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

April 8, 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 arise 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), 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

Here we apply queueing models to 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 non-idling (work-conserving); i.e., no customer waits in

queue if there is an available server, so that the 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 $G/GI/n$ queueing model with n homogeneous servers working in parallel. We assume that there is a target break duration θ. 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 calls having a mean duration of 3 minutes, so an hour is a time interval of length 20 with $E[S] = 1$. Very 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. The second goal leads to multiple-criteria decision making. 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 variance $SD(T)$. Alternatively, we could look at weighted averages $wE[T] + (1 - w)SD(T)$ for $0 < w < 1$.

1.2 Our Main Contributions

(i) The standard longest-idle-server-first (LISF) server-assignment rule and natural alternatives such as the random routing (RR) rule generate unannounced breaks, because we call all idle times exceeding θ breaks. However, we show that these rules generate breaks too infrequently.

(ii) Hence, 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 θ), 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 θ), 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 many-server heavy-traffic (MSHT) limits 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, all are optimal for the fluid model, minimizing $E[T]$ (in fact lexicographically optimal, first minimizing $E[T]$ and then minimizing $SD(T)$). Explicit formulas for the steady-state performance show that (i) the distribution of T is insensitive to the arrival process beyond its rate and (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 show that the lexicographical optimality criterion can play an important role by identifying another rule that also minimizes $E[T]$ for the MSHT fluid model, but produces much larger $SD(T)$. That rule is the natural myopic alternative to D_1 in which we first assign idle servers who have completed a break and then use the shortest-idle-server-first (SISF) rule, looking at the current level of the elapse idle time instead of the age.

(v) We also consider announced work breaks, for which we necessarily lose the non-idling property. (With announced breaks, 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 τ. (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 η on the number of servers that can be on break at any one time. If a server cannot be given a break, it is given high priority for a future break.

(vi) We propose a way to evaluate the tradeoff between the frequency of announced breaks and the resulting performance degradation for the customers being served. As a specific criterion, we propose minimizing a cost function that is a weighted sum of the proportion of customers experiencing a delay before starting service and the proportion of server idle time not devoted to announced breaks.

(vii) Finally, we report results of extensive simulation experiments. These simulation experiments show for the base case with $n = 100$ that the standard LISF server-assignment rule and the RR variant do not generate sufficient breaks, but the new server-assignment rules do. For large n,

the simulations confirm the MSHT fluid model formulas, but the MSHT fluid model provides only a crude approximation for the base case, so that simulation also provides an important contribution.

1.3 Related Literature and Organization

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 (2006b) 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).

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 et al. (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 introduce a general Markov process that describes the evolution of the system state for the D_1 server-assignment rule. It also can be used for other server-assignment rules that exploit the elapsed times since the last service completion and the last break. 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 an appendix.

2 The Stochastic Model for the D_1 Server-Assignment Rule

We consider the standard $M/GI/n$ multi-server queueing model with n 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 $\lambda \equiv \rho < 1$ 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 $EU = 1/\lambda$.

2.1 A Function-Valued Markov Process

Since we want to consider the D_1 server-assignment policy as well as alternatives, we extend the model. Let the target break duration be θ. We call the elapsed time since the last break (idle time of at least θ) the "age." Let $B(t, x, y)$ be the number of busy servers at time t with age at most x and elapsed current service time at most y and let $I(t, x, y)$ be 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 $x \geq y$ for $I(t, x, y)$). Let $Q(t)$ be the total number of customers in the system at time t; let $B(t) \equiv B(t, \infty, \infty)$ be the number of busy servers at time t; and let $I(t) \equiv I(t, \infty, \infty)$ be the number of idle servers at time t. We clearly have $B(t) = \min\{Q(t), n\}$ and $I(t) = \max\{n - Q(t), 0\}$.

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

$$(Q, B, I)_t \equiv (Q(t), B(t, \cdot, \cdot), I(t, \cdot, \cdot)) \equiv \{\{(Q(t), B(t, x, y), I(t, x, y) : x \geq 0, y \geq 0\} : t \geq 0\} \qquad (2.1)$$

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, for any initial state $(Q, B, I)_0$, $(Q, B, I)_t \Rightarrow (Q, B, I)$; i.e., as $t \to \infty$,

$$\{(Q(t), B(t, x, y), I(t, x, y)) : x \geq 0, y \geq 0\} \Rightarrow \{(Q, B(x, y), I(x, y)) : x \geq 0, y \geq 0\} \equiv (Q, B, I) \quad (2.2)$$

and when the initial state $(Q, B, 0)_0$ is the limit (Q, B, I), $(Q, B, I)_t$ becomes a stationary stochastic process. When we refer to the steady-state quantities, we omit the index t.

Remark 2.1 (*Relation between D_1 and LISF*) Because D_1 is a work-conserving server-assignment rule, the stochastic process $\{\{Q(t), B(t,\infty,y), I(t,\infty) : y \geq 0\} : t \geq 0\}$ is the same as for LISF or any other work-conserving server-assignment rule with $M/GI/n$ model. The D_1 rule only alters the server idle times and ages.

2.2 Conservation Laws

Conservation laws are important for understanding allocations of idleness in steady state (so we now omit t). Given that all arrivals are eventually served and that customer service times are not altered by any of the server-assignment rules, the following (well known) expressions for the steady-state mean values are valid:

$$E[B] = \rho n \quad \text{and} \quad E[I] = (1-\rho)n, \tag{2.3}$$

where $B \equiv B(\infty,\infty)$ and $I \equiv I(\infty,\infty)$. Formula (2.3) implies that, regardless of the server-assignment 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 (2.3) remains unchanged.

Let V denote the steady-state interval between successive service times, 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

$$1-\rho = \frac{E[V]}{E[V]+1}, \quad \text{so that} \quad E[V] = \frac{1-\rho}{\rho}. \tag{2.4}$$

From (2.4), we see that (i) the server-assignment rule cannot alter $E[V]$ and (ii) the target break $\theta = 5/3$ is 15 times larger than $E[V] = 0.1111$ in the base case with $\rho = 0.9$.

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

$$\beta = \frac{1}{E[T]}, \quad \pi_\beta = \frac{E[D]}{E[T]} \quad \text{and} \quad \pi_{\beta,I} = \frac{\pi_\beta}{1-\rho}. \tag{2.5}$$

We can combine (2.4) and (2.5) to deduce that idle times occur at rate $(1-\rho)/E[V] = \rho$, so the rate at which breaks occur can be represented as

$$\beta = \frac{(1-\rho)P(V \geq \theta)}{E[V]} = \rho P(V \geq \theta). \tag{2.6}$$

Lemma 2.1 (*upper bound on the rate of breaks*) *Given ρ and θ, the rate at which breaks occur is bounded above by*

$$\beta \leq \beta^* \equiv \frac{1-\rho}{\theta}. \tag{2.7}$$

which occurs if a proportion $p \equiv E[V]/\theta = (1-\rho)/\rho\theta$ *of the idle times are* θ *and the rest are* 0.

Proof. We can apply (2.6), observing the $P(V \geq \theta)$ is maximized over all possible distributions of V with mean fixed at $E[V] = (1-\rho)/\rho$ by the two-point distribution on θ and 0 that has the given mean. ∎

Remark 2.2 (*conservation laws in the fluid model*) The conservation laws in this section have natural analogs for the associated deterministic fluid model considered in §3. They are identical, except we remove the n in (2.3).

2.3 LISF and RR in the Base Case

We started by studying the idleness in the $M/GI/n$ model with the LISF and RR server-assignment rules. In the appendix we develop exact results and approximations for the steady-state distributions of: (i) the number of idle servers, (ii) the cumulative idleness in a time interval, and (iii) the idle-time distribution. For the $M/M/n$ base case with $n = 100$, $\rho = 0.9$, $E[S] = 1$ and $\theta = 5/3$, we find that the cumulative idleness over $[0, 40]$ is sufficient to produce effective work breaks, but the LISF and RR rules do not generate them frequently enough.

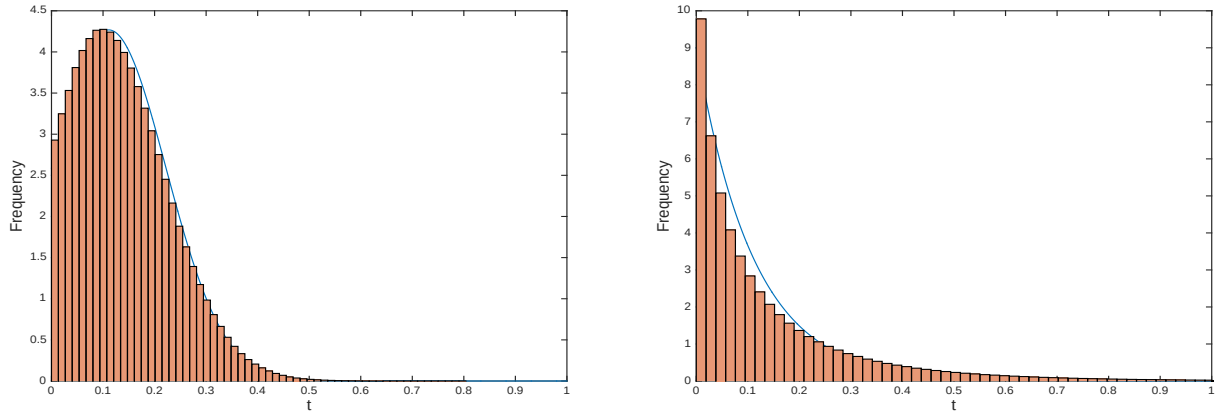
For example, in the base case, LISF produces a steady-state idle time V with approximately a truncated Gaussian distribution having $P(V = 0) = 0.215$, $E[V] = (1-\rho)/\rho = 0.1111$ and $SD(V) = 0.100$. Since $\theta = 5/3$ is 15.7 standard deviations above the mean, it is highly unlikely that an idle time will be a break.

In contrast, with RR, V has 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 θ is more than 9 standard deviations above the mean.

Figure 1 shows histograms estimated by simulation of the steady-state idle-time pdf with LISF and RR for the base case. In these figures 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 sufficient.

3 The MSHT Fluid Model for Age-Based Server-Assignment Rules

We can better understand why our server-assignment rules are attractive candidates for creating work breaks with large scale by considering the many-server heavy-traffic (MSHT) limiting fluid model,



(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.

which arises as the limit in a functional weak law of large numbers (FWLLN) for the stochastic model in §2. The age-based server-assignment rules are much easier to analyze for the fluid model because the discrete stochastic processes are replaced by continuous divisible deterministic processes, which we refer to as fluid processes. Thus, for the fluid model our proposed D_1 server-assignment rule achieves the maximum possible rate of breaks in Lemma 2.1.

3.1 Many-Server Heavy-Traffic (MSHT) Limits

For the MSHT FWLLN, we consider a sequence of $G/GI/n$ models indexed by n, where in model n the number of servers is n and the arrival rate is $\lambda_n = n\rho$ 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) many-server heavy-traffic 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.

The MSHT FWLLN states that

$$(\bar{Q}, \bar{B}, \bar{I})_{t,n} \Rightarrow (\bar{Q}, \bar{B}, \bar{I}) \quad \text{as} \quad n \to \infty \tag{3.1}$$

for each t (actually uniformly in t over bounded intervals), where we average for each n; i.e.,

$$(\bar{Q}, \bar{B}, \bar{I}) \equiv n^{-1}(Q, B, I)_{t,n} \tag{3.2}$$

with $(Q, B, I)_{t,n}$ being $(Q, B, I)_t$ defined in (2.1) above for model n and $(\bar{Q}, \bar{B}, \bar{I})$ is the limiting deterministic fluid process. We propose to approximate the performance of the stochastic process $(\bar{Q}, \bar{B}, \bar{I})$

for large n and t by the steady-state of the limiting fluid model, denoted by $(Q, B, I)_{\infty,\infty}$.

As depicted in Figure 2, there are actually four limits supporting this approximation. First, the

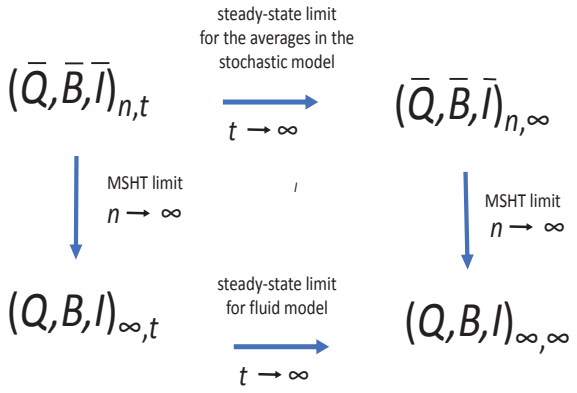


Figure 2: The four limits as $n \to \infty$ and $t \to \infty$ starting with the averages in the stochastic model (upper left) and leading to the steady-state of the MSHT fluid model (lower right).

assumed steady-state convergence in (2.2) implies associated limits as $t \to \infty$ for the averages in the stochastic model for each n, as shown in the top arrow. To get to the steady-state of the fluid, there are two possible iterated limits for the averages $(\bar{Q}, \bar{B}, \bar{I})_{t,n}$ in (3.2): $\lim_{n\to\infty} \lim_{t\to\infty}$ and $\lim_{t\to\infty} \lim_{n\to\infty}$. We will not derive these limits in this paper. Instead, to focus on the central applied issue, here we assume that these two iterated limits exist and coincide, and here derive the explicit form of the steady-state of the D_1 fluid model $(Q, B, I)_{\infty,\infty}$, which is shorthand for $\{(Q, B(x, y), I(x, y)) : x \geq 0, y \geq 0\}_{\infty,\infty}$, which we will hereafter refer to as $(Q, B(x, y), I(x, y))$.

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

We now consider the underloaded deterministic MSHT fluid model associated with the D_1 server-assignment rule. For the fluid model we let the capacity (maximum possible service rate) be 1 and refer to the model as the "G/GI" MSHT fluid model. (The fluid model for G arrivals is the same as for M.) The key parameters are the traffic intensity ρ (assumed to satisfy $0 < \rho < 1$), the target length of each break θ and the service-time cdf F (assumed to have a density and finite first two moments). We will focus on the steady-state behavior.

It is natural to think of the experience of individual atoms of fluid as following stochastic processes. For example, a major component of the $G/GI/n$ stochastic model for each n is a sequence of random service times. For each n, this is a sequence of i.i.d. random variables each distributed as a random variable S with cdf F, mean $E[S] = 1$ and a finite variance σ^2. It is natural to speak of random variables,

but the distributions should be interpreted as proportions in the fluid model. 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.

From Remark 2.1, we see that much of the fluid model is already contained in Whitt (2006a) and Liu and Whitt (2012a); we make the same technical smoothness assumptions here. A key insight is that, for the D_1 fluid model, we can directly apply the previous fluid results in those papers. .

The D_1 rule acts to first reassign all idle server fluid content with current idle time exceeding θ, with the fluid having largest idle time being assigned first, thus determining β. After that is accomplished, the D_1 rule assigns the idle server fluid content with current idle time less than θ, with the least age being assigned first. We will show that, with the continuous divisible deterministic fluid, the D_1 policy produces a remarkably simple steady-state solution, in which we achieve the maximum possible rate of breaks. For D_1, we show that there is a unique time τ^* such that all fluid that is in service beyond τ^* remains idle after a service completion for duration θ and thus receives a break, while all fluid that completes service before τ^* is immediately reassigned.

3.3 Relevant Renewal Theory

Given that we will consider 0-length idle times, we want to understand the implications of consecutive service times. An important role is played by the renewal counting process $N \equiv \{N(t) : t \geq 0\}$ associated with those service times, i.e.,

$$N(t) \equiv \max\{k \geq 0 : S_0 + S_1 + \cdots + S_k \leq t\}, \quad t \geq 0, \tag{3.3}$$

where $S_0 \equiv 0$.

We will exploit the mean of the renewal process, called the *renewal function*,

$$m(t) \equiv E[N(t)], \quad t \geq 0, \tag{3.4}$$

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

$$R(t) \equiv S_{N(t)+1} - t, \quad t \geq 0. \tag{3.5}$$

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

$$E[R(t)] = E[S](E[N(t)] + 1) - t = E[N(t)] + 1 - t \quad \text{for all} \quad t \geq 0. \tag{3.6}$$

or, equivalently,

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

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).

Because the service distribution has a density (and thus is nonlattice) with $\sigma^2 < \infty$, see Proposition 3.4.8 of Ross (1996),

$$R(t) \Rightarrow S_e \quad \text{as} \quad t \to \infty \tag{3.8}$$

and

$$E[R(t)] \to E[S_e] = \frac{ES^2]}{2E[S]} = \frac{E[S](c_s^2 + 1)}{2} \quad \text{as} \quad t \to \infty, \tag{3.9}$$

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

$$F_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]} \int_0^t P(S > u)\, du, \quad t \geq 0. \tag{3.10}$$

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

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}, \tag{3.11}$$

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

$$E[S_e] = \frac{E[S^2]}{2} = \frac{1 + c_s^2}{2}, \tag{3.12}$$

where $c_s^2 \equiv \sigma^2/E[S]^2 = \sigma^2$ and

$$Var(S_e) = E[S_e^2] - (E[S_e])^2 = \frac{E[S^3]}{3} - \left(\frac{E[S^2]}{2}\right)^2. \tag{3.13}$$

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

$$R(t) \approx S_e \quad \text{and} \quad E[R(t)] \approx E[S_e]. \tag{3.14}$$

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 §4.5 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).

3.4 The Performance of the Age-Based Server Assignment Rules

Recall that we consider the G/GI fluid model with: (i) fluid service capacity 1, (ii) arrival rate $\rho < 1$, (iii) service-time proportions with cdf $F(x) \equiv P(S \le x)$ having pdf f with mean 1 and finite variance σ^2, (iv) the D_1 server-assignment rule with target work breaks of length θ, where $m \equiv (1-\rho)/\rho < \theta$ and (v) in steady-state. We assume that the service-time renewal function $m(t)$ in (3.4) is strictly increasing and continuous on $[0, \infty)$. Let $\stackrel{\rm d}{=}$ 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 τ^* is given a break of length θ, and thus is assigned exactly θ time units later, while all fluid completing service with with age less than τ^* is reassigned instantaneously and so experiences 0 idle time. The critical time τ^* is the unique root of the equation*

$$m(\tau^*) = \frac{1}{p} - 1 > 0, \tag{3.15}$$

where $p \equiv (1-\rho)/\rho\theta < 1$ and $m(t)$ is the renewal function associated with the service-time cdf F in (3.4). As a consequence, work breaks (idle times of length at least θ) occur at the upper bound rate from Lemma 2.1,

$$\beta^* = \frac{1-\rho}{\theta} = p\rho, \tag{3.16}$$

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

$$T^* \stackrel{\rm d}{=} \tau^* + R(\tau^*) + \theta = N(\tau^*) + 1 + \theta, \tag{3.17}$$

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

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

(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

$$b(x) = \beta^* 1_{\{0 \le x < \tau^*\}} + \beta^* P(R(\tau^*) \ge x - \tau^*) 1_{\{\tau^* \le x < \infty\}} \tag{3.19}$$

and

$$g(x) = 0 \cdot 1_{\{0 \le x < \tau^*\}} + \beta^* P(R(\tau^*) \le x - \tau^*) 1_{\{\tau^* \le x < \tau^* + \theta\}}$$

$$+\beta^*(P(x-\tau^*-\theta \le R(\tau^*) \le x-\tau^*)1_{\{\tau^*+\theta \le x<\infty\}} \tag{3.20}$$

for β^* *in* (3.16)*,* τ^* *the solution of equation* (3.15) *and* $R(t)$ *the renewal excess in* (3.5)*. As a consequence, the associated cumulative functions satisfy*

$$0 = I(\tau^*, \infty) < I(x, \infty) < I(\infty, \infty) \equiv I = 1-\rho, \quad \tau^* < x < \infty, \tag{3.21}$$

and

$$B(\tau^*, \infty) = \beta^*\tau^* < B(x, \infty) < B(\infty, \infty) \equiv B = \rho, \quad \tau^* < x < \infty. \tag{3.22}$$

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

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 θ 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 τ^*) is immediately reassigned upon service completion. Moreover, by Lemma 2.1 and Remark 2.2, 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 a new "macro service-times," which combines the consecutive service times experienced between breaks. Given τ^*, the new combined service-time is $\tilde{S} \equiv \tau^* + R(\tau^*)$ with cdf is $\tilde{F}$ and pdf $\tilde{f}$. Thus, in the underloaded D_1 fluid model, each atom of fluid experiences alternating breaks of length θ, 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 if we consider the service-time cdf $\tilde{F}$ and a fluid arrival process with rate $\beta^* E[\tilde{S}]$. The higher arrival rate is balanced by the longer service time; i.e.,

$$b(x) = (\beta^* E[\tilde{S}])\tilde{f}_e(x) = (\beta^* E[\tilde{S}])(\tilde{F}^c(x)/E[\tilde{S}]) = \beta^* \tilde{F}^c(x), \tag{3.23}$$

which coincides with (3.19). The density b in (3.23) then coincides with (3.2) in Theorem 3.1 (a) of Whitt (2006a). The density g in (3.20) follows from observing that all idle fluid remains exactly for time θ after it arrived.

It remains to show that there exists a unique pair (τ^*, β^*) satisfying (3.15) and (3.16). To start, the renewal function has a unique inverse, because we have made assumptions that ensure it is continuous and strictly increasing. Thus, (3.15) necessarily has a unique solution.

On the other hand, given the form of the busy-server density $b(x)$ in (3.19), and the total busy server content $B = \rho$, we have $\rho = \beta^*\tau^* + \beta^* E[R(\tau^*)] = \beta^*(m(\tau^*) + 1)$, where β^* is the rate breaks occur. Hence,

$$\beta^* = \rho/(m(\tau^*) + 1). \tag{3.24}$$

Given the D_1 policy, For T^* in (3.17), we also have $T^* \stackrel{\mathrm{d}}{=} \tau^* + R(\tau^*) + \theta$, where $R(\tau^*)$ is the residual service time beyond τ^*, so that

$$\beta^* = \frac{1}{E[T^*]} = \frac{1}{m(\tau^*) + 1 + \theta}. \tag{3.25}$$

Combining (3.24) and (3.25), we obtain the unique solution with τ^* in (3.15) and β^* in (3.16). We remark that, as an alternative argument, we could also apply (2.4) and Remark 2.2: On average, each server experiences, $m(\tau^*)$ idle times of length 0 followed by one of length θ. Hence,

$$E[V] = \frac{\theta}{m(\tau^*) + 1} = m = \frac{1 - \rho}{\rho}, \tag{3.26}$$

from which we also obtain (3.15). Because there is a unique solution to equation (3.15), 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 τ^* 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. ∎

Corollary 3.1 (*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 $\eta \geq 1 - \rho$.*

Remark 3.1 (*the experience of individual servers*) Individual servers (atoms of fluid) experience alternating busy periods distributed as $T_B \stackrel{\mathrm{d}}{=} \tau^* + R(\tau^*)$ 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 (3.19) and (3.20) 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.

Remark 3.2 (*exponential service*) The solution in Theorem 3.1 simplifies if the service time S is a mean-1 exponential, $M(1)$, because then $m(\tau^*) = \tau^*$ and $R(x^*) \stackrel{\mathrm{d}}{=} M(1)$, so that $\tau^* = (1/p) - 1$ and $T^* \stackrel{\mathrm{d}}{=} \tau^* + \theta + M(1)$.

3.5 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 θ 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 θ (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 λ, δ and α 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 α_0 and α_θ be the rate of assignment of fluid that has been idle for time 0 and θ, respectively. If feasible, then we have $\alpha = \alpha_0 + \alpha_\theta$. By Lemma 2.1, the maximum possible value of breaks is $\alpha_\theta = \beta^* = p\rho$, 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 θ 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 θ), 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

$$T \approx \theta + \sum_{i=1}^{N(p)} S_i, \tag{3.27}$$

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 σ^2 that are independent of $N(p)$, so that

$$E[T] = \theta + \frac{1}{p} = \theta + \frac{\theta}{E[V]} = \frac{\theta}{1-\rho} = \frac{1}{\beta^*}, \tag{3.28}$$

as it should, and

$$Var(T) = Var(S)E[N(p)] + E[S]^2 Var(N(p)) = \frac{\sigma^2}{p} + \frac{1-p}{p^2} = \frac{p\sigma^2 + 1 - p}{p^2} = \left(\frac{\rho\theta}{(1-\rho)}\right)^2. \tag{3.29}$$

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 §4.1 and §4.2 we indicate how we implement the D_1 and $SISF$ server-assignment rules in the simulation. In §4.3 we discuss how we execute the simulation and perform the statistical estimates. In §4.4 we report simulation results for the $M/M/n$ model in the base case. In §4.5 we report additional results for the D_1 rule with a hyperexponential service-time distribution.

4.1 Implementing the D_1 Server-Assignment Rule

Let any idle time greater than or equal to θ be called an (unannounced) *break*. Following an object-oriented-programming approach, we treat each server as an "object" from a "server class;" e.g., see Horstmann (2002). Each server contains three "instance variables," namely its identity number, service completion time and break end time. To implement D_1 in a virtual environment, we maintain for each busy server a service-completion time; this value is infinity by default for idle servers. Similarly, for each idle server we maintain a break end time by acting as if its current idle period will eventually develop into a break; this value is infinity by default for busy servers.

We conduct a discrete-event simulation in which no change in the system occurs between consecutive events. Thus the simulation jumps in time from one event to the next. For D_1, three types of events can happen: (i) customer arrival, (ii) customer departure and (iii) end (completion) of break. The algorithm maintains (a) a FIFO queue for waiting customers, (b) a high-priority-queue (HPQ) containing all servers whose elapsed idle time exceeds θ and (c) a sorted list L with all the server other than those in the HPQ in the order of increasing ages.

At each arrival epoch, we look for idle servers in the HPQ. If any, assign the server at the head of the HPQ, reset its age to zero and move the server to the head of the list L. Otherwise, we scan through the list L to find an idle server with the shortest age. We make assignment if there exists such a server in L; otherwise the customer is put in queue.

For the selected idle server, the algorithm generates a service requirement S from the service-time distribution and resets its service completion time to $t + S$. Then we find the minimum service-completion time among all busy servers and update the departure time accordingly. Searching for the closest service completion time can be costly if the number of servers n is large. To accelerate the search, we arrange all busy servers in a binary heap where the root node is the server with the minimum service completion time. Computationally this is efficient, because it takes $O(1)$ operations to extract the minimum and $O(\log(n))$ operations to restore the heap structure as new elements enter.

At each departure epoch, we first look for customers in queue. The server gets assigned if the queue is nonempty. Otherwise the server becomes idle. At this time, we reset its service completion time to infinity, set the break-end time to $t + \theta$ and update the closest break-end time.

At the end of a break, we move the idle server to the back of the HPQ. That prevents a break from being much greater than θ, because we first assign idle servers from the HPQ.

4.2 Implementing the $SISF$ Server-Assignment Rule

To implement $SISF$, we stipulate that each server belongs to one of the three places: (i) the busy-server pool (BSP), (ii) the low-priority-queue (LPQ) for assignment or (iii) the high-priority-queue (HPQ) for assignment. For each busy server, we maintain the time for the current task to complete and set this value to infinity for idle servers. Similarly, for each idle server in the LPQ we maintain a break end time by assuming that its current idle period would eventually develop into an idle period of length θ; we set this value to infinity for busy servers as well as (idle) servers in the HPQ.

At each arrival epoch, we look to see if the HPQ is empty; if it is nonempty, we assign the server at the head of the HPQ. If the HPQ is empty, we look for idle servers in the LPQ and assign a server

(if any) from the back of the LPQ. We use the first-in first-out (FIFO) discipline in the HPQ, but the last-in first-out (LIFO) discipline in the LPQ. Because the HPQ is FIFO, we use a circular array to implement the HPQ. The LPQ is a LIFO queue except that when a break finishes the server at the head of the LPQ joins the back of the HPQ (at this time we reset its break end time to infinity). We therefore use a linked-list to efficiently implement the LPQ.

Once a server gets assigned, we put the server into the BSP and attach to it a service completion time by sampling from the service-time distribution. Here we calculate (update) the minimum service completion time and let it be the time of next departure. Again we use a binary heap as we did for rule D_1 to speed up the searches for the minimum service completion time among all busy servers.

If no customers wait in queue, each customer departure is followed by a removal of a server from the BSP and its joining the LPQ. At this time, we set its service completion time to infinity and schedule its next long-idle-period end time.

4.3 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 and use sample quantiles to construct the CI.

4.4 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. 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 Lemma 2.1. The fluid model provides the limiting case of $n = \infty$.

system	D_1		$SISF$	
size	$P(V_n \le 0.1)$	$P(V_n \ge \theta)$	$P(V_n \le 0.1)$	$P(V_n \ge \theta)$
$n = 25$	0.7917 ± 0.0018	0.0163 ± 0.0003	0.8257 ± 0.0012	0.0217 ± 0.0003
$n = 100$	0.8240 ± 0.0013	0.0223 ± 0.0004	0.8341 ± 0.0008	0.0293 ± 0.0004
$n = 250$	0.8498 ± 0.0007	0.0317 ± 0.0003	0.8698 ± 0.0005	0.0386 ± 0.0003
$n = 1000$	0.8896 ± 0.0008	0.0492 ± 0.0007	0.9028 ± 0.0005	0.0546 ± 0.0005
$n = 5000$	0.9155 ± 0.0002	0.0601 ± 0.0010	0.9236 ± 0.0003	0.0628 ± 0.0002
$n = \infty$	0.9333 ± 0.0000	0.0633 ± 0.0000	0.9333 ± 0.0000	0.0667 ± 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 §3.5, 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 $n = 100$.

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 3.1. Figure 3 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 (3.19) and (3.20). Table 3 reports estimations of the mean and standard deviation of these age random variables for D_1 as a

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

function of n. As before, the case $n = \infty$ corresponds to the fluid model.

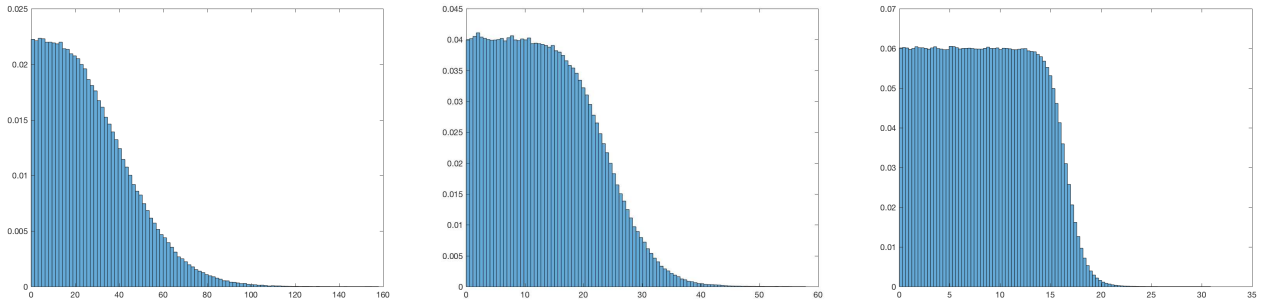
	Busy		Idle	
	$E[A_B]$	$std(A_B)$	$E[A_I]$	$std(A_I)$
$n = 100$	26.510 ± 0.051	19.146 ± 0.072	41.725 ± 0.068	19.725 ± 0.083
$n = 500$	13.178 ± 0.016	8.395 ± 0.033	24.858 ± 0.019	6.565 ± 0.024
$n = 1000$	10.518 ± 0.011	6.380 ± 0.018	20.865 ± 0.013	3.828 ± 0.017
$n = 5000$	8.399 ± 0.004	4.935 ± 0.011	17.378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

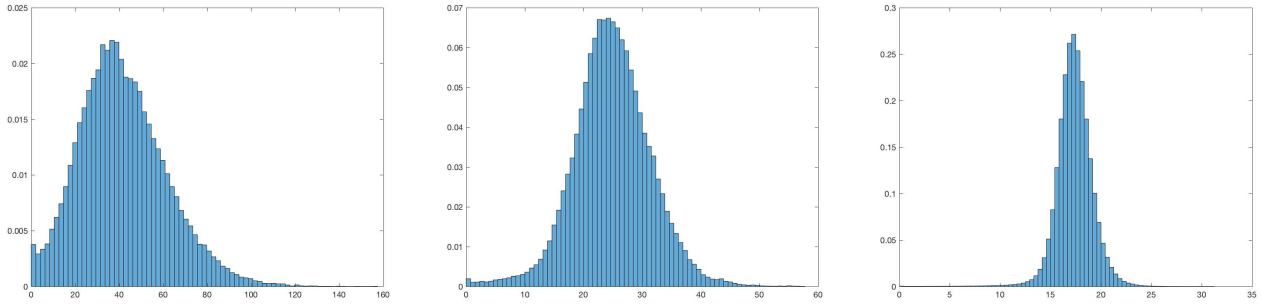
It is also useful to look at the pattern of successive idle times over a long horizon. Figure 4 displays successive idle-times for a set of randomly selected servers in the $M/M/n$ base case. The vertical axis measures the length of an idle-time and the horizontal axis indexes the successive idle times.

Figure 4 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 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



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



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

Figure 3: 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 $n = 5000$.

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 5, 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.

4.5 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 (3.14) 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 §3.3, 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 $SD(R(\tau^*)) = 1$.

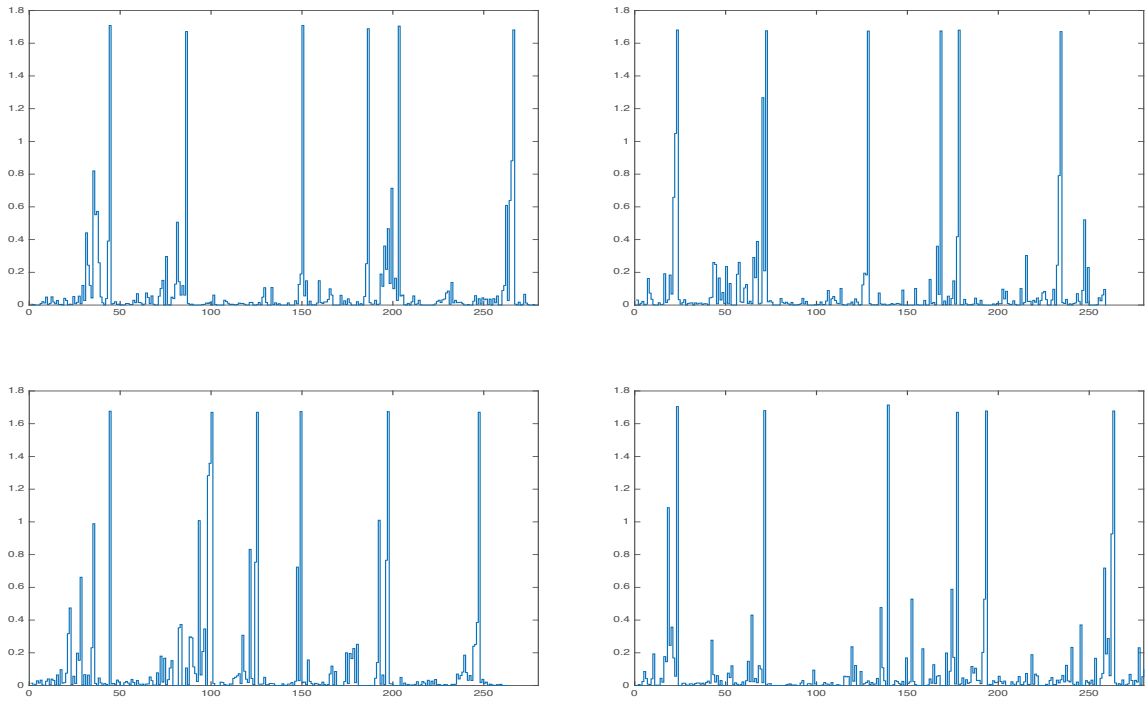
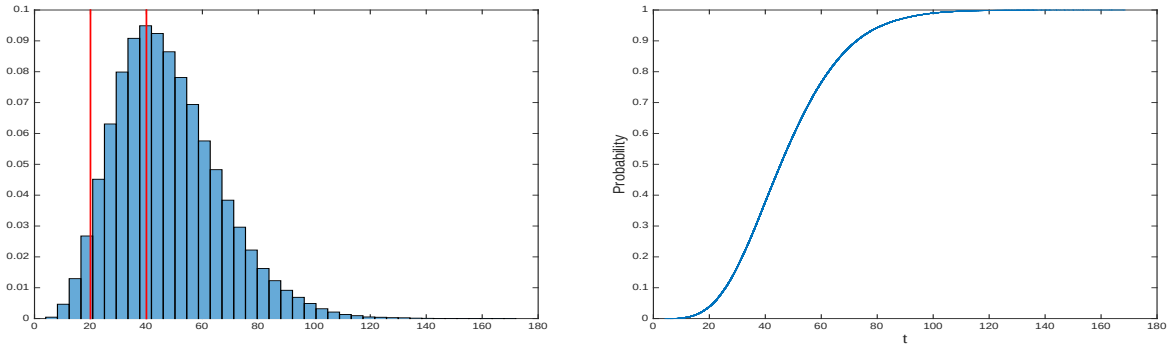


Figure 4: Four sample paths of successive idle times over a time interval of length 300 for D_1 in the base case.



(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 $\theta = 5/3$

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 (3.14), 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 ± 0.098	22.059 ± 0.102	37.622 ± 0.106	23.851 ± 0.115	41.663 ± 0.126	23.7531 ± 0.131
$n = 250$	18.277 ± 0.085	13.584 ± 0.092	29.473 ± 0.089	13.922 ± 0.079	31.748 ± 0.095	13.473 ± 0.104
$n = 1000$	10.813 ± 0.062	7.249 ± 0.071	20.031 ± 0.075	5.883 ± 0.058	20.495 ± 0.047	5.568 ± 0.072
$n = 5000$	8.765 ± 0.022	5.789 ± 0.030	17.017 ± 0.028	4.150 ± 0.025	16.725 ± 0.024	3.876 ± 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 τ, paralleling τ^*. For each server, we keep track of the age and announce a break when the age exceeds τ; the server is then off duty for time θ. (For a busy server, the break begins upon service completion; for an idle server, the break begins immediately.) Any breaks that occur before time τ are unannounced breaks.

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 η 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 η 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 τ and η to minimize the cost function in (5.1).

5.1 Implementing the D_2 Server-Assignment Rule

We consider five types of events: customer arrival, customer departure (service completion), due for a break, announced break completion and unannounced break completion. We first explain how to treat the control parameter τ with $\eta = \infty$, so it plays no role. Afterwards, we discuss the modifications to include η.

At each customer arrival epoch, we look for available servers. If any, assign the server with the shortest age. For the selected idle server, the algorithm generates a service requirement S from the

service-time distribution and resets its service completion time to $t + S$. Then we find the minimum service-completion time among all busy servers and update the departure time accordingly. If there are no servers available, the arriving customer waits in queue.

At each customer departure epoch, we look for customers in queue. If there is a customer waiting, assign the newly-available server to the head-of-line customer. Otherwise, let the newly-available server either become idle or start a break depending on whether or not a high priority designation (to be explained momentarily) was given. If a high priority designation was given, the break is announced and the server is off duty and not available to provide service for a duration θ after that time. Otherwise it remains idle.

At each break due time (when a server's age reaches τ), if the server is busy, then we give the server a high priority designation indicating that its next idle period will be replaced by an announced break. If the server is idle, then the server starts an announced break and goes off duty for the duration θ. (The elapse idle time at the time of the break is not included in the break, and is counted as part of the total idle time.)

At each announced-break-end time, we first reset the server's age to zero. We assign this newly-available server to a customer if the queue is not empty. Otherwise, the newly-available server stays idle.

At each unannounced-break-end time, we reset the server's age to zero. At this time the queue must be empty because this server was idle but on call.

We now discuss modifications to treat the bound η.

Each time a break is due, if the server is idle and the number of off-duty servers is less than η, then a break is announced and the server is not available to provide service for the duration θ. On the other hand, if the server is idle and the the number of off-duty servers equals η, then we give the server a high-priority designation and do not make the break announcement. Meanwhile, we keep track of the elapsed time since this high priority designation has been assigned.

At each customer departure epoch, if the queue is non-empty, then the server is assigned to the customer at the head of the queue. Hence, suppose that the queue is empty. If a high priority designation was given to that server and the number off-duty servers is less than η, then the break is announced and the server no longer provides service for the duration θ. Otherwise the server stays idle but on-call.

At each announced-break-end time, there is a newly-available server. We reset the server's age to zero. We assign this newly-available server to a customer if the queue is not empty. Otherwise, the

newly-available server stays idle. At the meantime we look for other idle servers with a high-priority designation. If any, choose the one with the longest elapsed time since it received this high priority level and announce the break.

5.2 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 τ and η 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 $p_D^* = 0.223$.)

	$\eta = 4$	$\eta = 6$	$\eta = 8$	$\eta = 10$
τ	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 τ and η for $n = 100$ and $\theta = 5/3$. The entries in bold are maximal over η for that τ.

	$\eta = 4$	$\eta = 6$	$\eta = 8$	$\eta = 10$
τ	p_D	p_D	p_D	p_D
$\tau = 15$	0.3368 ± 0.0018	0.4141 ± 0.0026	0.4860 ± 0.0020	0.5414 ± 0.0023
$\tau = 20$	0.3330 ± 0.0021	0.4076 ± 0.0021	0.4603 ± 0.0023	0.4855 ± 0.0021
$\tau = 25$	0.3319 ± 0.0022	0.3937 ± 0.0017	0.4218 ± 0.0020	0.4339 ± 0.0025
$\tau = 30$	0.3291 ± 0.0018	0.3739 ± 0.0025	0.3887 ± 0.0025	0.3974 ± 0.0024
$\tau = 35$	0.3246 ± 0.0021	0.3510 ± 0.0024	0.3598 ± 0.0022	0.3663 ± 0.0024
$\tau = 40$	0.3206 ± 0.0020	0.3342 ± 0.0027	0.3413 ± 0.0020	0.3449 ± 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 τ and η for $n = 100$ and $\theta = 5/3$.

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 θ, 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.

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

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

These tables show that η makes much greater difference than τ. Moreover, there is a strong tradeoff in the choice of η. All three of p_D, p_A and p_B are monotone in τ, but p_A and p_B are not monotone in η for fixed τ. The entries in bold show that optimal η for each τ. The values of η where these maximal proportions occur are decreasing in τ. The corresponding plots for other weights w are shown in the appendix. Figure 6 shows the cost in (5.1) as a function of τ and η 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 τ, 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 τ, but otherwise uses the standard LISF rule for 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,

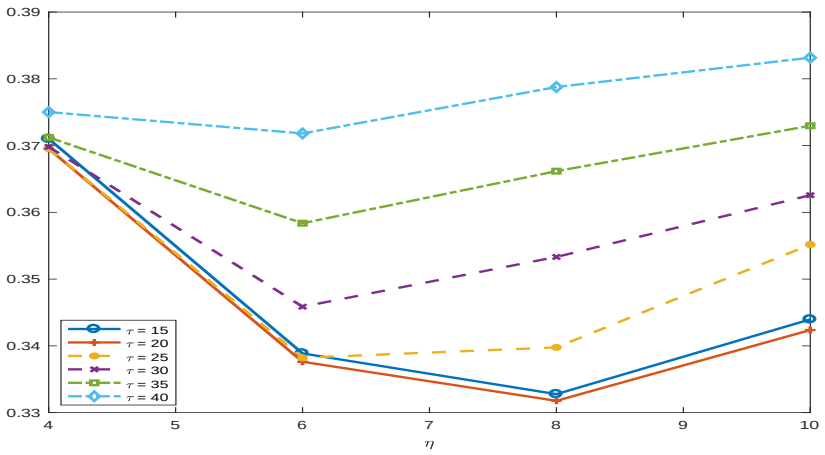


Figure 6: The cost in (5.1) as a function of τ and η for $D_2(\theta, \tau, \eta)$ in the base case with $n = 100$, $\theta = 5/3$ and $w = 0.5$

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 η.

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 §2.3, 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 shows 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 Lemma 2.1. However, in §3.5 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 τ and η 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. Finally, there remain many 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.

Horstmann C (2002) Big java early objects. *Interfaces* 9(10):10.

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, 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.

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.

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 (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.