Abstract

Sun and Whitt (2017) developed a class of server-assignment rules that can help create effective breaks for servers from naturally available idleness. They found that the standard longest-idle-server-first rule and the alternative random routing generate breaks too infrequently. This paper provides theoretical justifications for these empirical findings by establishing the many-server heavy-traffic limits for server idle times with customary assignment rules. We gain insights into the server idle-busy activity patterns resulting from the server-assignment schemes in the system.

Keywords: multiple-server queues; many-server heavy-traffic limits for queues; server-assignment rules; idle-busy activity pattern.

1. Introduction

This is a sequel to the paper [10] in which the authors 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) and the random-routing (RR) alternative do not generate breaks often enough, the authors studied the one-parameter rule $D_1$ yielding unannounced work breaks while maintaining work conservation and then studies three-parameter refined rule $D_2$ producing announced work breaks by sacrificing work-conservation.

In this paper, we provide theoretical support for the performance of the LISF rule and the RR rule by analyzing them in the quality-driven (QD) many-server heavy-traffic (MSHT) regime in which the number of servers and the arrival rate increase toward infinity, while the traffic intensity (workload per server) is held fixed at $<1$. We deliberately consider exponential services due to technical considerations on which we elaborate momentarily. Our analysis indicates that, with the standard

* Xu Sun

Email address: xs2235@columbia.edu (Xu Sun)
LISF policy, an idle period between two successive service times tends to be a constant value subjected to a Gaussian noise. In contrast, the RR rule produces idle periods which we show to be approximately exponentially distributed. These MSHT limits, along with the $D_1$ fluid model considered in [10] allow us to gain important insights into how server idle-busy activity patterns can be shaped by server-assignment schemes in the system.

The paper is structured as follows. In §2 we describe the queueing model. In §3 we establish the MSHT functional central limit theorem (FCLT) and the functional weak law of large numbers (FWLLN) for the idleness (number of idle servers) process as well as the central limit theorem (CLT) and weak law of large numbers (WLLN) for the steady state idleness. These results holds true for a class of server-assignment rules to which the LISF rule and the RR rule are two special cases. for the LISF policy. We derive the MSHT limits of the steady-state idle time under the RR policy in §5.

Notation and Conventions. We denote by $\mathbb{R}$ the set of real numbers. For real numbers $a$ and $b$, $a \wedge b \equiv \min(a,b)$, $a \vee b \equiv \max(a,b)$ and $[a]^+ \equiv a \vee 0$. We use $\lfloor a \rfloor$ to denote the maximum integer that is less than or equal to $a$. For an event (set) $A$, let $1_A(\cdot)$ denote the indicator function of the event $A$ and $A^c$ be its complement. Let $\Rightarrow$ denote convergence in distribution or convergence under the Lévy-Prokhorov metric. Let $\mathcal{D}$ be the usual function space of right-continuous real-valued functions on the interval $[0, \infty)$ with left limits, as in [14]. Convergence $x_n \to x$ in $\mathcal{D}$ at continuous limits $x$ is equivalent to uniform convergence over bounded intervals. Let $\eta \equiv \eta(t) \equiv 1$, $t \geq 0$, be the unit function in $\mathcal{D}$. Let $\mathcal{D}^k$ be the $k$-fold product space of $\mathcal{D}$ with itself with the usual product topology. We use $\overset{d}{=} \equiv$ to denote equal in distribution. Accordingly, we use $\overset{d}{\approx}$ to represent approximately equal in distribution. Lastly, we use $N(\mu, \sigma^2)$ to represent the normal distribution with mean $\mu$ and variance $\sigma^2$ and $\mathcal{B}(p)$ the Bernoulli distribution with parameter $p$. We abbreviate “random variable” to “r.v.”.

2. A Family of $GI/M/n$ Models

We consider the standard $GI/M/n$ multi-server queueing model with $n$ homogeneous servers working in parallel and a centralized queue with unlimited waiting space. Customers in queue are served in a first-come first-served (FCFS) order.

We start with customer arrivals. To formally describe the arrival process $A_n$ for the $n$th stochastic model, first introduce a sequence of independent r.v.s $\{U_i; i \geq 1\}$ taking values in $\mathbb{R}$, with $\{U_i; i \geq 2\}$ identically distributed having a cumulative distribution function (CDF) $G(x)$ with mean $\mathbb{E}[U_2] = 1/\rho$ and coefficient of variation (CV) $c_a$. The first element $U_1$ is assumed to follow the associated stationary-excess distribution $G^e(x)$; i.e.,

$$G^e(x) \equiv P(U_1 \leq x) = \frac{G^e(x)}{\mathbb{E}[U_2]}$$
with \( G^c(x) \equiv 1 - G(x) \). Throughout the paper we will assume that the distribution \( G(\cdot) \) has finite third moment. Thus, the distribution \( G^c(\cdot) \) has finite second moment.

Let \( S_k \equiv \sum_{i=1}^k U_i \) be the partial sum associated with the sequence \( \{U_i; i \geq 1\} \) and \( A \) be the associated counting process. That is,

\[
A(t) \equiv \min_k \{k | S_k \geq t\}.
\]

Then \( \{A(t); t \geq 0\} \) is a delayed renewal process and is time-stationary; see §3.5 of [9] for backgrounds. We directly assume the arrival process for the \( n \)th system to be

\[
A_n \equiv A(nt) = \min_k \{k | S_k \geq nt\}.
\]

Suppose that \( \{U_{n,i}; i \geq 1\} \) are successive inter-arrival times of the counting process \( A_n \). It is a simple exercise to argue that \( \{U_{n,i}; i \geq 2\} \) are i.i.d. having a CDF \( G_n(x) \equiv G(nx) \) with mean \( \mathbb{E}[U_{n,1}] = 1/(n\rho) \) and CV \( c_a \). Similarly, the first element \( U_{n,1} \) has a CDF \( G_{n}^e(x) \equiv G^e(nx) \). One can easily verify that the asymptotic mean and variance of the sum satisfy Lyapunov’s condition. An application of the functional Lindeberg-Feller central limit theorem applied to a triangular array gives us

\[
\sqrt{n} \left( \sum_{i=1}^{[nt]} U_{n,i} - \rho^{-1}t \right) \Rightarrow \rho^{-1}c_a W(t) \quad \text{in } \mathcal{D} \quad \text{as } n \to \infty
\]

where we have used \( W(\cdot) \) to represent a standard Brownian motion; see Theorem 4.1 of [8]. By the continuous mapping theorem (CMT) with inverse mapping (see, e.g., Theorem 13.7.1 of [14]) and (3), we obtain that

\[
n^{-1/2}[A_n(t) - n\rho t] \Rightarrow -\sqrt{\rho c_a^2} W(t) \quad \text{in } \mathcal{D}
\]

jointly with (3). Note that \( W \) in (4) is the same Brownian motion as given in (3).

Paralleling the base case considered in [10], we assume service times to be exponentially distributed with rate \( \mu = 1 \), independent of the scaling parameter \( n \). Our assumption about exponential services is based on two considerations. (i) The assumption guarantees that each idle period is initialized in the steady state of the queueing system; see Lemma 3.1; (ii) for \( GI/M/n \) type models it can be shown that the limit of steady-state idleness distribution coincides with the steady-state distribution of the limiting idleness process; see the discussion preceding Proposition 3.1.

The stochastic system evolves over time as a consequence of arrivals, service completions and server assignments. Hence, to understand the full evolution of the system, it only remains to specify how the servers are assigned to customers when multiple idle servers are present in the system. Server-assignment rules can be classified into two types: work-conserving or non-work-conserving. Work-conserving policies immediately assign one of the idle servers to a customer whenever there is a
customer in need of service and there is an idle server. Non-work-conserving policies
might let the customer wait in queue and assign to it an available server until a later
time.

In the present study we restrict ourselves to the family of work-conserving poli-
cies. In particular, we consider two customary server-assignment rules, namely the
longest-idle-server-first (LISF) rule and the randomized-routing (RR) rule. The
LISF rule assigns the entering customer to the idle server that has been idle the
longest among all idle servers (if any). If no server is idle, the customer waits in
queue. In contrast, the RR rule assigns the entering customer to each server in
the idle-server pool equally likely. That is, if there are \( k > 0 \) idle servers when an
customer arrives, each idle server will be chosen with probability \( \frac{1}{k} \). If no server
is idle, the customer waits in queue.

3. Supporting QD MSHT Limits

The results developed in this section are useful for all work-conserving server-
assignment policies. Let \( X_n(t) \) represent the number of customers in system at time
\( t \) for the \( n \)th model. With non-exponential inter-arrival times, the time sequence
\( \{X_n(t); t \geq 0\} \) does not form a continuous-time Markov chain (CTMC). But the
stationary variable \( X_n(\infty) \) can be obtained by considering the embedded discrete-
time Markov chain (DTMC) at arrival epochs. To maintain the Markov property
in continuous time, one could expand the state space by inserting the excess or
residual inter-arrival time \( U^*_n(t) \) as the second state variable. Then the process
\( X_n(t) \equiv (X_n(t), U^*_n(t)) \) is a CTMC. The following lemma is useful for the analysis
of the idle-time distribution in §4 and §5.

**Lemma 3.1.** (departures that see time averages) Suppose that the system is initial-
ized in steady state; i.e., \( X_n(0) \) follows the stationary distribution of the embedded
DTMC. Then each departure after time 0 sees the system as in steady state.

**Proof of Lemma 3.1.** With exponential service, the reverse Lack of Bias Assumption
(LBA) in [5] trivially holds for the departures; see, e.g., Example 2 on p.165 of [5].
An application of Theorem 6 in [5] allows us to conclude that each departure observes
the system as if in steady state without that leaving customer.

It is well known since Iglehart [3] that, for the \( GI/M/n \) model, the QD MSHT
limit of the number-in-system process \( X_n(t) \) is asymptotically equivalent to what it
is for the corresponding \( GI/M/\infty \) infinite-server (IS) model.

Let \( I_n(t) \) be the number of idle servers at time \( t \). In general,

\[
I_n(t) = [n - X_n(t)]^+.
\]  

Hence, the process \( I_n \) can be treated as a deterministic functional of the process
\( X_n \). Note that in the QD MSHT limit, we can act as if \( X_n(t) \leq n \), so that \( I_n(t) =
n - X_n(t) \geq 0 \).
We now state the basic QD MSHT limits and justify the heuristic reasoning above. For that purpose, let the diffusion-scaled queue-length and idleness (number of idle servers) processes be
\[
(\hat{X}_n(t), \hat{I}_n(t)) \equiv n^{-1/2} (X_n(t) - n\rho, I_n(t) - n(1 - \rho)), \quad t \geq 0,
\]
and let the associated fluid-scaled processes be
\[
(\bar{X}_n(t), \bar{I}_n(t)) \equiv n^{-1} (X_n(t), I_n(t)), \quad t \geq 0.
\]

Then, as reviewed on p.175 of [13], the basic MSHT QD limit theorems for \( X_n(t) \) are an FCLT and its FWLLN consequence. In the following result, we extend it to a joint FCLT and FWLLN for the pair \((X_n(t), I_n(t))\). For \( x \in \mathcal{D} \), let the supremum function be defined by
\[
x^{\uparrow}(t) \equiv \sup_{0 \leq s \leq t} \{x(s)\}, \quad t \geq 0,
\]
and let the uniform norm for functions on \([0, t]\) be defined by
\[
\|x\|_t \equiv \sup_{0 \leq s \leq t} \{|x(t)| = |x^{\uparrow}(t)|, \quad t \geq 0.
\]

**Theorem 3.1.** (QD MSHT FCLT and FWLLN for the idleness process) For the sequence of GI/M/n models in the QD MSHT regime with \( \rho_n \equiv \rho < 1 \), if
\[
\hat{X}_n(0) \equiv n^{-1/2} (X_n(0) - n\rho) \Rightarrow X(0) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty,
\]
where \( X(0) \) can be either a constant or a real-value r.v., then
\[
(\hat{X}_n, \hat{I}_n) \Rightarrow (X, I) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty,
\]
where \( I \equiv -X \) and \( X \equiv \{X(t) : t \geq 0\} \) is an Ornstein-Uhlenbeck (OU) diffusion process with infinitesimal drift function \( \mu(x) \equiv -x \) and constant infinitesimal variance function \( \sigma^2(x) \equiv \rho(1 + c_a^2) \), so that the steady-state distributions of \( X \) and \( I \) are given by
\[
I(\infty) = -X(\infty) \overset{d}{=} N(0, -\sigma^2(1)/2\mu(1)) = N(0, \rho(1 + c_a^2)/2)
\]
and
\[
(\tilde{X}_n, \tilde{I}_n) \Rightarrow (\rho\eta, (1 - \rho)\eta) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty.
\]
Proof of Theorem 3.1. The FCLT for \( X_n(t) \) in (10) is well known, following from the two-parameter MSHT limit for the \( G/GI/\infty \) model in [7]; see also §7.3 in [6] for the treatment of \( G1/M/\cdot \) type models via martingale approach. We now justify the joint FCLT in (10). By the continuous mapping theorem and Theorem 13.4.1 of [14] for the supremum function in (8),

\[
\hat{X}_n^\uparrow \Rightarrow X^\uparrow \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.
\]

(13)

However, note that

\[
\hat{X}_n^\uparrow(t) = n^{-1/2}[X_n^\uparrow(t) - n\rho], \quad t \geq 0.
\]

(14)

Because the limit in (13) is OU, \( P(X^\uparrow(t) < \infty) = 1 \) for all \( t > 0 \). Hence, for any \( t > 0 \) and \( \epsilon > 0 \), there exists \( K \equiv K(t, \epsilon) \) such that \( P(X^\uparrow(t) > K/2) < \epsilon/2 \). Then, by (13), for all \( t \) and \( \epsilon > 0 \), there exists \( n_0 \equiv n_0(t, \epsilon) \), such that

\[
P(X_n^\uparrow(t) > n\rho + \sqrt{n}K) = P(\hat{X}_n^\uparrow(t) > K) \\
= P(\hat{X}_n^\uparrow(t) - X^\uparrow(t) > K) \\
\leq P(X^\uparrow(t) > K/2) + P(\hat{X}_n^\uparrow(t) - X^\uparrow(t) > K/2) \\
< \epsilon/2 + \epsilon/2 = \epsilon \quad \text{for all} \quad n \geq n_0.
\]

(15)

As a consequence, for each \( t > 0 \),

\[
P(X_n^\uparrow(t) > n) \to 0 \quad \text{as} \quad n \to \infty,
\]

(16)

Combining (5) and (16) yields, for each \( t > 0 \),

\[
\|\hat{I}_n + \hat{X}_n\|_t \Rightarrow 0 \quad \text{as} \quad n \to \infty,
\]

(17)

which implies the joint convergence in (10) by Theorem 11.4.7 of [14]. Then (12) is an immediate consequence.

We now consider the associated heavy-traffic limit theorem for the steady-state distributions, i.e., for \( I_n(\infty) \). For that purpose, let

\[
\bar{I}_n(\infty) \equiv I_n(\infty)/n \quad \text{and} \quad \hat{I}_n(\infty) \equiv n^{1/2}(\bar{I}_n(\infty) - (1 - \rho))
\]

(18)

be the fluid- and diffusion- scaled quantities. We should expect that limit of \( \{\hat{I}_n(\infty); n \geq 1\} \) as \( n \to \infty \) coincides with the limiting distribution of the diffusion process \( \{I(t); t \geq 0\} \) as \( t \to \infty \). But that is not automatic as an interchange of limits is involved.

A justification for the interchange of iterated limits is often difficult. Fortunately, for \( GI/M/n \) type models, the desired result can be established fairly easily as elaborated in §3 of [2]; see also the proof of Theorem 2.3 in [13] in the context of overloaded queues. The idea is to show that the sequence of normalized steady-state r.v.s \( \{\hat{I}_n(\infty); n \geq 1\} \) is tight by constructing upper and lower bounding processes.
that have proper limits as \( n \to \infty \) (see the proof of Theorem 3 in [2] for further details). The tightness guarantees that every subsequence has a convergent sub-subsequence. It remains to show that if all convergent subsequences are shown to have the same limit. Here a subsequence can serve as the sequence of initial distributions in the condition of Theorem 3. Since these initial distributions are also stationary, the limiting distribution must be a stationary distribution for the limiting OU process. In view of the fact that the OU process has a unique stationary distribution, all convergent subsequences must have the same limiting distribution. Proposition 3.1 formalizes the foregoing discussion.

**Proposition 3.1.** (QD MSHT WLLN and CLT for the steady-state idleness under work-conserving server-assignment rules) Let \( \bar{I}_n(\infty) \) and \( \hat{I}_n(\infty) \) be given by (18). Then

\[
(\bar{I}_n(\infty), \hat{I}_n(\infty)) \Rightarrow (1 - \rho, I(\infty)) \quad \text{in} \quad \mathbb{R}^2 \quad \text{as} \quad n \to \infty,
\]

where \( I(\infty) \) is the stationary distribution of the OU process \( I \) as in (10).

4. QD MSHT Limits for Idle Times with Rule LISF

The primary goal of this section is to derive the weak limit of the server idle time under the LISF server-assignment rule as \( n \) grows to infinity. We are going to show that, as \( n \to \infty \),

\[
V_n^{\text{LISF}} \Rightarrow c \quad \text{in} \quad \mathbb{R} \tag{19}
\]

for some constant value \( c \). In doing so, we obtain a stronger result that can be seen as a second-order refinement for the steady-state idle-time distribution.

Under the LISF routing policy, an idle period of a typical server starts at his or her service completion time \( \tilde{t} \) and ends when all servers that are idle at time \( \tilde{t} \) get reassigned. By Lemma 3.1, the steady-state idle time \( V_n^{\text{LISF}} \) equals, in distribution,

\[
V_n^{\text{LISF}} \overset{d}{=} \sum_{i=1}^{I_n(\infty)} U_{n,i} \tag{20}
\]

where we recall that the first element \( U_{n,1} \) follows the stationary-excess distribution \( G_{\infty}^e(x) \) and \( I_n(\infty) \) represents the number of idle servers in steady state.

**Theorem 4.1.** (QD MSHT WLLN and CLT for the steady-state idle time with rule LISF) For each \( n \geq 1 \), define

\[
\hat{V}_n^{\text{LISF}} \equiv n^{1/2}(V_n^{\text{LISF}} - (1 - \rho)/\rho), \tag{21}
\]

where \( V_n^{\text{LISF}} \) represents the steady-state server idle time under the LISF policy. Then

\[
\hat{V}_n^{\text{LISF}} \Rightarrow N\left(0, \left(c^2_n(1 - \rho/2) + \rho/2\right)/\rho^2\right) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty. \tag{22}
\]

Besides, (19) holds with \( c = (1 - \rho)/\rho \).
Proof of Theorem 4.1. The proof is similar to the proof of Theorem 6.4 in [11]. But we need Lemma 3.1 to treat non-exponential inter-arrival times. First, some algebraic manipulation of the equation (21) yields

\[
\hat{V}_{n}^{\text{LISF}} = \sqrt{n} \left( \sum_{i=1}^{[nt]} \frac{U_{n,i}}{\rho} - \frac{I_{n}(\infty)}{\rho} \right) + \sqrt{n} \left( \frac{\bar{I}_{n}(\infty) - (1 - \rho)}{\rho} \right) / \rho
\]

(23)

We start with the first component. First, the joint convergence in (3) gives us

\[
\sqrt{n} \left( \sum_{i=1}^{[nt]} \frac{U_{n,i} - t/\rho}{\rho} \right) \Rightarrow \rho^{-1} c_{a} W(t) \quad \text{in } D \quad \text{as } n \to \infty.
\]

(24)

Then use Proposition 3.1 and the continuous mapping theorem (CMT) with composition (see, e.g., Proposition 13.2.1 of [14]) to obtain

\[
\sqrt{n} \left( \sum_{i=1}^{[nt]} \frac{U_{n,i} - \bar{I}_{n}(\infty)}{\rho} \right) \Rightarrow \rho^{-1} c_{a} W(1 - \rho) \quad \text{in } \mathbb{R} \quad \text{as } n \to \infty,
\]

yielding the limit for the first part.

It is important to note that the two components on the right side of the second equality in (23) are statistically independent. This is because the inter-arrival times \{\{U_{n,i}; i \geq 1\}\} are realized after the departure epoch (service completion time) \(\hat{t}\). We can therefore apply the converging-together criteria as in Theorem 11.4.4 in [14] to conclude

\[
\left( \sqrt{n} \left( \sum_{i=1}^{[nt]} \frac{U_{n,i} - \bar{I}_{n}(\infty)}{\rho} \right), \frac{\bar{I}_{n}(\infty)}{\rho} \right) \Rightarrow \left( \rho^{-1} c_{a} W(1 - \rho), I(\infty)/\rho \right) \quad \text{in } \mathbb{R}^2
\]

(25)

as \(n \to \infty\). Finally, an application of the CMT to (25) gives

\[
\hat{V}_{n}^{\text{LISF}} \Rightarrow \rho^{-1} c_{a} W(1 - \rho) + I(\infty)/\rho \overset{d}{=} N(0, c_{a}^2 (1 - \rho)/\rho^2) + N(0, (c_{a}^2 + 1)/(2\rho))
\]

(26)

as \(n \to \infty\). The sum of two normal r.v.s in (26) simplifies, yielding the desired result in (22). Then WLLN limit in (19) holds as an immediate consequence. \(\blacksquare\)

Remark 4.1. (Approximation of steady-state idle times) The MSHT limit provides support for approximating the idle period by a truncated normal distribution. Indeed, for the nth stochastic model we can use (22) in Theorem 4.1 to obtain

\[
V_{n}^{\text{LISF}} \overset{d}{=} \left[ (1 - \rho)/\rho + n^{-1/2} \xi_{M} \right]^+.
\]

(27)
where
\[ \xi_M \overset{d}{=} N\left(0, \left(\frac{c_2^2(1 - \rho/2) + \rho/2}{\rho^2}\right)\right). \] (28)

We use Figure 1 to illustrate the accuracy of the approximation by displaying the histogram and the probability density curve (in blue) on the same plot. We also examined the case of a mean-\(\rho^{-1}\) hyper-exponential (H2) distribution with \(c_2^2 = 2\) and balanced means, as in §3.1 of [12]. These plots show that the MSHT approximation derived above is effective.

Remark 4.2. (Engineering approximation with general service times) The proof of Theorem 4.1 requires \(\{I_n(\infty); n \geq 1\}\) to satisfy the the WLLN and the CLT, which in turn, is obtained by interchanging the iterated limits. For sequence of GI/G/n systems in the QD MSHT regime, it is not impossible to have the limit of the steady-state distributions be equal in distribution to the steady state of the limiting process; see, e.g., [1]. Here we do not establish such a result for the GI/G/n model. Nevertheless, we shall provide an approximating formula for the idle time under the LISF policy. The approximation draws upon the process limits in [7] for a GI/G/\(\infty\) queue. In particular, we use Theorem 4.1 in [7] to get
\[ V_n^{LISF} \overset{d}{=} \left[ (1 - \rho)/\rho + n^{-1/2}\xi_G \right]^+ \]

where
\[ \xi_G \overset{d}{=} N\left(0, \frac{c_2^2(1 - \rho)/\rho^2 + 1/\rho + \rho^{-1}(c_2^2 - 1)}{\rho} \int_0^\infty G^c(u)^2du\right). \] (29)

Note that (29) reduces to (28) when \(G^c(x) \equiv 1 - G(x) = e^{-x}\).

5. QD MSHT Limits for Idle Times with Rule RR

Let \(V_n^{RR}\) denote the steady-state server idle time under the RR policy for the \(n\)th stochastic model. At crux of the analysis is an argument embedding the idle period \(V_n^{RR}\) of a typical server into a counting process \(\tilde{A}_n\) such that this idle period coincides with the first inter-arrival time of the process \(\tilde{A}_n\). We do it for each \(n \in \mathbb{N}\). We thus obtain a sequence of point processes \(\{\tilde{A}_n; n \geq 1\}\). We are able to show that \(\{V_n^{RR}; n \geq 1\}\) converges weakly to an exponential r.v. with rate parameter \(\rho/(1 - \rho)\) as \(n \to \infty\) by arguing that the sequence of processes \(\tilde{A}_n\) converges weakly to a Poisson point process with rate parameter \(\rho/(1 - \rho)\).

For each \(n \in \mathbb{N}\), we construct the counting process \(\tilde{A}_n\) in the following way. The counting process \(\tilde{A}_n\) is initialized at a departure epoch \(t_0\). Without loss of generality, assume that \(t_0 = 0\). Suppose server \(k\) is the one that completes a service at \(t_0\). Then server \(k\) becomes the flagged (marked) idle server until server \(k\) gets reassigned under the RR policy. Once the reassignment occurs, say at time \(t_1 > t_0\),
we unflag (remove the flag from) server $k$ and randomly choose an idle server (if any), say server $k'$, from the pool of idle servers to be the newly flagged idle server. At this time, $\tilde{A}_n$ increments by one. In the same fashion, server $k'$ would pass the flag randomly to another idle server (if any) at the time server $k'$ is reassigned by the RR rule. We thus obtain a sequence of flagged idle servers over time and in the meantime create a process $\tilde{A}_n(t)$ counting the number of flagged servers having been reassigned up to time $t$. Note that if there is no other idle server at the moment when the flagged server starts a new service, the flag is passed to the next server that completes a service and is about to be idle. More formally, we can spell out the auxiliary counting process $\tilde{A}_n$ through the arrival process $A_n$ and the idleness process $I_n$:

$$\tilde{A}_n(t) \overset{d}{=} 1\{I_n(0)=0\} + \sum_{k=1}^{A_n(t)} X_k$$

(30)
where \( A_n(t) \) is the total number of arrivals up to time \( t \) and

\[
X_k = \begin{cases} 
\mathcal{B}_k(1/I_n(t_k)) & \text{if } I_n(t_k) > 0 \\
0 & \text{if } I_n(t_k) = 0 
\end{cases}
\]  

(31)

where we recall that \( \mathcal{B}_k(p_k) \) denotes a Bernoulli r.v. with parameter \( p_k \) and \( t_k \) denotes the \( k \)th arrival time. It is important to note that the sequence of Bernoulli r.v.s are conditionally independent given the realizations \( \{I_n(t_k); k \geq 1\} \) and so are the sequence of r.v.s \( \{X_k; k \geq 1\} \).

With this preparation, we are able to represent the idle time of a typical server under the RR assignment rule as the first inter-arrival time of the process \( \tilde{A}_n \).

**Theorem 5.1.** (Poisson limit for the auxiliary point process \( \tilde{A}_n \)) Let \( \tilde{A}_n \) be constructed as above for the \( n \)th GI/M/n model. Then

\[
\tilde{A}_n(\cdot) \Rightarrow \mathcal{P}(\rho \cdot / (1 - \rho)) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty,
\]

(32)

where \( \mathcal{P}(\cdot) \) represents a standard Poisson point process.

**Proof of Theorem 5.1.** In our setting, weak convergence of the processes with non-decreasing sample paths to a Poisson point process in \( \mathcal{D} \) is equivalent to convergence of all finite-dimensional distributions; see VI.3.37 of [4]. That is, for any integer \( m \), any \( m \)-tuple of disjoint subintervals \( \{(u_{i,1}, u_{i,2}); 1 \leq i \leq m\} \) and any \( m \)-tuple of nonnegative integers \( \{j_i; 1 \leq i \leq m\} \), one needs to show that

\[
P \left( \tilde{A}_n(u_{i,1}, u_{i,2}) = j_i; 1 \leq i \leq m \right) \to \prod_{i=1}^{m} \frac{e^{-\rho \Delta u_i/(1 - \rho)} (\rho \Delta u_i/(1 - \rho))^{j_i}}{j_i!} \quad \text{as} \quad n \to \infty
\]

(33)

where we have defined \( \tilde{A}_n(u_{i,1}, u_{i,2}) \equiv \tilde{A}_n(u_{i,2}) - \tilde{A}_n(u_{i,1}) \) and \( \Delta u_i \equiv u_{i,2} - u_{i,1} \).

Here we establish (33) via integral transform. Let

\[
\psi_n(s) \equiv E \left[ \exp \left\{ -\sum_{i=1}^{m} s_i \tilde{A}_n(u_{i,1}, u_{i,2}) \right\} \right]
\]

be the Laplace transform of the \( m \)-tuple of r.v.s \( \{\tilde{A}_n(u_{i,1}, u_{i,2}); 1 \leq i \leq m\} \). Then the goal is to show that

\[
\psi_n(s) \to \prod_{i=1}^{m} \exp \left\{ -\frac{\rho \Delta u_i (1 - e^{-s_i})}{(1 - \rho)} \right\} \quad \text{as} \quad n \to \infty.
\]

(34)

First we choose \( t \) such that \( (u_{i,1}, u_{i,2}] \subset [0, t] \) for \( 1 \leq i \leq m \). Now for an arbitrarily chosen \( \epsilon > 0 \), choose \( n \) large enough such that

\[
P \left( \| I_n - (1 - \rho) \eta \|_t > \epsilon \right) < \epsilon/2 \quad \text{and} \quad P \left( |\tilde{A}_n(t) - pt| > \epsilon \right) < \epsilon/2
\]
where $\| \cdot \|_t$ denotes the supreme norm on the interval $[0, t]$. Let $\mathcal{E}_{n,t}$ denote the event in the brackets

$$\mathcal{E}_{n,t} \equiv \{ \| \bar{I}_n - (1 - \rho) \eta \|_t > \epsilon \} \cup \{ | \bar{A}_n(t) - \rho t | > \epsilon \}. \quad (35)$$

Note that $P(\mathcal{E}_{n,t}) < \epsilon$ and

$$\mathcal{E}_{n,t}^c \implies \begin{cases} (1 - \rho) - \epsilon \leq \bar{I}_n(t_k) \leq (1 - \rho) + \epsilon \quad \text{for all} \quad k \leq A_n(t) \quad \text{and} \\ \rho s - \epsilon \leq \bar{A}_n(s) \leq \rho s + \epsilon \quad \text{for all} \quad s \leq t. \end{cases} \quad (36)$$

We can then decompose the Laplace transform in the following way

$$\psi_n(s) = E \left[ \exp \left\{ - \sum_{i=1}^{m} s_i \bar{A}_n(u_{i,1}, u_{i,2}) \right\} 1_{\mathcal{E}_{n,t}}^c \right] + E \left[ \exp \left\{ - \sum_{i=1}^{m} s_i \bar{A}_n(u_{i,1}, u_{i,2}) \right\} 1_{\mathcal{E}_{n,t}} \right] \equiv C_{n,t,1} + C_{n,t,2}. \quad (37)$$

Using (30) and (36), we obtain bounds for the first component

$$\prod_{i=1}^{m} \left[ 1 + \frac{e^{-s_i} - 1}{(1 - \rho - \epsilon)n} \right]^{(\rho \Delta u_i + \epsilon)n} \leq C_{n,t,1} \leq \prod_{i=1}^{m} \left[ 1 + \frac{e^{-s_i} - 1}{(1 - \rho + \epsilon)n} \right]^{(\rho \Delta u_i - \epsilon)n}. \quad (38)$$

Based on (38), one can further deduce that

$$\prod_{i=1}^{m} \exp \left\{ - \frac{(\rho \Delta u_i + \epsilon)(1 - e^{-s_i})}{1 - \rho - \epsilon} \right\} - \epsilon'_{1} \leq C_{n,t,1} \leq \prod_{i=1}^{m} \exp \left\{ - \frac{(\rho \Delta u_i - \epsilon)(1 - e^{-s_i})}{1 - \rho + \epsilon} \right\} + \epsilon'_{2}. \quad (39)$$

where $\epsilon'_{1}$ and $\epsilon'_{2}$ are due to the approximation error. For the second component, we have

$$0 < C_{n,t,2} \leq E \left[ 1_{\mathcal{E}_{n,t}} \right] < \epsilon. \quad (40)$$

taking into account the fact that $\{ s_i; 1 \leq i \leq m \}$ are all nonnegative real numbers. Combining (37), (39) and (40) yields

$$\prod_{i=1}^{m} \exp \left\{ - \frac{(\rho \Delta u_i + \epsilon)(1 - e^{-s_i})}{1 - \rho - \epsilon} \right\} - \epsilon'_{1} < \psi_n(s) < \prod_{i=1}^{m} \exp \left\{ - \frac{(\rho \Delta u_i - \epsilon)(1 - e^{-s_i})}{1 - \rho + \epsilon} \right\} + \epsilon'_{2} + \epsilon.$$

Because $\epsilon$ is arbitrary and $\epsilon'_{1}$ and $\epsilon'_{2}$ can be arbitrarily small by choosing $n$ large enough, we obtain (34) and hence (33). We thus accomplish what we set out to do by showing that the sequence of auxiliary counting processes $\bar{A}_n$ converges weakly to a Poisson point process with rate parameter $\rho/(1 - \rho)$ as $n \to \infty$. \hfill \blacksquare
Figure 2: Histograms estimated by simulation (with the atom at 0 removed) of the steady-state idle-time distribution with RR for the \( M/M/n \) model and the \( H_2/M/n \) with \( \mu = 1 \) and \( \rho = 0.9 \).

**Remark 5.1.** The assumption of Theorem 4.1 allows a general sequence of arrival processes, but they are required to satisfy an FCLT because the primary concern there is establishing the MSHT CLT for the steady-state idle time. In Theorem 5.1, the FCLT condition can be weakened to having only an FWLLN because the proof of 5.1 only requires the MSHT FWLLN of the arrival processes and the MSHT FWLLN of the idleness processes; see (35).

By our construction of the auxiliary counting process \( \tilde{A}_n \), the idle time of a typical server \( V_{n_{RR}} \) can be seen as the first inter-arrival time of the process \( \tilde{A}_n \). The result of Theorem 5.1 necessarily implies the convergence of \( \{V_{n_{RR}}; n \geq 1\} \) to an exponential distribution.

**Corollary 5.1.** (QD MSHT WLLN for the steady-state idle time with rule RR) Let \( V_{n_{RR}} \) denote the steady-state server idle time under the RR policy for the \( n \)th stochastic model with \( \rho_n \equiv \rho \). Then

\[
V_{n_{RR}} \Rightarrow \mathcal{M}(\rho/(1-\rho)) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty,
\]

(41)

where \( \mathcal{M}(\rho/(1-\rho)) \) represents an exponential r.v. with rate parameter \( \rho/(1-\rho) \).
In Figure 2, we show the histogram of server idle periods for both the $M/M/n$ model and the $H_2/M/n$ model with the blue curve showing the probability density of an exponential distribution with rate $\rho/(1 - \rho) = 9$. Consistent with our analysis, the idle time between successive busy periods can be well approximated by an exponential distribution.

Acknowledgment

The author is indebted to the anonymous referee(s) and the associate editor for their careful review and suggestions. The author would like to thank Dr. Ward Whitt for pointing out the problem of interchange of iterated limits and suggesting references on arrivals that see time averages. Research support was received from NSF (CMMI 1634133).

References


