9$, $E[S] = 1$, $n = 100$ and $\theta = 5/3$. Table 1 provides simulation estimates of the probability of short and large idle times as a function of the scale $n$. We call idle times small is they are less than 0.1, an arbitrary number less than the mean 0.1111; we call idle times greater than or equal to $\theta$ large. Figure 1 in the online supplement shows that the idle-time distribution with $D_1$ tends to be like the two-point extremal distribution for the fluid model.

Table 1 shows that the performance of the two rules is very similar, but $SISF$ produces an idle-time distribution slightly closer to the desired two-point extremal distribution in Theorem 2.1. The fluid model provides the limiting case of $n = \infty$.

| system | $D_1$ | | $SISF$ | |
|---|---|---|---|---|
| size | $P(V_n \leq 0.1)$ | $P(V_n \geq \theta)$ | $P(V_n \leq 0.1)$ | $P(V_n \geq \theta)$ |
| $n = 25$ | $0.7917 \pm 0.0018$ | $0.0163 \pm 0.0003$ | $0.8257 \pm 0.0012$ | $0.0217 \pm 0.0003$ |
| $n = 100$ | $0.8240 \pm 0.0013$ | $0.0223 \pm 0.0004$ | $0.8341 \pm 0.0008$ | $0.0293 \pm 0.0004$ |
| $n = 250$ | $0.8498 \pm 0.0007$ | $0.0317 \pm 0.0003$ | $0.8698 \pm 0.0005$ | $0.0386 \pm 0.0003$ |
| $n = 1000$ | $0.8896 \pm 0.0008$ | $0.0492 \pm 0.0007$ | $0.9028 \pm 0.0005$ | $0.0546 \pm 0.0005$ |
| $n = 5000$ | $0.9155 \pm 0.0002$ | $0.0601 \pm 0.0010$ | $0.9236 \pm 0.0003$ | $0.0628 \pm 0.0002$ |
| $n = \infty$ | $0.9333 \pm 0.0000$ | $0.0633 \pm 0.0000$ | $0.9333 \pm 0.0000$ | $0.0667 \pm 0.0000$ |

Table 1: Simulation estimates of the probability of short and large idle times as a function of the scale $n$ for the server-assignment rules $D_1$ and SISF in the base $M/M/n$ case with $\rho = 0.9$, $E[S] = 1$ and $\theta = 5/3$. The fluid model provides the limiting case of $n = \infty$.

Table 2 shows simulation estimates of the mean and standard deviation of the interval between breaks, $T_n$, as a function of the scale $n$ for the server-assignment rules $D_1$ and SISF in the base $M/M/n$ case. As for the fluid model in §3.4, the means are very similar, but the standard deviation is much smaller for $D_1$. The fluid model is very helpful for understanding the advantage of $D_1$ over $SISF$, but the fluid model does not yield accurate approximations for the base case of $n = 100$.

| system | $D_1$ | | $SISF$ | |
|---|---|---|---|---|
| size | $E[T_n]$ | $SD(T_n)$ | $E[T_n]$ | $SD(T_n)$ |
| $n = 25$ | $66.29 \pm 1.12$ | $38.04 \pm 0.71$ | $51.44 \pm 0.49$ | $52.31 \pm 0.89$ |
| $n = 100$ | $48.06 \pm 0.79$ | $18.73 \pm 0.41$ | $37.85 \pm 0.49$ | $36.68 \pm 0.52$ |
| $n = 250$ | $33.45 \pm 0.33$ | $9.47 \pm 0.35$ | $28.62 \pm 0.21$ | $27.01 \pm 0.28$ |
| $n = 1000$ | $20.84 \pm 0.30$ | $3.06 \pm 0.12$ | $20.28 \pm 0.16$ | $18.54 \pm 0.16$ |
| $n = 5000$ | $16.75 \pm 0.07$ | $1.38 \pm 0.03$ | $17.28 \pm 0.05$ | $15.59 \pm 0.06$ |
| $n = \infty$ | $16.67 \pm 0.00$ | $1.00 \pm 0.00$ | $16.67 \pm 0.00$ | $15.00 \pm 0.00$ |

Table 2: Simulation estimates of the mean and standard deviation of the interval between breaks, $T_n$, as a function of the scale $n$ for the server-assignment rules $D_1$ and SISF in the base $M/M/n$ case with $\rho = 0.9$, $E[S] = 1$ and $\theta = 5/3$. The fluid model provides the limiting case of $n = \infty$.

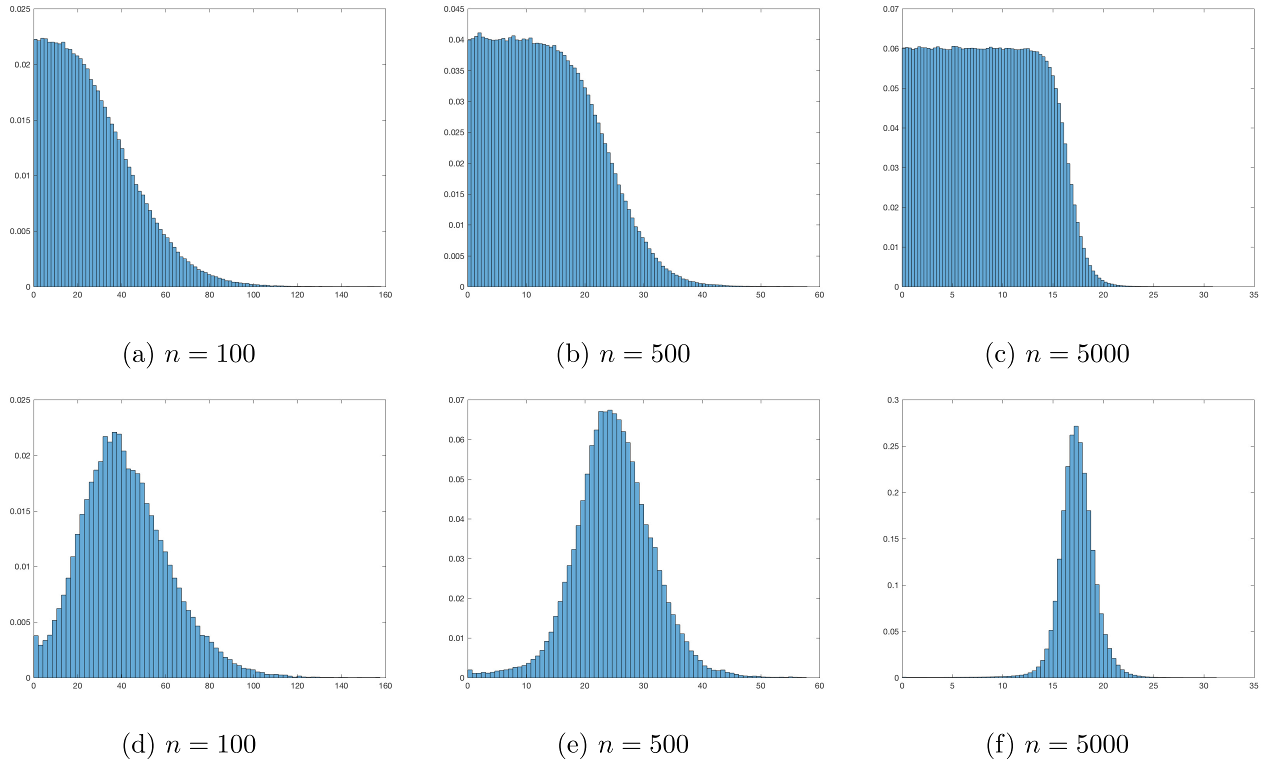
Let $A_B$ ($A_I$) be a random variable with the distribution of the age of a busy (idle) server at an arbitrary time in steady state, as discussed in Remark 3.6. Figure 2 shows histograms of these ages estimated from the simulation results. The vertical $y$ axis has been scaled so that the area under each histograms is 1, making the histogram an estimate of the density.

From the MSHT fluid model with rule $D_1$, we expect that the ages $A_B$ and $A_I$ have densities much like their fluid counterparts $b(x)/\rho$ and $g(x)/(1 - \rho)$ for $b(x)$ and $g(x)$ in (3.12) and (3.13). Table 3 reports estimations of the mean and standard deviation of these age random variables for $D_1$ as a function of $n$. As before, the case $n = \infty$ corresponds to the fluid model.

| | Busy | | Idle | |
|---|---|---|---|---|
| | $E[A_B]$ | $std(A_B)$ | $E[A_I]$ | $std(A_I)$ |
| $n = 100$ | $26.510 \pm 0.051$ | $19.146 \pm 0.072$ | $41.725 \pm 0.068$ | $19.725 \pm 0.083$ |
| $n = 500$ | $13.178 \pm 0.016$ | $8.395 \pm 0.033$ | $24.858 \pm 0.019$ | $6.565 \pm 0.024$ |
| $n = 1000$ | $10.518 \pm 0.011$ | $6.380 \pm 0.018$ | $20.865 \pm 0.013$ | $3.828 \pm 0.017$ |
| $n = 5000$ | $8.399 \pm 0.004$ | $4.935 \pm 0.011$ | $17.378 \pm 0.004$ | $1.797 \pm 0.007$ |
| $n = \infty$ | $7.533 \pm 0.000$ | $4.392 \pm 0.000$ | $15.833 \pm 0.000$ | $1.108 \pm 0.000$ |

Table 3: Simulation estimates of the mean and standard deviation of the ages $A_B$ and $A_I$ in the base case as a function of $n$.

It is also useful to look at the pattern of successive idle times over a long horizon. Figure 3 displays successive idle-times for a set of randomly selected servers in the $M/M/n$ base case. The vertical axis measures the length of an idle-time and the horizontal axis indexes the successive idle times.



(a) $n = 100$ (b) $n = 500$ (c) $n = 5000$

(d) $n = 100$ (e) $n = 500$ (f) $n = 5000$

Figure 2: Histograms of the ages $A_B$ of a busy server (top) and $A_I$ of an idle server (bottom) estimated from computer simulation for the in the base $M/M/n$ model with rule $D_1$ for three values of $n$: $n = 100$, $n = 500$ and $n = 5000$.

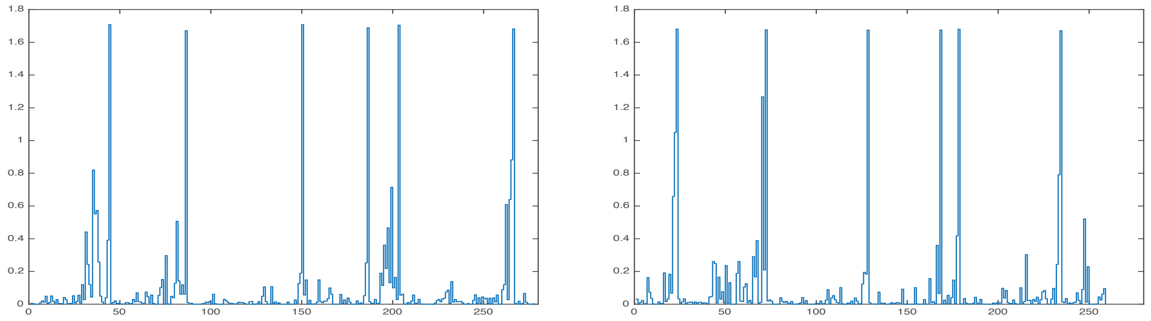
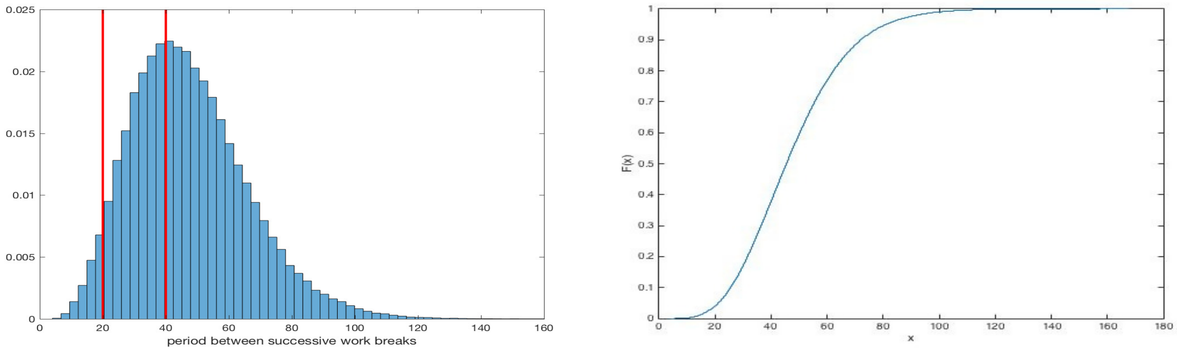


Figure 3: Two sample paths of successive idle times over a time interval of length 300 for $D_1$ in the base case.

Figure 3 shows that $D_1$ generates occasional long idle times with many very short ones in between. Over a long horizon, these work breaks occur fairly regularly.

From the results above, we conclude that, unlike LISF and RR, the $D_1$ server-assignment rule can achieve the desired work breaks. Nevertheless, there are three serious drawbacks in $D_1$. First, Figure

4 shows 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 48 for $\theta = 5/3$. The full distribution is shown in Figure 4, with a histogram on the left and the empirical cumulative distribution function (ecdf) on the right. Finally, we want to announce the work breaks so that the server can be off duty during the breeak, which is not possible with $D_1$.



(a) Histogram

(b) ECDF

Figure 4: The histogram (left) and ecdf (right) estimated from simulation of the distribution of $T$, the time between breaks, with rule $DP1$ for $\theta = 5/3$

## 4.3 The $D_1$ Rule with a Different Service-Time Distribution

We also examined $D_1$ with non-exponential service-time distributions. We illustrate by briefly discussing the case of a mean-1 hyperexponential ($H_2$) distribution with variance $\sigma^2 = 4$ and balanced means, as in §3.1 of Whitt (1982); additional discussion for this example appears in the appendix.

From (3.25) and Theorem 3.1, the key quantities for the fluid model are

$$E[R(\tau^*)] \approx E[S_e] = 2.50 \quad \text{and} \quad SD(R(\tau^*)) \approx SD(S_e) = 3.71 \tag{4.1}$$

At the end of §3.2, we noted that $S_e$ is an upper bound for $R(t)$ in stochastic order, because the $H_2$ cdf is DFR. The numerical values in (4.1) should be compared to the corresponding values for $M(1)$: $E[R(\tau^*)] = 1$ and $SD(R(\tau^*)) = 1$.

Table 4 shows simulation estimates of the mean and standard deviation of $A_B$, $A_I$ and $T$ as a function of $n$ in the $M/H_2/n$ model with rule $D_1$, $\rho = 0.9$ and $\theta = 5/3$.

Tables 2-4 provide important confirmation of the fluid model with non-exponential service-time distribution and the approximation $R(\tau^*) \approx S_e$ in (3.25), because the estimates for $n = 5000$ are

|  | $E[A_B]$ | $std(A_B)$ | $E[A_I]$ | $std(A_I)$ | $E[T_n]$ | $std(T_n)$ |
|---|---|---|---|---|---|---|
| $n = 100$ | $27.145 \pm 0.098$ | $22.059 \pm 0.102$ | $37.622 \pm 0.106$ | $23.851 \pm 0.115$ | $41.663 \pm 0.126$ | $23.7531 \pm 0.131$ |
| $n = 250$ | $18.277 \pm 0.085$ | $13.584 \pm 0.092$ | $29.473 \pm 0.089$ | $13.922 \pm 0.079$ | $31.748 \pm 0.095$ | $13.473 \pm 0.104$ |
| $n = 1000$ | $10.813 \pm 0.062$ | $7.249 \pm 0.071$ | $20.031 \pm 0.075$ | $5.883 \pm 0.058$ | $20.495 \pm 0.047$ | $5.568 \pm 0.072$ |
| $n = 5000$ | $8.765 \pm 0.022$ | $5.789 \pm 0.030$ | $17.017 \pm 0.028$ | $4.150 \pm 0.025$ | $16.725 \pm 0.024$ | $3.876 \pm 0.030$ |

Table 4: Simulation estimates of the mean and standard deviation of $A_B$, $A_I$ and $T$ as a function of $n$ in the $M/H_2/n$ model with rule $D_1$, $\rho = 0.9$ and $\theta = 5/3$.

close to the analytical values for $n = \infty$. In particular, consistent with the fluid model, Tables 2-4 indicate that the mean of $T^*$ is independent of the additional service-time variability, while the standard deviation increases in the variability. The estimated value for $SD(T)$ of 3.88 from simulation for $n = 5000$ is well approximated by $SD(S_e) = 3.71$ in (4.1). However, as before, the fluid model approximations for $n = 100$ are not accurate.

## 5 The $D_2(\theta, \tau, \eta)$ Rule for Announced Work Breaks

Theorem 3.1 for the fluid model suggests a natural way to modify $D_1$ to create a rule for announced breaks: introduce a threshold control parameter $\tau$, paralleling $\tau^*$. For each server, we keep track of the age and announce a break when the age exceeds $\tau$; the server is then off duty for time $\theta$. (For a busy server, the break begins upon service completion; for an idle server, the break begins immediately.) Any breaks that occur before time $\tau$ are unannounced breaks.

We first observe that Theorem 3.1 implies that $D_2$ is also optimal for the fluid model provided we choose the correct parameters.

**Corollary 5.1** (*equivalence for $D_2$ with appropriate parameters*) *Under the conditions of Theorem* 3.1*, for the $G/GI$ fluid model, the server assignment rule $D_2(\theta, \tau)$ coincides with the $D_1(\theta)$ rule if $\tau = \tau^*$ and $\eta \geq 1 - \rho$.*

Because the servers that are on break are off duty, there can be servers not serving a customer even though there are customers waiting in queue; i.e., now there is inevitably some level of performance degradation for customers. To control that performance degradation for customers, we further modify $D_2$ by imposing an upper bound $\eta$ on the number of servers that can be on break at any time. A server due a break when the number of servers on break is $\eta$ is given high priority for a break in the future.

Clearly, the additional parameters complicate the control. We propose introducing a cost function to measure the tradeoff between the cost to servers of not getting enough announced breaks and the cost to customers of performance degradation. We illustrate how such cost functions can be constructed by using a cost function that is a function two steady-state proportions: (i) the proportion of the idle time per server spent on an announced break, $p_A$, and the proportion of customers delayed, $p_D \equiv P(Q \geq n)$, measured relative the value $p_D^*$ with no degradation at all.

Specifically, the proposed cost function is

$$C \equiv C(\tau, \eta) = w(1 - p_A) + (1 - w)(p_D - p_D^*), \tag{5.1}$$

where the performance measures $p_A$ and $p_D$ are functions of the control parameters, while the weight $w$ with $0 \leq w \leq 1$ represent our relative concern about the two factors. We have used simulation to study the performance of the $D_2(\theta, \tau, \eta)$ rule as a function of the parameters, including choosing the optimal $\tau$ and $\eta$ to minimize the cost function in (5.1).

## 5.1 Simulation Results for the Base Case

We start by showing in Tables 5 and 6 how the two performance measures $p_A$ and $p_D$ depend on the control parameters $\tau$ and $\eta$ for the base $M/M/n$ model with $n = 100$ and $\rho = 0.9$. (For this base case, the delay probability without extra degradation is $p_D^* = 0.223$.)

|  | $\eta = 4$ | $\eta = 6$ | $\eta = 8$ | $\eta = 10$ |
|---|---|---|---|---|
| $\tau$ | $p_A$ | $p_A$ | $p_A$ | $p_A$ |
| $\tau = 15$ | $0.3714 \pm 9 \times 10^{-4}$ | $0.5130 \pm 7 \times 10^{-4}$ | $0.5971 \pm 6 \times 10^{-4}$ | $\mathbf{0.6301 \pm 8 \times 10^{-4}}$ |
| $\tau = 20$ | $0.3706 \pm 9 \times 10^{-4}$ | $0.5090 \pm 8 \times 10^{-4}$ | $0.5734 \pm 8 \times 10^{-4}$ | $\mathbf{0.5774 \pm 7 \times 10^{-4}}$ |
| $\tau = 25$ | $0.3694 \pm 9 \times 10^{-4}$ | $0.4939 \pm 8 \times 10^{-4}$ | $\mathbf{0.5189 \pm 9 \times 10^{-4}}$ | $0.5002 \pm 9 \times 10^{-4}$ |
| $\tau = 30$ | $0.3661 \pm 9 \times 10^{-4}$ | $0.4588 \pm 9 \times 10^{-4}$ | $\mathbf{0.4587 \pm 9 \times 10^{-4}}$ | $0.4489 \pm 9 \times 10^{-4}$ |
| $\tau = 35$ | $0.3588 \pm 9 \times 10^{-4}$ | $\mathbf{0.4109 \pm 9 \times 10^{-4}}$ | $0.4041 \pm 9 \times 10^{-4}$ | $0.3970 \pm 9 \times 10^{-4}$ |
| $\tau = 40$ | $0.3472 \pm 9 \times 10^{-4}$ | $\mathbf{0.3672 \pm 9 \times 10^{-4}}$ | $0.3604 \pm 9 \times 10^{-4}$ | $0.3552 \pm 7 \times 10^{-4}$ |

Table 5: 95% confidence intervals for the proportion of idle time spent on announced work breaks, $p_A$, for rule $D_2(\theta, \tau, \eta)$ as a function of and $\tau$ and $\eta$ for $n = 100$ and $\theta = 5/3$. The entries in bold are maximal over $\eta$ for that $\tau$.

In addition to the announced breaks, there also are unannounced breaks. Paralleling Table 5, Table 7 shows the proportion of idle time spent on idle periods of size at least $\theta$, denoted by $p_B$, with rule $D_2(\theta, \tau, \eta)$. The proportions are larger in Table 7, because both unannounced and announced breaks are included.

These tables show that $\eta$ makes much greater difference than $\tau$. Moreover, there is a strong tradeoff in the choice of $\eta$. All three of $p_D$, $p_A$ and $p_B$ are monotone in $\tau$, but $p_A$ and $p_B$ are not monotone

|  | $\eta = 4$ | $\eta = 6$ | $\eta = 8$ | $\eta = 10$ |
|---|---|---|---|---|
| $\tau$ | $p_D$ | $p_D$ | $p_D$ | $p_D$ |
| $\tau = 15$ | $0.3368 \pm 0.0018$ | $0.4141 \pm 0.0026$ | $0.4860 \pm 0.0020$ | $0.5414 \pm 0.0023$ |
| $\tau = 20$ | $0.3330 \pm 0.0021$ | $0.4076 \pm 0.0021$ | $0.4603 \pm 0.0023$ | $0.4855 \pm 0.0021$ |
| $\tau = 25$ | $0.3319 \pm 0.0022$ | $0.3937 \pm 0.0017$ | $0.4218 \pm 0.0020$ | $0.4339 \pm 0.0025$ |
| $\tau = 30$ | $0.3291 \pm 0.0018$ | $0.3739 \pm 0.0025$ | $0.3887 \pm 0.0025$ | $0.3974 \pm 0.0024$ |
| $\tau = 35$ | $0.3246 \pm 0.0021$ | $0.3510 \pm 0.0024$ | $0.3598 \pm 0.0022$ | $0.3663 \pm 0.0024$ |
| $\tau = 40$ | $0.3206 \pm 0.0020$ | $0.3342 \pm 0.0027$ | $0.3413 \pm 0.0020$ | $0.3449 \pm 0.0028$ |

Table 6: 95% confidence intervals for the steady-state delay probability $p_D$ associated with $D_2(\theta, \tau, \eta)$ as a function of and $\tau$ and $\eta$ for $n = 100$ and $\theta = 5/3$.

|  | $\eta = 4$ | $\eta = 6$ | $\eta = 8$ | $\eta = 10$ |
|---|---|---|---|---|
| $\tau$ | $p_B$ | $p_B$ | $p_B$ | $p_B$ |
| $\tau = 15$ | $0.5041 \pm 6 \times 10^{-4}$ | $0.5731 \pm 5 \times 10^{-4}$ | $0.6212 \pm 6 \times 10^{-4}$ | **$0.6407 \pm 8 \times 10^{-4}$** |
| $\tau = 20$ | $0.5043 \pm 7 \times 10^{-4}$ | $0.5684 \pm 6 \times 10^{-4}$ | $0.6022 \pm 9 \times 10^{-4}$ | **$0.6032 \pm 6 \times 10^{-4}$** |
| $\tau = 25$ | $0.5021 \pm 7 \times 10^{-4}$ | $0.5587 \pm 6 \times 10^{-4}$ | **$0.5671 \pm 7 \times 10^{-4}$** | $0.5616 \pm 9 \times 10^{-4}$ |
| $\tau = 30$ | $0.4991 \pm 9 \times 10^{-4}$ | **$0.5349 \pm 9 \times 10^{-4}$** | $0.5333 \pm 7 \times 10^{-4}$ | $0.5278 \pm 6 \times 10^{-4}$ |
| $\tau = 35$ | $0.4944 \pm 7 \times 10^{-4}$ | **$0.5091 \pm 8 \times 10^{-4}$** | $0.5045 \pm 9 \times 10^{-4}$ | $0.5009 \pm 7 \times 10^{-4}$ |
| $\tau = 40$ | $0.4832 \pm 8 \times 10^{-4}$ | **$0.4872 \pm 5 \times 10^{-4}$** | $0.4829 \pm 7 \times 10^{-4}$ | $0.4797 \pm 7 \times 10^{-4}$ |

Table 7: 95% confidence intervals for the proportion of idle time spent on idle periods of size at least $\theta$, $p_B$, with rule $D_2(\theta, \tau, \eta)$ as a function of $\tau$ and $\eta$ for $n = 100$ and $\theta = 5/3$. The entries in bold are maximal over $\eta$ for that $\tau$.

in $\eta$ for fixed $\tau$. The entries in bold show that optimal $\eta$ for each $\tau$. The values of $\eta$ where these maximal proportions occur are decreasing in $\tau$. The corresponding plots for other weights $w$ are shown in the appendix. Figure 5 shows the cost in (5.1) as a function of $\tau$ and $\eta$ for the base case with weight $w = 0.5$. Overall, we see that the cost is minimized by choosing $\eta = 8$ with $\tau = 15$ or $\tau = 20$. For higher $\tau$, the optimal choice shifts to $\eta = 6$.

**Remark 5.1** (*a larger system*) The appendix shows corresponding results for a large $M/M/n$ system with $n = 1000$, but still $\rho = 0.9$ and $\theta = 5/3$.

**Remark 5.2** (*an alternative more elementary server-assignment rule*) We identified an alternative rule that is easier to implement and has comparable performance. This alternative rule still lets servers go on break when their age exceeds the threshold $\tau$, but otherwise uses the standard LISF rule for

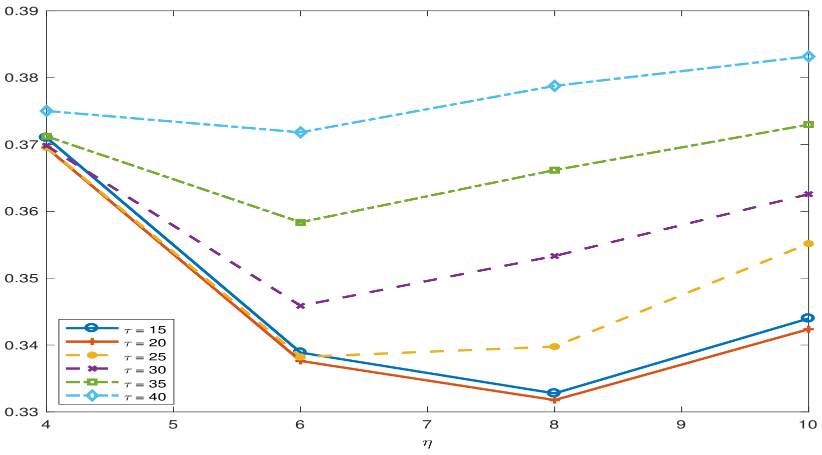


Figure 5: The cost in (5.1) as a function of $\tau$ and $\eta$ for $D_2(\theta, \tau, \eta)$ in the base case with $n = 100$, $\theta = 5/3$ and $w = 0.5$

server assignment. Tables and plots for this alternative LISF-based alternative to $D_2(\theta, \tau, \eta)$ are shown in the appendix.

**Remark 5.3** (*comparison to the $M/M/(n-b)$ model with a fixed number $b$ on break*) It is interesting to compare the server-assignment rule $D_2$ to what happens with a fixed number of servers on break. The appendix shows that the $D_2$ outperforms the alternative with a fixed number $b$ of servers on break, where a range of $b$ is considered ranging from the greatest integer less than or equal to the average number on break to the bound $\eta$.

## 6 Conclusions

In this paper we developed new rules for assigning idle servers to customers requesting service in a contact center in order to create effective work breaks from available idleness. After showing that the standard longest-idle-server-first (LISF) rule and the random routing (RR) alternative generate breaks too infrequently in §2.5, we studied the one-parameter rule $D_1 \equiv D_1(\theta)$ yielding unannounced breaks while maintaining work conservation in §3 and §4, and then studied the three-parameter refined rule $D_2 \equiv D_2(\theta, \tau, \eta)$ yielding announced breaks by sacrificing work-conservation in §5.

We provided strong theoretical support for these proposed server-assignment rules in §3 by analyzing them in the many-server heavy-traffic (MSHT) fluid model for the $G/GI/n$ model, which arises as the MSHT limit as the number of servers $n$ and the arrival rate increase toward infinity, while the traffic intensity (workload per server) is held fixed at $\rho < 1$ (the quality-driven MSHT regime). Theorem

3.1 and Corollary 5.1 show that both rules are optimal for this fluid model, minimizing $E[T]$, the steady-state mean interval between breaks, yielding the upper bound on the rate of breaks, established in Theorem 2.1. However, in §3.4 we show that there are multiple rules that achieve this optimal mean. Among all rules that achieve this minimum mean $E[T]$, the rules $D_1$ and $D_2$ minimize the standard deviation $SD(T)$.

Since announced breaks are likely to be preferred, there is interest in the rule $D_2(\theta, \tau, \eta)$, but it is complicated because it causes performance degradation for customers and has more parameters. In §5 we show the the parameters $\tau$ and $\eta$ can be chosen by formulating an optimization that expresses the tradeoff between the interests of servers and customers.

Finally, we conducted extensive simulation experiments evaluating the new server-assignment rules $D_1$ and $D_2$. First, the simulation experiments reported in §4 confirm the fluid limit and show that the rule $D_1$ is effective for generating unannounced breaks in an $M/M/n$ base case with $n = 100$ servers and $\rho = 0.9$. Second, the simulation results in §5 show that simulation can be used to solve the optimization problems yielding the control paramters.

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, it remains to investigate other server-assignment rules. Third, it remains to establish the MSHT FWLLN showing that the sequence of stochastic models converges to the MSHT fluid model as the scale $n$ increases; the authors hope to report results for that in the near future. Finally, there remain many other analytical challenges, such as deriving explicit formulas and establishing optimality for the stochastic models.

## Acknowledgment

Research support was received from NSF (CMMI 1634133).

## References

Abate J, Whitt W (1992) The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10:5–88.

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* 63(2):163–191.

Brown M (1980) Bounds, inequalities and monotonicity properties for some specialized renewal processes. *Annals of Probability* 8(2):227–240.

Brown M (1981) Further monotonicity properties for specialized renewal processes. *Annals of Probability* 9(5):891–895.

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.

Eick SG, Massey WA, Whitt W (1993) The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41:731–742.

Feller W (1971) *An Introduction to Probability Theory and its Applications* (New York: John Wiley), second edition edition.

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

Jackson JR (1957) Networks of waiting lines. *management Science* 5(4):518–521.

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.

Kaspi H, Ramanan K (2011) Law of large numbers limits for many-server queues. *Ann. Applied Probab.* 21:33–114.

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

Liu Y, Whitt W (2011) Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. *Queueing Systems* 67:145–182.

Liu Y, Whitt W (2012a) The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71:405–444.

Liu Y, Whitt W (2012b) A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Oper. Res. Letters* 40:307–312.

Liu Y, Whitt W (2014) Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability* 24(1):378–421.

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).

Ni EC, Henderson SG (2015) How hard are steady-state queueing simulations? *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 25(4):27.

Pang G, Whitt W (2010) Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* 65:325–364.

Puterman ML (2005) *Markov Decison Processes: Discrete Stochastic Dynamic Programming* (Hoboken, NJ: Wiley), second edition.

Ross SM (1996) *Stochastic Processes* (New York: Wiley), second edition.

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.

Srikant R, Whitt W (1996) Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 6(1):7–52.

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).

Walpole RE, Myers RH, Myers SL, Ye K (1993) *Probability and statistics for engineers and scientists*, volume 5 (Macmillan New York).

Whitt W (1982) Approximating a point process by a renewal process, I: two basic methods. *Oper. Res.* 30:125–147.

Whitt W (1989) Planning queueing simulations. *Management Science* 35(11):1341–1366.

Whitt W (1991) A review of $L = \lambda W$. *Queueing Systems* 9:235–268.

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

Whitt W (2006a) Fluid models for multiserver queues with abandonments. *Operations Research* 54(1):37–54.

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