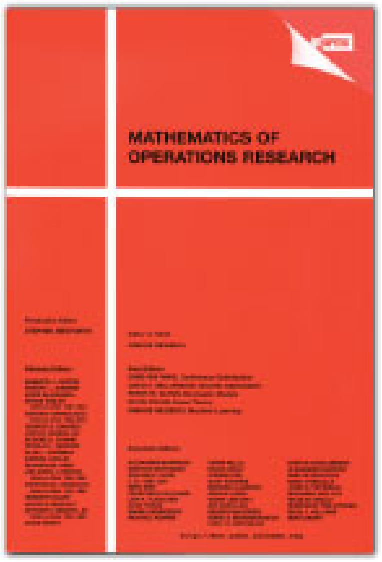

Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

An #-Nash Equilibrium with High Probability for Strategic Customers in Heavy Traffic

Rami Atar, Subhamay Saha

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–21
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

https://doi.org/10.1287/moor.2016.0820

An ε-Nash Equilibrium with High Probability for Strategic Customers in Heavy Traffic

Rami Atar
Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel,
atar@ee.technion.ac.il

Subhamay Saha
Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India,
subhamay585@gmail.com

A multiclass queue with many servers is considered, where customers make a join-or-leave decision upon arrival based on queue length information, without knowing the state of other queues. A game theoretic formulation is proposed and analyzed, that takes advantage of a phenomenon unique to heavy traffic regimes, namely, Reiman's snaphshot principle, by which waiting times are predicted with high precision by the information available upon arrival. The payoff considered is given as a *random variable*, which depends on the customer's decision, accounting for waiting time in the queue and penalty for leaving. The notion of an equilibrium is only meaningful in an asymptotic framework, which is taken here to be the Halfin-Whitt heavy traffic regime. The main result is the identification of an ε-Nash equilibrium with probability approaching 1. On the way to proving this result, new diffusion limit results for systems with finite buffers are obtained.

Keywords: Halfin-Whitt heavy traffic regime; Reiman's snapshot principle; strategic customers; ε-Nash equilibrium with high probability
MSC2000 subject classification: Primary: 60F17, 60J60, 60K25, 91A06, 93E20
OR/MS subject classification: games/group decisions: stochastic; queue: diffusion models
History: Received February 9, 2016; revised July 27, 2016. Published online in *Articles in Advance* December 14, 2016.

1. Introduction. Equilibrium behavior of strategic customers in queueing systems has been the subject of great interest since the work of Naor [16] (see the book by Hassin and Haviv [13] for a survey), and has been a particularly active research area in recent years. As far as heavy traffic analysis is concerned, not a great deal of attention has been drawn to game theoretic aspects such as the asymptotic study of Nash equilibria, unlike, for example, control theoretic treatment, to which much work has been devoted. In this paper we propose and analyse a game theoretic formulation of strategic customers in a multiclass queueing system that takes advantage of phenomena specific to heavy traffic regimes. The formulation is based on associating with each customer a payoff that reflects the customer's actual waiting time rather than its expectation. The notion of equilibrium addressed, namely, an ε-Nash equilibrium with high probability (w.h.p.), becomes meaningful only as scaling limits are taken. An additional aspect that is unique to this setting regards the relatively small level of information required for the players. In game theoretic analysis of queueing models, it is usually the case that when partial information of the system's state is available to the player, the unobservable states are assumed to be in stationarity. In the setting of this paper, customers are aware of the queue length of their own class (as well as the system parameters, specifically the rate of arrival, or at least a first order approximation thereof) but not those of other classes. However, stationarity assumptions are not required. Moreover, while this paper assumes that the system operates under a scheduling policy of one out of two specific types (see below), it should be noted that the scheduling policy is not known to the customers when they make their decisions.

Motivating applications include the following. Consider a call center where customers call to get various kinds of services Aksin et al. [1]. There are several classes of customers associated with service requirements of different types. Customers that call are notified how many customers of their class are waiting in line, and naturally, they are not aware of the scheduling policy. They then decide whether to stay in line or leave based on this piece of information. Another motivation comes from on-demand usage of resources in cloud computing. The cloud service provider offers computational resources to different classes of customers, at possibly different rates. Delay is often a significant factor taken into account side by side with the cost of usage, and usage decisions are made by the customers based on this information. For more on the latter application see Atar et al. [6] and references therein; this application has also been the main motivation for a control problem formulation in heavy traffic in Atar and Shifrin [5].

The model considered consists of a fixed number of customer classes, that differ in their service rates, and n identical, exponential servers that work in parallel. Upon arrival of a class-i customer, the ith queue length is revealed and, based on this information, he decides whether to join or leave. Accordingly, the customer's payoff is given by $h_i(\mathrm{WT})$ or r_i, respectively, where $h_i\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a given function, WT is the time the particular customer will wait in line before being admitted into service, and $r_i \in (0, \infty)$ is a cost for not receiving service (both h_i and r_i depend on the class, i). Because WT is a random variable the value of which is not known at the time of decision, the payoff is in fact a *random* function of the customer's decision, as well as other customers' decisions. Establishing an equilibrium based on random payoffs is made possible thanks to the consideration of the game in an asymptotic regime.

The asymptotic setting considered is the Halfin-Whitt (HW) heavy traffic regime (Halfin and Whitt [12]), in which the number of servers, n, grows without bound, and the arrival processes accelerate accordingly so as to keep the system critically loaded. The customers considered are those that arrive during a fixed, finite time interval. Thus the number of participating players also grows without bound.

The specific feature due to which a random payoff formulation is tractable in this regime (and potentially in other heavy traffic regimes), is *Reiman's snapshot principle* (RSP) (Reiman [17]), which, when specialized to the present setting, states that the waiting time a customer will experience is asymptotically equal to the queue length at the time of arrival divided by the overall rate at which customers from the class are served (see Section 5 for a precise statement). While this principle has been proved in a number of settings, it does not always hold (as explained in Remark 2.4). In particular, its validity depends on the scheduling policy. Our equilibrium results, that are based on this principle, can therefore only be obtained under some assumptions on the scheduling. We address this aspect by considering two families of scheduling policies under which, as we show, RSP holds: *fixed priority* (FP), where a server that becomes available will always pick the customer at the head of the line of the buffer with least index among nonempty buffers, and *serve the longest queue* (SLQ), where the buffer with longest queue is picked. Our main result shows that if all customers adopt a strategy that uses RSP as a prediction for the waiting time, an ε-Nash equilibrium w.h.p. is obtained.

On the way to proving the main result we prove new diffusion limit results for the above two policies, for systems in which customers join only when the queue length of the corresponding buffer is below a threshold, an element that can otherwise be described by finite buffers. A nonstandard aspect of the diffusion scale analysis required toward proving the main result is that one must take into account different behaviours of customers, so as to allow for scenarios where one of the customers deviates from the strategy that is to be shown to lead to an equilibrium. In particular, properties on which the proof is based, such as the C-tightness of some of the processes involved, are proved to hold uniformly over such scenarios.

Other works where a game theoretic equilibrium is considered in conjunction with heavy traffic analysis of a queueing model are Allon and Gurvich [2], Gopalakrishnan et al. [9], and Zhan and Ward [18]. The latter two works do not address a Nash equilibrium w.h.p., but rather provide an asymptotic analysis *subsequently* to establishing the prelimit equilibrium, and the notion of equilibrium is based on deterministic, steady state payoffs, and complete state information. Both works study *servers* that act strategically. Specifically, in Gopalakrishnan et al. [9], servers choose their service rate in order to optimize a trade-off between an effort cost and value of idleness. The focus of Gopalakrishnan et al. [9] is on the study of the implications of such strategic behaviour on staffing and routing, as the size of the system becomes large. In Zhan and Ward [18], servers are paid based on their service speed as well as service quality, and each choose their own service speed in order to maximize expected utility. The work of Allon and Gurvich [2] addresses a notion of an ε-Nash equilibrium for prelimit models and studies their convergence in fluid and diffusion scales to equilibria characterized by certain fluid and diffusion games, respectively. The notion addressed in Allon and Gurvich [2] is different from ours, where the main difference lies in the fact that the game is formulated at steady state, and thus dynamics of the stochastic processes involved do not show up. An additional significant difference is that the number of players in the game considered in Allon and Gurvich [2] is fixed.

As far as our convergence results and RSP are concerned, the closest work is by Gurvich and Whitt [11], where a parallel server system, with multiple classes as well as multiple server pools, is considered in the HW regime, under the *fixed queue and idleness ratio* policy. This policy aims at keeping queue lengths as well as idleness levels at the different server pools at predetermined fixed ratios. When specialized to the case of a single server pool, and equal queue length ratios, this setting is similar to one of the two settings studied in this paper, namely, SLQ. There are, however, two important differences in terms of the technical treatment. First, as already mentioned, the estimates required to deduce the main result must be uniform over scenarios. A second difference is that finite buffers are not covered by Gurvich and Whitt [11]. Although it may seem that this aspect requires only simple adaptations to cover convergence results, this is not the case. In fact, diffusion limits do not always exist under our assumptions, as is the case under SLQ if the buffers are of equal size (this issue is developed further in Atar and Saha [4]). Hence, considerations beyond the infinite buffer model are necessarily significant here.

As an additional small sample of recent work on strategic behaviour in queueing systems, we mention Guo and Hassin [10], that analyse the response of customers to shutting down service when the queue is empty, and resuming when the queue length exceeds a threshold; and Manou et al. [15], that studies a natural model for the behaviour of customers in a transportation station. In both cases, Nash equilibria are determined under various assumptions on the level of information.

We use the following notation. For $a, b \in \mathbb{R}$, the maximum (resp., minimum) is denoted by $a \vee b$ (resp., $a \wedge b$), and $a^+ = a \vee 0$, $a^- = (-a) \vee 0$. For $x, y \in \mathbb{R}^k$ (k a positive integer), $x \cdot y$ and $\|x\|$ denote the usual scalar product and ℓ_2 norm, respectively. Write $\{\mathbf{e}_i\}$, $i = 1, \dots, k$ for the standard basis in $\mathbb{R}^k$ and $\mathbf{1}$ for $\sum_{i=1}^k \mathbf{e}_i$. Denote $\mathbb{R}_+ = [0, \infty)$. For $f\colon \mathbb{R}_+ \to \mathbb{R}^k$, $\|f\|_T = \sup_{t\in[0,T]} \|f(t)\|$, and, for $\theta > 0$,

$$w_T(f, \theta) = \sup_{0\le s<u\le s+\theta\le T} \|f_u - f_s\|.$$

For a Polish space $\mathcal{S}$, let $\mathbb{C}_{\mathcal{S}}([0, T])$ and $\mathbb{D}_{\mathcal{S}}([0, T])$ denote the set of continuous and, respectively, càdlàg functions $[0, T] \to \mathcal{S}$. Write $\mathbb{C}_{\mathcal{S}}$ and $\mathbb{D}_{\mathcal{S}}$ for the case where $[0, T]$ is replaced by $\mathbb{R}_+$. Endow $\mathbb{D}_{\mathcal{S}}$ with the Skorohod J_1 topology. Write $X_n \Rightarrow X$ for convergence in distribution. A sequence of processes X_n with sample paths in $\mathbb{D}_{\mathcal{S}}$ is said to be *C-tight* if it is tight and every subsequential limit has, with probability 1, sample paths in $\mathbb{C}_{\mathcal{S}}$. For a sequence of processes ξ^n, $n \in \mathbb{N}$, with sample paths in $\mathbb{D}_{\mathbb{R}^k}$, C-tightness is characterized (see Jacod and Shiryaev [14, VI.3.26]) by the following:

C1. The sequence of random variables $\|\xi^n\|_T$ is tight for every fixed $T < \infty$.

C2. For every $T < \infty$, $\varepsilon > 0$, and $\eta > 0$ there exist n_0 and $\theta > 0$ such that

$$n \ge n_0 \quad \text{implies} \quad \mathbb{P}(w_T(\xi^n, \theta) > \eta) < \varepsilon.$$

For a positive integer k, $m \in \mathbb{R}^k$ and a symmetric, positive matrix $A \in \mathbb{R}^{k\times k}$, *an* (m, A)*-Brownian motion* (BM) is a k-dimensional BM starting from zero, having drift m and infinitesimal covariance matrix A.

This paper is organized as follows. The model and the equilibrium result appear in Section 2. Sections 3 and 4 analyse the behavior of the system under FP and SLQ, respectively, and along the way also obtain diffusion limit results, that may be interesting by their own right. Section 5 addresses RSP in these two settings and proves the main result.

2. Model and main result.

We start by introducing the probabilistic model and the HW scaling. Then we provide the game theoretic setting and state the main result.

A sequence of queueing models is considered, indexed by $n \in \mathbb{N}$. The nth system has N buffers and n identical servers. Customers from N distinct classes arrive at the system and, upon arrival, each customer is informed about the queue length at the buffer that corresponds to its own customer class, and, based on this information only, makes a decision whether to join or leave the system. If a customer of class i decides to join, he goes directly for service on the event that any of the servers is available, and otherwise he is queued in buffer i. As far as the service policy is concerned, we consider FP and SLQ (that is, however, unknown to the customers). In the first case, the servers serve according to the rule given by $1 > 2 > \cdots > N$. Thus, when a server becomes available, it admits into service a customer in the buffer with highest priority (that is, least index) among all buffers that are nonempty at that instant. Under SLQ, the buffer that currently has the most customers receives highest priority (where ties are broken arbitrarily). At each buffer, the customers are always taken from the head of the line. We assume the nonidling condition, that is, that no server will idle as long as any customers are in the queue.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, on which all the random variables (r.v.s) introduced below are to be defined. The arrivals in each class occur according to independent renewal processes. Let parameters $\lambda_i^n > 0$, $i \in \{1, 2, \dots, N\}$, be given, representing the mean interarrival times of class-i customers in the nth system. Let $\{\mathrm{IA}_i(l)\colon l \in \mathbb{N}\}_i$ be independent sequences of strictly positive i.i.d. r.v.s with mean 1 and variance $C^2_{\mathrm{IA}_i}$. Let

$$E_i^n(t) = \sup\left\{ l \ge 0\colon \sum_{k=1}^{l} \frac{\mathrm{IA}_i(k)}{\lambda_i^n} \le t \right\}, \qquad t \ge 0. \tag{1}$$

Then E_i^n counts the number of class-i arrivals up to time t. The parameters λ_i^n satisfy

$$\lambda_i^n = n\lambda_i + \sqrt{n}\hat{\lambda}_i + o(\sqrt{n}), \tag{2}$$

where $\lambda_i > 0$ and $\hat{\lambda}_i \in \mathbb{R}$ are fixed. The service times of class-i customers are assumed to be exponential with mean μ_i. The potential service processes, denoted by $\{S_i\}_{i=1,2,\dots,N}$, are thus assumed to comprise a collection of N mutually independent Poisson process, with rates μ_i, $i = 1, 2, \dots, N$, respectively. They are assumed to have right-continuous sample paths. While the arrival rates are accelerated with n, the individual service rates are not. However, the capacity of the service pool grows due to the increase of the number of servers, n. The resulting traffic intensity is thus asymptotically given by $\sum_i \rho_i$, where $\rho_i = \lambda_i/\mu_i$. We will assume the following critical load condition:

$$\sum_i \rho_i = 1. \tag{3}$$

The initial conditions,

$$Q^n(0) = (Q_1^n(0), Q_2^n(0), \ldots, Q_N^n(0)), \qquad \Psi^n(0) = (\Psi_1^n(0), \Psi_2^n(0), \ldots, \Psi_N^n(0)),$$

are $\mathbb{Z}_+^N$-valued r.v.s representing the number of customers initially in the buffers and in service, respectively. It is assumed that the initial configuration satisfies $\mathbf{1} \cdot Q^n(0) > 0$ implies $\mathbf{1} \cdot \Psi^n(0) = n$, reflecting the nonidling condition.

For each n, the three objects

$$\{E_i^n\}_i, \{S_i\}_i, \qquad (Q^n(0), \Psi^n(0)) \tag{4}$$

are assumed to be mutually independent. The triplet (4) will be referred to as the *stochastic primitives* of the model. All r.v.s introduced below, describing the system dynamics, will be given as functions of the stochastic primitives and of the collection of decisions taken by the strategic customers.

Thus, before describing the system dynamics, we introduce the notation for the decision variables. The customers initially in the system do not participate in the game formulation, and therefore in what follows, unless otherwise stated, the term *customer* will refer to those customers that arrive after time zero. A customer will be identified by a pair (i, j), where $i \in \{1, 2, \ldots, N\}$ is its class, and $j \in \mathbb{N}$ is its serial number in order of arrival. The collection of decision variables $\delta = \{\delta_{ij}\colon i \in \{1, 2, \ldots, N\}, j \in \mathbb{N}\}$, where $\delta_{ij} \in \{0, 1\}$, specifies the decision of each of the customers. Having $\delta_{ij} = 1$ (resp., 0) specifies that the jth class-i customer to arrive decides to join (resp., leave) the system. Let

$$J_i^n(t) = \sum_{j=1}^{E_i^n(t)} \delta_{ij}, \qquad R_i^n(t) = \sum_{j=1}^{E_i^n(t)} (1 - \delta_{ij}) \tag{5}$$

denote counting processes for joining and reneging customers. Let $Q_i^n(t)$ be the number of class-i customers waiting at the ith buffer at time t, and let $B_i^n(t)$ be the number of class-i customers routed to the service pool by that time. Then we have

$$Q_i^n(t) = Q_i^n(0) + E_i^n(t) - B_i^n(t) - R_i^n(t). \tag{6}$$

Let $\Psi_i^n(t)$ denote the number of class-i customers in service at time t. Then

$$\Psi_i^n(t) = \Psi_i^n(0) + B_i^n(t) - D_i^n(t), \tag{7}$$

where the departure process D_i^n counts the number of completed services of class-i jobs since time 0 (including initial customers). It is assumed that the departure process is given, in terms of the potential service process, by

$$D_i^n(t) = S_i\biggl(\int_0^t \Psi_i^n(u)\,du\biggr). \tag{8}$$

The nonidling condition is expressed by requiring

$$\text{for every } t, \qquad \mathbf{1} \cdot Q^n(t) > 0 \quad \text{implies} \quad \mathbf{1} \cdot \Psi^n(t) = n. \tag{9}$$

Under the FP policy we have

$$\int_{[0,\infty)} \sum_{k=1}^{i-1} Q_k^n(t)\,dB_i^n(t) = 0, \qquad i = 2, 3, \ldots, N. \tag{10}$$

And under SLQ, a server that becomes available at time t chooses class i_0, where $i_0 \in \arg\max_i Q_i^n$ (where ties are broken in an arbitrary, but concrete way), namely,

$$\int_{[0,\infty)} \mathbb{I}_{\{Q_i^n(t-) < \max_k Q_k^n(t-)\}}\,dB_i^n(t) = 0, \qquad i = 1, 2, \ldots, N. \tag{11}$$

The collection of Equations (5)–(9) and either (10) or (11), along with the primitives and the decision variables δ, uniquely define the processes Q^n, X^n, Ψ^n, B^n, and D^n under each of the two policies. Note that these processes are right-continuous by construction.

Now let

$$\mathrm{JT}_i^n(t) = \inf\{s \ge t \colon J_i^n(s) > J_i^n(t-)\}, \tag{12}$$

(where, by convention, $\mathrm{JT}_i^n(0-) = 0$), represent the time of arrival of the first class-i customer to join the system at or after time t. Let also

$$\mathrm{RT}_i^n(t) = \inf\big\{s > t \colon B_i^n(s) \ge B_i^n(\mathrm{JT}_i^n(t)) + Q_i^n(\mathrm{JT}_i^n(t))\big\}. \tag{13}$$

Then $\mathrm{RT}_i^n(t)$ gives the time when the customer joining at $\mathrm{JT}_i^n(t)$ enters service. The time that particular customer waits in the queue is then given by

$$\mathrm{WT}_i^n(t) = \mathrm{RT}_i^n(t) - \mathrm{JT}_i^n(t). \tag{14}$$

Note that, as a consequence,

$$Q_i^n(\mathrm{JT}_i^n(t)) = B_i^n(\mathrm{JT}_i^n(t) + \mathrm{WT}_i^n(t)) - B_i^n(\mathrm{JT}_i^n(t)). \tag{15}$$

(JT, RT, WT, as well as AT defined below, are mnemonics for joining time, routing time, waiting time, and arrival time.) We shall also need notation of arrival time and waiting time of the jth class-i customer. These are obtained as follows:

$$\mathrm{AT}_{ij}^n = \mathrm{inv}E_i^n(j) = \inf\{t \ge 0 \colon E_i^n(t) \ge j\},$$
$$\mathrm{WT}_{ij}^n = \mathrm{WT}_i^n(\mathrm{AT}_{ij}^n).$$

Note that while WT_{ij}^n is well defined for all (i, j), it only gives the waiting time for those customers (i, j) that have actually joined the system; this concept is indeed meaningless for the reneging customers. Scaled versions of the main stochastic processes introduced above are defined as follows:

$$\begin{gathered}
\bar{\Psi}_i^n(t) = \frac{\Psi_i^n(t)}{n}, \\
\hat{Q}_i^n(t) = \frac{Q_i^n(t)}{\sqrt{n}}, \qquad \hat{B}_i^n(t) = \frac{B_i^n(t) - n\lambda_i t}{\sqrt{n}}, \\
\hat{R}_i^n(t) = \frac{R_i^n(t)}{\sqrt{n}}, \qquad \hat{S}_i^n(t) = \frac{S_i(nt) - n\mu_i t}{\sqrt{n}}, \\
\hat{D}_i^n(t) = \hat{S}_i^n\left(\int_0^t \bar{\Psi}_i^n(u)\,du\right), \qquad \hat{E}_i^n(t) = \frac{E_i^n(t) - \lambda_i^n t}{\sqrt{n}}, \qquad \hat{\Psi}_i^n(t) = \frac{\Psi_i^n(t) - \rho_i n}{\sqrt{n}}.
\end{gathered} \tag{16}$$

Also define,

$$\widehat{\mathrm{WT}}_i^n(t) = \sqrt{n}\mathrm{WT}_i^n(t), \qquad \widehat{\mathrm{WT}}_{ij}^n = \sqrt{n}\mathrm{WT}_{ij}^n. \tag{17}$$

It is assumed that the scaled initial condition converges in distribution:

$$(\hat{Q}^n(0), \hat{\Psi}^n(0)) \Rightarrow (0, \Psi(0)), \tag{18}$$

where $\Psi(0)$ is an $\mathbb{R}^N$-valued r.v. with $\sum_i \Psi_i(0) \le 0$.

This completes the description of the stochastic processes of interest. We denote the collection of processes, that we will sometimes refer to as *dynamics*, by

$$\mathscr{S}^n = \mathscr{S}^n[\delta] = (J^n, R^n, Q^n, B^n, \Psi^n, D^n, \mathrm{JT}^n, \mathrm{RT}^n, \mathrm{WT}^n),$$

where we emphasize the dependence of these processes on the decision variables δ. We will use similar notation to emphasize the dependence of each of the components of $\mathscr{S}^n$ on δ, as, for example, $Q^n[\delta]$.

Now we come to the game theoretic setting. It is described for fixed n. In the game, the dynamics described above will serve as the game's state. The game is played by the customers to arrive up to time $\bar{T}$, where $\bar{T} \in (0, \infty)$ is fixed throughout. A decision is made by each customer once the queue length of the corresponding class at the time of arrival is revealed to it. Thus for our purpose, a *strategy* is a mapping $\sigma \colon \mathbb{Z}_+ \to \{0, 1\}$. We denote the set of all such mappings by Σ. A *strategy profile* is an element of $\bar{\Sigma} := \Sigma^{\{1,2,\ldots,N\}\times\mathbb{N}}$. Let a strategy profile $\sigma = \{\sigma_{ij}\} \in \bar{\Sigma}$ be given. We say that *the game is played with the strategy profile* σ if one has

$$\begin{cases} \mathscr{S}^n = \mathscr{S}^n[\Delta^n], & \text{(specifically, } Q^n = Q^n[\Delta^n]), \\ \Delta_i^n(j) = \sigma_{ij}(Q_i^n(\mathrm{AT}_{ij}^n -)), & i \in \{1, 2, \ldots, N\},\ j \in \mathbb{N}. \end{cases} \tag{19}$$

Thus $\mathscr{S}^n$ is the dynamics resulting from having each customer (i, j) adopt the strategy σ_{ij}, and $\Delta^n_i(j)$ is a r.v. representing the action taken by customer (i, j) in that situation. An argument by induction on the times of arrival shows that the system of Equations (19) has a unique solution, and thus $\mathscr{S}^n$ and Δ^n are well-defined r.v.s. We will also need a notation for the dynamics $\mathscr{S}^n$, thus determined by (19), as a function of the strategy profile σ. We write it as $\mathscr{S}^n(\sigma)$.

We formulate the payoff for customer (i, j) by accounting for a cost associated with not receiving service (in case of reneging) and a function of the waiting time (in case of joining). To this end, we are given constants $r_i > 0$, $i \in \{1, 2, \dots, N\}$ and functions h_i: $\mathbb{R}_+ \to \mathbb{R}_+$, assumed to be continuous, strictly increasing, and to vanish at zero. For a strategy profile $\sigma = \{\sigma_{ij}\}$, denote $\sigma^{ij} = \{\sigma_{k,l}: (k, l) \neq (i, j)\}$. The payoff for customer (i, j), when the strategy profile σ is played, is given by

$$C^n_{ij}(\sigma_{ij}, \sigma^{ij}) = \begin{cases} r_i, & \Delta^n_i(j) = 0,\ \mathrm{AT}^n_{ij} \le \bar{T}, \\ h_i(\widehat{\mathrm{WT}}^n_{ij}), & \Delta^n_i(j) = 1,\ \mathrm{AT}^n_{ij} \le \bar{T}, \\ 0, & \mathrm{AT}^n_{ij} > \bar{T}. \end{cases} \tag{20}$$

Thus, according to the payoff definition, the game neglects all customers arriving after time $\bar{T}$. Note that if we let $h^n_i(x) = h_i(\sqrt{n}x)$, $x \ge 0$, then the expression $h_i(\widehat{\mathrm{WT}}^n_{ij})$ from (20) can be written in terms of the unnormalized waiting time as $h^n_i(\mathrm{WT}^n_{ij})$. Thus, for example, when h_i are linear, and given by $h_i(x) = c_i x$, $x \ge 0$, our setting corresponds to assuming that a class-i customer incurs a holding cost of $c_i\sqrt{n}$ per unit time.

For fixed n and $\varepsilon > 0$, and an event $\tilde{\Omega} \in \mathscr{F}$, a strategy profile $\sigma = \{\sigma_{ij}\}$ is said to be an *ε-Nash equilibrium on the event* $\tilde{\Omega}$ if

$$\forall (i, j), \quad \forall \tau \in \Sigma, \qquad C^n_{ij}(\sigma_{ij}, \sigma^{ij}) \le C^n_{ij}(\tau, \sigma^{ij}) + \varepsilon \tag{21}$$

holds on $\tilde{\Omega}$. A sequence of strategy profiles $\{\sigma^n\}_{n\in\mathbb{N}}$ is said to be an *ε-Nash equilibrium w.h.p.*, if there exist events $\tilde{\Omega}^n$, $n \in \mathbb{N}$, such that, for every n, σ^n is an ε-Nash equilibrium on $\hat{\Omega}^n$, and $\mathbb{P}(\tilde{\Omega}^n) \to 1$ as $n \to \infty$.

For each n and (i, j), consider the strategy

$$\sigma^n_{ij}(q) = \begin{cases} 1, & \text{if } h_i\left(\dfrac{q}{\sqrt{n}\lambda_i}\right) \le r_i, \\ 0, & \text{otherwise,} \end{cases} \qquad q \in \mathbb{Z}_+. \tag{22}$$

Theorem 2.1. *For any $\varepsilon > 0$, under each of the two scheduling policies defined above, the sequence of strategy profiles $\{\sigma^n\}$ defined in (22) is an ε-Nash equilibrium w.h.p.*

Remark 2.2 (Relation to Naor's Result). The decision threshold expressed by (22) is closely related to that from Naor's celebrated result (Naor [16]; see also Hassin and Haviv [13], Section 2.1), that addresses an *expected* delay cost, and a nonasymptotic regime. One can, in fact, recover Naor's threshold from (22). To this end, consider Theorem 2.1 in the case of a single class ($N = 1$), and assume that $h(x) = cx$, $x \ge 0$. This assumption corresponds to a cost $c^n := c\sqrt{n}$ per unit time incurred by a customer who decides to join. By (22), using the heavy traffic condition $\lambda = \mu$, the customer joins if and only if $c(Q^n(t)/(\sqrt{n}\mu)) \le r$, namely, $Q^n(t) \le (rn\mu)/c^n$. Since μ^n is asymptotic to $n\mu$, the above threshold is asymptotic to $r\mu^n/c^n$, which gives Naor's threshold (cf. with Hassin and Haviv [13, Equation (2.1)]).

Remark 2.3 (Information on Which Decisions Are Based). The decision rule (22) involves the ratio $q/(\sqrt{n}\lambda_i)$. If we translate (22) to a decision based on the cost functions h^n_i, which correspond to the actual (unnormalized) waiting times, we see that the ratio $q/(n\lambda_i)$ is required to be computed in order to make the decision. In particular, customers need access to the current state of the system, namely, the queue length, and the system parameters, specifically $n\lambda_i$. In practical applications, this means that the system manager should provide information on the rate of arrival. Since the rate of arrival, λ^n_i, is only needed up to a first order approximation, $n\lambda_i$, it seems natural to achieve such an approximation by counting recent arrivals over an interval of time, in applications where such a procedure is feasible. The interval should be sufficiently long for the law of large numbers to yield effective estimates. An interesting open question that arises from this discussion is whether one could obtain results similar to ours when system parameters such as λ_i and μ_i are not available to the decision makers.

Remark 2.4 (RSP Does Not Always Hold). One of the main issues we address is the validity of RSP under the scheduling policies considered. To prove the main result, this principle needs to hold in a strong form, namely, that, w.h.p., *every* customer arriving, and joining, in the given time interval $[0, \bar{T}]$, experiences a delay given, with high precision, by the ratio between queue length and arrival rate. It should be noted that this property is not valid for arbitrary scheduling. For example, consider a scheduling that prioritizes class 1 over class 2 up to a certain fixed time, t_0, and then switches to the a priority of 2 over 1. The standard prediction is that the diffusion scale waiting time for a class-i customer is approximately given by $\widehat{\mathrm{WT}} \approx \lambda_i^{-1}\hat{Q}_i = (\rho_i\mu_i)^{-1}\hat{Q}_i$, where $\hat{Q}_i$ is the diffusion scale queue length at the arrival time. Now, consider a class-2 customer present in the buffer at time t_0. Such a customer will be sent to service approximately $(\rho_1\mu_1+\rho_2\mu_2)^{-1}\hat{q}$ units of time after t_0, where $\hat{q}=n^{-1/2}q$, and q is its position in line at t_0, because when 2 has priority, every server in the pool to become available will pick a customer from buffer 2. Hence, w.h.p., most customers that are in buffer 2 at time t_0, that are, in fact, $O(\sqrt{n})$ in number, will experience a delay significantly different than that predicted by RSP. This number increases even further under a policy that switches priority many times during the time interval in question. While these policies may not be particularly interesting by their own right, this discussion shows that there is content in the assertion that the principle does hold for the policies of interest.

Remark 2.5 (Individual Decisions May Have Long Term Effect). The analysis must take into account the possible behaviour of customers that do not follow the proposed rule. At the technical level, the estimates that lead to existence of diffusion limits are dealt with for different behaviours of customers. It may seem that it is enough to consider the behavior of the system when all customers follow the proposed rule, and then argue that the behaviour of a single customer will have a negligible effect. It should be noted, however, that the decision of one customer may significantly affect the waiting time of other customers. As a simple example for that, consider a two-class system under FP, where, at a certain time, a high-priority customer arrives to find an empty buffer of its own class. If he decides to leave, and for a little while there are no new arrivals, then the first-in-line customer at the low-priority class will get served as soon as a server becomes available. If he joins, it is possible that a large number of high-priority customers will join soon after, so that the waiting time of the low-priority customer referred to above will delay considerably. Hence a single player's decision may have a significant effect on other players.

The proof of Theorem 2.1 is based on analysis at the diffusion scale. On the way to proving it, we obtain diffusion limit results for the two policies under consideration, namely, Proposition 3.3 for FP, and Proposition 4.3 for SLQ.

3. Fixed priority.

This section is devoted to a convergence result in the case where the servers implement the FP scheduling. It provides the main estimates that determine the limiting behaviour of the fluid and diffusion scaled processes, that are later used to prove RSP.

Throughout, $\sigma^n = \{\sigma^n_{ij}\}$ denotes the strategy profile (22). Given (i,j), denote by $\bar{\sigma}^n_{ij}\in\Sigma$ the strategy $\bar{\sigma}^n_{ij} = 1-\sigma^n_{ij}$, that acts precisely as the negation of σ^n_{ij}. We begin by noting that in order to show that σ^n is an ε-Nash equilibrium w.h.p., it suffices to consider (21) with $\tau=\bar{\sigma}^n_{ij}$ only. Indeed, given (i,j) and $\tau\in\Sigma$, define $A=\{q\in\mathbb{Z}_+\colon \tau(q)\neq\sigma^n_{ij}(q)\}$. Then we have

$$C^n_{ij}(\tau,\sigma^{n,ij}) = \begin{cases} C^n_{ij}(\bar{\sigma}^n_{ij},\sigma^{n,ij}), & \text{if } Q^n_i(\mathrm{AT}^n_{ij}-)\in A,\\ C^n_{ij}(\sigma^n_{ij},\sigma^{n,ij}), & \text{if } Q^n_i(\mathrm{AT}^n_{ij}-)\in A^c,\end{cases}$$

and so the validity of (21) for $\tau=\bar{\sigma}^n_{ij}$ and $\tau=\sigma^n_{ij}$ (the latter being trivial) implies the validity of this inequality for $\tau\in\Sigma$.

We will use the term *scenario* for the collection of processes obtained under any one of the strategy profiles $(\bar{\sigma}^n_{ij},\sigma^{n,ij})$. More precisely, let us fix n. Recall that, for $\sigma\in\bar{\Sigma}$, $\mathscr{S}^n(\sigma)$ denotes the dynamics obtained when a strategy profile σ is played. Let

$$\mathfrak{S} = \{(i,j)\colon i\in\{1,2,\dots,N\},\ j\in\mathbb{N}\}.$$

For $s=(i,j)\in\mathfrak{S}$, the *scenario* s is defined to be $\mathscr{S}^n(\bar{\sigma}^n_{ij},\sigma^{n,ij})$, namely, the dynamics corresponding to player (i,j) playing $\bar{\sigma}^n_{ij}$ and all other players (k,l) playing σ^n_{kl}. In addition, *scenario* 0, that we will also call the *reference* scenario, is defined as $\mathscr{S}^n(\sigma^n)$. Scenarios are thus indexed by the set $\mathfrak{S}_0 := \mathfrak{S}\cup\{0\}$. As we have just argued, the main result will follow once we show that there exist events $\tilde{\Omega}^n$ such that, for every n, on $\tilde{\Omega}^n$,

$$\forall (i,j)\quad C^n_{ij}(\sigma^n_{ij},\sigma^{n,ij}) \le C^n_{ij}(\bar{\sigma}^n_{ij},\sigma^{n,ij})+\varepsilon, \tag{23}$$

and $\mathbb{P}(\tilde{\Omega}^n)\to 1$ as $n\to\infty$. We thus work in what follows with scenarios. To address all scenarios simultaneously, the dependence of the processes on the scenario has to be reflected in the notation. For each of the processes introduced above, except for the stochastic primitives and their scaled versions, an additional superscript s will indicate that the process is considered under scenario $s\in\mathfrak{S}_0$. Thus, for example, $Q^{n,s}=Q^n(\bar{\sigma}^n_{ij},\sigma^{n,ij})$ if $s=(i,j)$, and $Q^{n,s}=Q^n(\sigma^n)$ if $s=0$.

Throughout what follows, we adopt the convention that $e^{n,s}(t)$ (or sometimes $e_i^{n,s}(t)$), $t \in [0,T]$, denotes a generic family of processes, indexed by $n \in \mathbb{N}$ and $s \in \mathfrak{S}_0$, that can change from one appearance to another, and has the property

$$\sup_s \|e^{n,s}\|_T \to 0, \quad \text{in probability, as } n \to \infty. \tag{24}$$

The balance Equations (6)–(8) have the following form when translated to the diffusion scale, namely,

$$\hat{Q}_i^{n,s}(t) = \hat{Q}_i^n(0) + \hat{E}_i^n(t) - \hat{B}_i^{n,s}(t) - \hat{R}_i^{n,s}(t) + n^{-1/2}(\lambda_i^n - n\lambda_i)t, \tag{25}$$

$$\hat{\Psi}_i^{n,s}(t) = \hat{\Psi}_i^n(0) + \hat{B}_i^{n,s}(t) - \hat{S}_i^n\left(\int_0^t \bar{\Psi}_i^{n,s}(u)\,du\right) - \mu_i \int_0^t \hat{\Psi}_i^{n,s}(u)\,du. \tag{26}$$

Let $X_i^{n,s} = Q_i^{n,s} + \Psi_i^{n,s}$ represent the total number of class-i customers in the system, and let its scaled version be defined by

$$\hat{X}_i^{n,s}(t) = \frac{X_i^{n,s}(t) - \rho_i n}{\sqrt{n}} = \hat{Q}_i^{n,s}(t) + \hat{\Psi}_i^{n,s}(t). \tag{27}$$

Then by the assumptions on the initial conditions we have

$$\hat{X}^n(0) \to \Psi(0) =: X_0.$$

Our first estimate addresses the scaled queue lengths of the high-priority classes.

Lemma 3.1. *For $i = 1, 2, \ldots, N-1$ and for any $T < \infty$ we have*

$$\sup_s \|\hat{Q}_i^{n,s}\|_T \to 0, \quad \textit{in probability.}$$

Proof. From the functional central limit theorem we have,

$$(\hat{E}^n, \hat{S}^n) \Rightarrow (W_1, W_2), \tag{28}$$

where W_1 and W_2 are independent N-dimensional BMs, with W_1 a $(0, A_1)$-BM and W_2 a $(0, A_2)$-BM, $A_1 = \operatorname{diag}(\lambda_i C_{\mathrm{IA}_i}^2)$, and $A_2 = \operatorname{diag}(\mu_i)$ (see Billingsley [7, Section 17]). In particular, the sequence $(\hat{E}^n, \hat{S}^n)$ is C-tight.

Fix $\epsilon > 0$. Define the event

$$\Omega^n = \left\{ \sum_{i=1}^{N-1} \hat{Q}_i^n(0) \le \frac{\epsilon}{4} \text{ and } \bar{\Psi}_i^n(0) \ge \rho_i - \frac{\varepsilon_i}{4}, \text{ for all } i \in \{1, 2, \ldots, N-1\} \right\},$$

where $\varepsilon_i = \epsilon/(\mu_i(N-1))$. Then by the assumption (18) on the initial conditions we have $\mathbb{P}(\Omega^n) \to 1$. For $s \in \mathfrak{S}_0$ define

$$\tau_1^{n,s} = \inf\left\{ t \ge 0 \colon \sum_{i=1}^{N-1} \hat{Q}_i^{n,s}(t) \ge \epsilon \text{ or } \bar{\Psi}_i^{n,s}(t) \le \rho_i - \varepsilon_i, \text{ for some } i \in \{1, 2, \ldots, N-1\} \right\}.$$

Let $A^{n,s} = \{\tau_1^{n,s} \le T\}$. Now let

$$A_1^{n,s} = \left\{ \omega \in A^{n,s} \colon \sum_{i=1}^{N-1} \hat{Q}_i^{n,s}(\tau_1^{n,s}) \ge \epsilon \right\} \cap \Omega^n,$$

$$A_2^{n,s,i} = \left\{ \omega \in A^{n,s} \colon \sum_{k=1}^{N-1} \hat{Q}_k^{n,s}(\tau_1^{n,s}) < \epsilon \text{ and } \bar{\Psi}_i^{n,s}(\tau_1^{n,s}) \le \rho_i - \varepsilon_i \right\} \cap \Omega^n, \quad i \le N-1.$$

For $\omega \in A_1^{n,s}$ there exists $\eta_1^{n,s} = \eta_1^{n,s}(\omega)$ such that

$$\sum_{i=1}^{N-1} \hat{Q}_i^{n,s}(\eta_1^{n,s}) \le \frac{\epsilon}{2}, \qquad \text{and} \qquad \text{on } I_1^{n,s} := [\eta_1^{n,s}, \tau_1^{n,s}], \quad \sum_{i=1}^{N-1} \hat{Q}_i^{n,s} > 0. \tag{29}$$

Throughout, for $0 \le t_1 \le t_2 < \infty$, $I = [t_1, t_2]$ and $f \colon \mathbb{R}_+ \to \mathbb{R}$, we use the notation

$$f[t_1, t_2] = f[I] = f(t_2) - f(t_1).$$

By (6) and the fact that R_i^n is nondecreasing, we have on $A_1^{n,s}$

$$\frac{\epsilon\sqrt{n}}{2} \le \sum_{i=1}^{N-1} Q_i^{n,s}[I_1^{n,s}] \le \sum_{i=1}^{N-1} E_i^n[I_1^{n,s}] - \sum_{i=1}^{N-1} B_i^{n,s}[I_1^{n,s}]. \tag{30}$$

By (29) and (9), $\mathbf{1}\cdot\Psi^{n,s}(t) = n$ for $t = \eta_1^{n,s}$ and $t = \tau_1^{n,s}$. Thus by (7), $\mathbf{1}\cdot B^{n,s}[I_1^{n,s}] = \mathbf{1}\cdot D^{n,s}[I_1^{n,s}]$. Moreover, since by (29) the high-priority buffers are nonempty on the time interval of interest, the priority rule expressed by (10) dictates that the process $B_N^{n,s}$ does not increase on that interval. As a result, the last term in (30) equals $\mathbf{1}\cdot D^{n,s}[I_1^{n,s}]$, and

$$\frac{\epsilon}{2} \le \sum_{i=1}^{N-1} \hat{E}_i^n[I_1^{n,s}] + \sum_{i=1}^{N-1} \frac{\lambda_i^n(\tau_1^{n,s} - \eta_1^{n,s})}{\sqrt{n}} - \sum_{i=1}^{N} \hat{D}^{n,s}[I_1^{n,s}] - n^{-1/2}\sum_{i=1}^{N}\mu_i \int_{\eta_1^{n,s}}^{\tau_1^{n,s}} \Psi_i^{n,s}(u)\,du.$$

On the time interval under consideration we have for $i < N$ that $\Psi_i^{n,s} \ge n\delta_i$, where $\delta_i = \rho_i - \varepsilon_i$. Thus, denoting $\mu_{\min} = \min_i \mu_i > 0$,

$$\begin{aligned}
\sum_{i=1}^{N} \mu_i \Psi_i^{n,s} &= \sum_{i=1}^{N-1} \mu_i(n\delta_i + \Psi_i^{n,s} - n\delta_i) + \mu_N \Psi_N^{n,s}\\
&\ge n\left(\sum_{i=1}^{N-1}\lambda_i - \epsilon\right) + \mu_{\min}\sum_{i=1}^{N-1}(\Psi_i^{n,s} - n\delta_i) + \mu_{\min}\Psi_N^{n,s}\\
&= n\left(\sum_{i=1}^{N-1}\lambda_i - \epsilon + \mu_{\min}\rho_N + \mu_{\min}\sum_{i=1}^{N-1}\varepsilon_i\right),
\end{aligned}$$

where the last equality uses the fact that $\sum_{i=1}^{N}\Psi_i^{n,s} = n$ that is true thanks to the nonidling condition (9) and the fact that, by (29), the queues are not all empty. Therefore for ϵ small enough there exists a $\delta > 0$, such that

$$\sum_{i=1}^{N}\mu_i \int_{\eta_1^{n,s}}^{\tau_1^{n,s}} \Psi_i^{n,s}(u)\,du \ge n\left(\sum_{i=1}^{N-1}\lambda_i + \delta\right)(\tau_1^{n,s} - \eta_1^{n,s}).$$

Hence we have

$$\begin{aligned}
\frac{\epsilon}{2} &\le \sum_{i=1}^{N-1}\hat{E}_i^n[I_1^{n,s}] - \sum_{i=1}^{N}\hat{D}^{n,s}[I_1^{n,s}]\\
&\quad + \sum_{i=1}^{N-1}\frac{(\lambda_i^n - n\lambda_i)(\tau_1^{n,s} - \eta_1^{n,s})}{\sqrt{n}} - \sqrt{n}\delta(\tau_1^{n,s} - \eta_1^{n,s}).
\end{aligned}$$

Let $r_n > 0$ be a sequence such that $r_n \to 0$ and $\sqrt{n}r_n \to \infty$. If $\tau_1^{n,s} - \sigma_1^{n,s} \le r_n$ then

$$\frac{\epsilon}{2} \le \sum_{i=1}^{N-1} w_T(\hat{E}_i^n, r_n) + \sum_{i=1}^{N} w_T(\hat{S}_i^n, r_n) + Kr_n,$$

where K is a constant and, throughout, for $f\colon \mathbb{R}_+ \to \mathbb{R}^k$ (k a positive integer),

$$w_T(f, a) = \sup\bigl\{\|f(t) - f(s)\|\colon s, t \in [0, T],\ |t - s| \le a\bigr\}, \qquad a > 0.$$

On the other hand, if $\tau_1^{n,s} - \sigma_1^{n,s} > r_n$ then

$$\frac{\epsilon}{2} \le 2\sum_{i=1}^{N-1}\|\hat{E}_i^n\|_T + KT + 2\sum_{i=1}^{N-1}\|\hat{S}_i^n\|_T - \sqrt{n}\delta r_n.$$

Hence by (28) and the resulting C-tightness of $\hat{E}_i^n$ and $\hat{S}_i^n$, we have

$$\mathbb{P}\left(\bigcup_s A_1^{n,s}\right) \to 0, \qquad \text{as } n \to \infty. \tag{31}$$

Next, on $A_2^{n,s,i}$, for $i \le N-1$ fixed, again there exists a time $\eta_2^{n,s} = \eta_2^{n,s}(\omega)$ such that

$$\bar{\Psi}_i^{n,s}(\eta_2^{n,s}) \ge \rho_i - \frac{\varepsilon_i}{2} \qquad \text{and} \qquad \text{on } I_2^{n,s} := [\eta_2^{n,s}, \tau_1^{n,s}], \quad \bar{\Psi}_i^{n,s} \le \rho_i.$$

Thus $X_i^{n,s}[I_2^{n,s}] \le \sqrt{n}\epsilon - n\varepsilon_i/2$, or

$$E_i^n[I_2^{n,s}] - D_i^{n,s}[I_2^{n,s}] - R_i^{n,ss}[I_2^{n,s}] \le \sqrt{n}\epsilon - \frac{n\varepsilon_i}{2},$$

and therefore

$$\hat{E}_i^n[I_2^{n,s}] - \hat{D}_i^{n,s}[I_2^{n,s}] + \frac{(\lambda_i^n - n\lambda_i)(\tau_1^{n,s} - \eta_1^{n,s})}{\sqrt{n}} \le \epsilon - \frac{\sqrt{n}\varepsilon_i}{2} + \frac{1}{\sqrt{n}},$$

whence

$$-2\|\hat{E}_i^n\|_T - \|\hat{S}_i^n\|_T - KT \le \epsilon - \frac{\sqrt{n}\varepsilon_i}{2} + \frac{1}{\sqrt{n}}.$$

Therefore by the tightness of $\|\hat{E}_i^n\|_T$ and $\|\hat{S}_i^n\|_T$, $n \in \mathbb{N}$ (for T fixed), we have

$$\mathbb{P}\Big(\bigcup_s A_2^{n,s,i}\Big) \to 0, \quad \text{as } n \to \infty,\ i \le N-1. \tag{32}$$

Putting together the estimates (31) and (32), we obtain $\mathbb{P}(\bigcap_s (A^{n,s})^c) \to 1$. Since $\varepsilon > 0$ is arbitrary, the result follows. □

Define

$$\theta_i = \lambda_i h_i^{-1}(r_i), \tag{33}$$

and note that these constants are positive. By (22), under the reference scenario, class-i customers always renege when the scaled queue length $\hat{Q}_i^n$ is in the interval $(\theta_i, \theta_i + 1/\sqrt{n}]$ and therefore the scaled queue length never exceeds that bound. Under any other scenario, there is at most one customer that does not follow the rule (22), and so we have

$$\hat{Q}_i^{n,s}(t) \le \theta_i + 2n^{-1/2}, \quad t \ge 0, \quad n \in \mathbb{N}, \quad s \in \mathfrak{S}_0, \quad i = 1, \dots, N. \tag{34}$$

Conversely, a class-i reneging will never take place when $h_i(\hat{Q}_i^{n,s}/\lambda_i) < r_i$, except, possibly, by a single customer.

Lemma 3.2. (i) *For $i = 1, 2, \dots, N-1$,*

$$\sup_s \hat{R}_i^{n,s}(T) \to 0, \quad \textit{in probability, as } n \to \infty. \tag{35}$$

(ii) *For $i = 1, \dots, N$,*

$$\sup_s \|\bar{\Psi}_i^{n,s} - \rho_i\|_T \to 0, \quad \textit{in probability, as } n \to \infty. \tag{36}$$

Proof. (i) By the discussion preceding the Lemma, $R_i^{n,s}(T) \le 1$ on the event that $\|\hat{Q}_i^{n,s}\|_T < \theta_i$. Hence (35) follows from Lemma 3.1.

(ii) We begin by proving the result for the high-priority classes. Thus, fix $i \le N-1$. We have by (6),

$$\begin{aligned}\bar{Q}_i^{n,s}(t) &= \bar{Q}_i^n(0) + \bar{E}_i^n(t) - \bar{B}_i^{n,s}(t) - \bar{R}_i^{n,s}(t)\\ &= \bar{Q}_i^n(0) + (\bar{E}_i^n(t) - \lambda_i t) - (\bar{B}_i^{n,s}(t) - \lambda_i t) - \bar{R}_i^{n,s}(t).\end{aligned}$$

By the functional law of large numbers, $\sup_{0\le t\le T} |\bar{E}_i^n(t) - \lambda_i t| \to 0$ in probability. Hence the estimates of Lemma 3.1 give (recall the convention (24))

$$\bar{B}_i^{n,s}(t) = \lambda_i t + e_t^{n,s}. \tag{37}$$

Next, by (7) and (8), using the identity $\rho_i = \lambda_i/\mu_i$,

$$\begin{aligned}\bar{\Psi}_i^{n,s}(t) - \rho_i = \bar{\Psi}_i^n(0) - \rho_i + (\bar{B}_i^{n,s}(t) - \lambda_i t) - \frac{1}{n}\bigg[S_i\bigg(\int_0^t n\bar{\Psi}_i^{n,s}(u)\,du\bigg) - \mu_i \int_0^t n\bar{\Psi}_i^{n,s}(u)\,du\bigg]\\ - \mu_i \int_0^t (\bar{\Psi}_i^{n,s}(u) - \rho_i)\,du.\end{aligned}$$

Using the fact $\bar{\Psi}_i^{n,s} \le 1$ we have, for $t \in [0, T]$,

$$|\bar{\Psi}_i^{n,s}(t) - \rho_i| \le |\bar{\Psi}_i^n(0) - \rho_i| + \beta_T^n + n^{-1/2}\|\hat{S}_i^n\|_T + \mu_i \int_0^T \sup_s \sup_{0\le r\le u} |\bar{\Psi}_i^{n,s}(r) - \rho_i|\,du.$$

It follows from (18) that $\|\bar{\Psi}^n(0)-\rho\|\to 0$ and from the law of large numbers for the Poisson process, that $n^{-1/2}\|\hat{S}_i^n\|_T\to 0$, in probability. Using these facts along with (37), the result (36), for $i\le N-1$, follows upon applying Gronwall's lemma.

Next we consider the class N. Because $\sum\rho_i=1$ and $\sum\bar{\Psi}_i^{n,s}\le 1$, we have from the validity of (36) for $i\le N-1$,

$$\sup_s\sup_{0\le t\le T}(\bar{\Psi}_N^{n,s}(t)-\rho_N)^+\to 0$$

in probability as $n\to\infty$. Using this and the assumption on the initial conditions, the probability of $\Omega_1^n:=\{\gamma^n<\varepsilon/16\}\cap\{|\bar{\Psi}_N^n(0)-\rho_N|<\varepsilon/2\}$ converges to 1, where $\gamma^n=\sup_s\sum_{i\le N-1}\|\bar{\Psi}_i^{n,s}-\rho\|_T$. Now let

$$\Omega^{n,s}=\Big\{\omega\colon \inf_{0\le t\le T}\bar{\Psi}_N^{n,s}(t)\le\rho_N-\epsilon\Big\}.$$

Then for $\omega\in\Omega^{n,s}\cap\{|\bar{\Psi}_N^n(0)-\rho_N|<\varepsilon/2\}$, there exist times $0\le\eta_3^{n,s}(\omega)\le\tau_3^{n,s}(\omega)\le T$ such that

$$\bar{\Psi}_N^{n,s}(\eta_3^{n,s})>\rho_N-\frac{\epsilon}{2},\qquad \bar{\Psi}_N^{n,s}(\tau_3^{n,s})\le\rho_N-\varepsilon,\qquad\text{and}\qquad \bar{\Psi}_N^{n,s}(t)\le\rho_N-\frac{\epsilon}{8},\quad\text{for all } t\in I^{n,s}:=[\eta_3^{n,s},\tau_3^{n,s}].$$

Also, on the event $\Omega^{n,s}\cap\{\gamma^n<\varepsilon/16\}$,

$$\sum_{i=1}^{N-1}\bar{\Psi}_i^{n,s}(t)\le\sum_{i=1}^{N-1}\rho_i+\frac{\epsilon}{16},\qquad\text{for all } t\in I^{n,s}.$$

Thus on $I^{n,s}$ we have $\sum_{i=1}^N\bar{\Psi}_i^{n,s}(t)\le 1-\epsilon/16<1$, which implies by the nonidling assumption that, on this time interval, we have $\sum_{n=1}^N Q_i^{n,s}(t)=0$. As a result, on this time interval there is no reneging under the reference scenario, and there is at most one reneging under any other scenario. Recalling that $X^{n,s}=Q^{n,s}+\Psi^{n,s}$, and using (6) and (7), we obtain, for a given scenario s, on the event $\Omega_1^n\cap\Omega^{n,s}$,

$$\begin{aligned}-\frac{n\epsilon}{2}&\ge X_N^{n,s}[I^{n,s}]\ge E_N^n[I^{n,s}]-D_N^{n,s}[I^{n,s}]-1\\ &=\sqrt{n}\hat{E}_N^n[I^{n,s}]+\lambda_N^n(\tau_3^{n,s}-\eta_3^{n,s})-\sqrt{n}\hat{D}_N^{n,s}[I^{n,s}]-n\mu_N\int_{I^{n,s}}\bar{\Psi}_N^{n,s}(u)\,du-1.\end{aligned}$$

Note that

$$\mu_N\int_{I^{n,s}}\bar{\Psi}_N^{n,s}(u)\,du\le\mu_N\rho_N(\tau_3^{n,s}-\eta_3^{n,s})=\lambda_N(\tau_3^{n,s}-\eta_3^{n,s}),$$

thus

$$-\frac{\epsilon}{2}\ge-2\frac{\|\hat{E}_N^n\|_T}{\sqrt{n}}-2\frac{\|\hat{D}_N^n\|_T}{\sqrt{n}}+\frac{(\lambda_N^n-n\lambda_N)(\tau_3^{n,s}-\eta_3^{n,s})}{n}-\frac{1}{n}.$$

Since $0\le\bar{\Psi}_N^{n,s}\le 1$, we have $\|\hat{D}^{n,s}\|_T\le\|\hat{S}^n\|_T$. Also, $n^{-1/2}(\lambda_N^n-n\lambda_N)$ converges. Hence

$$-\frac{\varepsilon}{2}\ge-2\frac{\|E_N^n\|_T}{\sqrt{n}}-2\frac{\|S_N^n\|_T}{\sqrt{n}}-\frac{KT}{\sqrt{n}}-\frac{1}{n}.$$

By the tightness of $\|\hat{E}_N^n\|_T$ and $\|\hat{S}_N^n\|_T$ for $n\in\mathbb{N}$ (and T fixed) and the fact that $\mathbb{P}(\Omega_1^n)\to 1$, we obtain $\mathbb{P}(\bigcup_s\Omega^{n,s})\to 0$. Since $\varepsilon>0$ is arbitrary, the result follows. □

Consider a stochastic differential equation with reflection, for a process Y that lives in

$$G=\{y\in\mathbb{R}^N\colon \mathbf{1}\cdot y\le\theta_N\},$$

and reflects on the boundary of G in the direction $-\mathbf{e}_N$. Let $\{W(t)\}$ be a $(\hat{\lambda},A)$-BM, where $A=\operatorname{diag}(\lambda_1(C_{IA_1}^2+1),\dots,\lambda_N(C_{IA_N}^2+1))$. Let $b\colon\mathbb{R}^N\to\mathbb{R}^N$ be given by

$$b(y)=-\big(\mu_1y_1,\dots,\mu_{N-1}y_{N-1},\mu_N(y_N-(\mathbf{1}\cdot y)^+)\big).\tag{38}$$

Let (X,L) be the unique pair of processes that is adapted to the filtration $\sigma(X_0)\vee\sigma\{W(u),u\le t\}$, where X has sample paths in $\mathbb{C}_G$, L has nondecreasing sample paths in $\mathbb{C}_{\mathbb{R}_+}$, and the pair satisfies a.s.,

$$\begin{gathered}X(t)=X_0+W(t)+\int_0^t b(X(u))\,du-L(t)\mathbf{e}_N,\qquad t\ge 0,\\ \int_{[0,\infty)}1_{\{\mathbf{1}\cdot X(t)<\theta_N\}}\,dL(t)=0.\end{gathered}\tag{39}$$

The existence and uniqueness of such a pair follows from Proposition 3 of Anderson and Orey [3] on noting that b is Lipschitz. We call this pair the solution to the SDE (39).

Define $\Gamma\colon \mathbb{D}_{\mathbb{R}^N}([0,T]) \to \mathbb{D}_{\mathbb{R}^N}([0,T])$ by

$$\Gamma(f)(t) = f(t) - g(t)\mathbf{e}_N, \quad g(t) = \sup_{0\le u\le t}(\theta_N - \mathbf{1}\cdot f(u))^-. \tag{40}$$

The following two properties follow directly from the definition, namely, there exists a constant C such that

$$\|\Gamma(f) - \Gamma(\tilde f)\|_T \le C\|f - \tilde f\|_T, \quad f, \tilde f \in \mathbb{D}_{\mathbb{R}^N}([0,T]), \tag{41}$$

and

$$w_T(\Gamma(f),\cdot) \le C w_T(f,\cdot), \quad f \in \mathbb{D}_{\mathbb{R}^N}([0,T]). \tag{42}$$

Given $z \in \mathbb{D}_{\mathbb{R}^N}$, $z(0) \in G$, we say that $(y,\ell) \in \mathbb{D}_{\mathbb{R}^N} \times \mathbb{D}_{\mathbb{R}}$ solves the Skorohod problem (SP) in G, with reflection in the direction $-\mathbf{e}_N$, for data z, if $y(t) \in G$ for all t, ℓ is nonnegative and nondecreasing, and

$$y = z - \ell\mathbf{e}_N, \qquad \int_{[0,\infty)} 1_{\{\mathbf{1}\cdot y < \theta_N\}}\, d\ell = 0.$$

It is well known that for z as above, a necessary and sufficient condition for (y,ℓ) to be a solution is that $y = \Gamma(z)$ (this follows, e.g., as a special case of the much broader result of Dupuis and Ishii [8]). This will be used in the proof below.

Denote

$$\hat W_i^{n,s}(t) = \hat E_i^n(t) + \frac{\lambda_i^n - n\lambda_i}{\sqrt{n}}t - \hat S_i^n\biggl(\int_0^t \bar\Psi_i^{n,s}(u)\,du\biggr). \tag{43}$$

Recall conditions C1–C2 from Section 1 that characterize C-tightness. We will say that a sequence of processes $\{\xi^{n,s}\}$, $n \in \mathbb{N}$, $s \in \mathfrak{S}_0$, with sample paths in $\mathbb{D}_{\mathbb{R}^k}$, is *$C$-tight, uniformly in s* if

C1′. The sequence of random variables $\|\xi^{n,s}\|_T$ is tight for every fixed $T < \infty$, and

C2′. For every $T < \infty$, $\varepsilon > 0$, and $\eta > 0$ there exist n_0 and $\theta > 0$ such that

$$n \ge n_0 \ \text{ implies } \ \mathbb{P}\Bigl(\sup_s w_T(\xi^{n,s},\theta) > \eta\Bigr) < \varepsilon.$$

Proposition 3.3. *The sequence $(\hat W^{n,s}, \hat X^{n,s}, \hat R^{n,s}, \hat Q^{n,s}, \hat\Psi^{n,s})$ is C-tight, uniformly in s. Moreover, $(\hat W^{n,0}, \hat X^{n,0}, \hat R^{n,0}, \hat Q^{n,0}, \hat\Psi^{n,0})$ converges in distribution to $(W, X, L\mathbf{e}_N, Q, \Psi)$, where (X, L) form the solution to the SDE (39), and*

$$Q = (\mathbf{1}\cdot X)^+\mathbf{e}_N, \qquad \Psi = X - Q.$$

Proof. The C-tightness of $\hat W^{n,s}$, uniformly in s, follows from (43) using (28) and the fact that $\bar\Psi_i^{n,s} \le 1$. By (25)–(27),

$$\hat X_i^{n,s} = \hat X_i^n(0) + \hat W_i^{n,s} - \mu_i\int_0^\cdot \hat\Psi_i^{n,s}(u)\,du - \hat R_i^{n,s}.$$

Thus

$$\hat\Psi_i^{n,s} = \hat X_i^n(0) + \hat W_i^{n,s} - \hat Q_i^{n,s} - \mu_i\int_0^\cdot \hat\Psi_i^{n,s}(u)\,du - \hat R_i^{n,s}, \quad i = 1,\dots,N-1,$$

and, noting that by (9) one has $\mathbf{1}\cdot\hat Q^{n,s} = (\mathbf{1}\cdot\hat X^{n,s})^+$,

$$\begin{aligned}
\hat X_N^{n,s} &= \hat X_N^n(0) + \hat W_N^{n,s} - \mu_N\int_0^\cdot (\hat X_N^{n,s}(u) - (\mathbf{1}\cdot\hat X^{n,s}(u))^+)\,du - \mu_N\int_0^\cdot \sum_{i=1}^{N-1}\hat Q_i^{n,s}(u)\,du - \hat R_N^{n,s}\\
&= \hat X_N^n(0) + \hat W_N^{n,s} - \mu_N\int_0^\cdot\biggl(\hat X_N^{n,s}(u) - \Bigl(\hat X_N^{n,s}(u) + \sum_{i=1}^{N-1}\hat\Psi_i^{n,s}(u)\Bigr)^+\biggr)\,du\\
&\quad - \mu_N\int_0^t \sum_{i=1}^{N-1}\hat Q_i^{n,s}(u)\,du + \mu_N\int_0^\cdot\biggl\{(\mathbf{1}\cdot\hat X^{n,s}(u))^+ - \biggl(\hat X_N^{n,s}(u) + \sum_{i=1}^{N-1}\hat\Psi_i^{n,s}(u)\biggr)^+\biggr\}\,du - \hat R_N^{n,s}.
\end{aligned}$$

Defining $Y_i^{n,s} = \hat\Psi_i^{n,s}$, $i = 1,\dots,N-1$, and $Y_N^{n,s} = \hat X_N^{n,s}$, we have, using Lemmas 3.1 and 3.2(i),

$$Y_i^{n,s} = \hat X_i^n(0) + \hat W_i^{n,s} - \mu_i\int_0^\cdot Y_i^{n,s}(u)\,du + e_i^{n,s}, \quad i = 1,\dots,N-1, \tag{44}$$

$$Y_N^{n,s} = \hat X_N^n(0) + \hat W_N^{n,s} - \mu_N\int_0^\cdot (Y_N^{n,s}(u) - (\mathbf{1}\cdot Y^{n,s}(u))^+)\,du - \hat R_N^{n,s} + e^{n,s}. \tag{45}$$

Let

$$F^{n,s} = \mathbf{1}\cdot Y^{n,s}\wedge\theta_N - \mathbf{1}\cdot Y^{n,s}. \tag{46}$$

Then

$$\mathbf{1}\cdot Y^{n,s} = \hat X_N^{n,s} + \sum_{i=1}^{N-1}\hat\Psi_i^{n,s} = \hat Q_N^{n,s} + \mathbf{1}\cdot\hat\Psi^{n,s} \le \hat Q_N^{n,s} \le \theta_N + \frac{2}{\sqrt{n}}, \tag{47}$$

by (34). Thus $|F^{n,s}|\le 2/\sqrt{n}$. Further define $\tilde Y_i^{n,s} = Y_i^{n,s} + (1/N)F^{n,s}$, $i=1,\dots,N$. Then $\tilde Y^{n,s}$ satisfies

$$\tilde Y^{n,s}(t)\in G,\qquad t\ge 0, \tag{48}$$

and, as follows from (38), (44), and (45),

$$\tilde Y^{n,s} = \hat X^n(0) + \hat W^{n,s} + \int_0^{\cdot} b(\tilde Y^{n,s}(u))\,du - \hat R_N^{n,s}\mathbf{e}_N + e^{n,s}. \tag{49}$$

Under the reference scenario, no class-N reneging occurs when $\hat Q^{n,0}<\theta_N$, that is,

$$\int 1_{\{\hat Q_N^{n,0}(t-)<\theta_N\}}\,d\hat R_N^{n,0}(t) = 0.$$

As a result, the same is true with $\hat Q_N^{n,0}(t-)$ replaced by $\hat Q_N^{n,0}(t)$. Under any other scenario, there may be one customer that does not follow the rule. For $s=(N,j)$, $j\in\mathbb{N}$, write $\tilde R_N^{n,s}$ for the normalized reneging count of all class-N customers except for customer (N,j) (if it reneges). For any other $s\in\mathfrak{S}_0$, let $\tilde R_N^{n,s} = \hat R_N^{n,s}$. Then $\tilde R_N^{n,s}$ is nondecreasing and satisfies

$$|\tilde R_N^{n,s} - \hat R_N^{n,s}| \le n^{-1/2}, \tag{50}$$

and

$$\int 1_{\{\hat Q_N^{n,s}(t)<\theta_N\}}\,d\tilde R_N^{n,s}(t) = 0.$$

Let us show that $\mathbf{1}\cdot\tilde Y^{n,s}<\theta_N$ implies $\hat Q_N^{n,s}<\theta_N$. Indeed, by (46), the former implies that $\mathbf{1}\cdot Y^{n,s}<\theta_N$. Now, $\mathbf{1}\cdot Y^{n,s} = \hat Q_N^{n,s} + \mathbf{1}\cdot\hat\Psi^{n,s}$, by (47). Thus either $\hat Q_N^{n,s}=0$, or $\hat Q_N^{n,s}>0$ in which case $\mathbf{1}\cdot\Psi^{n,s}=0$ by the nonidling condition (9). In both cases, $\hat Q_N^{n,s}<\theta_N$. It thus follows that

$$\int 1_{\{\mathbf{1}\cdot\tilde Y^{n,s}<\theta_N\}}\,d\tilde R_N^{n,s} = 0. \tag{51}$$

By (49) and (50),

$$\tilde Y^{n,s} = \hat X^n(0) + \hat W^{n,s} + \int_0^{\cdot} b(\tilde Y^{n,s}(u))\,du - \tilde R^{n,s}\mathbf{e}_N + e^{n,s}. \tag{52}$$

Hence from (48), (51), and (52), $(\tilde Y^{n,s}, \tilde R_N^{n,s})$ solves the aforementioned SP for the data

$$\hat X^n(0) + \hat W^{n,s} + \int_0^{\cdot} b(\tilde Y^{n,s}(u))\,du + e^{n,s}.$$

Therefore

$$\tilde Y^{n,s} = \Gamma\biggl(\hat X^n(0) + \hat W^{n,s} + \int_0^{\cdot} b(\tilde Y^{n,s}(u))\,du + e^{n,s}\biggr), \tag{53}$$

$$\tilde R^{n,s}\mathbf{e}_N = (I-\Gamma)\biggl(\hat X^n(0) + \hat W^{n,s} + \int_0^{\cdot} b(\tilde Y^{n,s}(u))\,du + e^{n,s}\biggr). \tag{54}$$

The convergence of $\hat X^n(0)$, the uniform C-tightness of $\hat W^{n,s}$, the Lipschitz property of b and the Lipschitz property of Γ, as expressed by (41), imply tightness of the r.v.s $\sup_s\|\tilde Y^{n,s}\|_T$, upon an application of Gronwall's lemma to (53). Hence, using again (53), along with the property (42), shows that the processes $\tilde Y^{n,s}$ are C-tight, uniformly in s. As a result, $\tilde R_N^{n,s}$ are also C-tight, uniformly in s. By Equations (52)–(54), any subsequential weak limit of $(\hat W^{n,0}, \tilde Y^{n,0}, \tilde R_N^{n,0})$ must be equal in distribution to (W,X,L). As a result, $(\hat W^{n,0}, \tilde Y^{n,0}, \tilde R_N^{n,0})\Rightarrow(W,X,L)$. From the definition of $\tilde Y^{n,s}$ and Lemma 3.1 it follows that $\hat X^{n,s} = \tilde Y^{n,s} + e^{n,s}$. Moreover, since by Lemma 3.2, $\hat R_i^{n,s} = e^{n,s}$ for $i\le N-1$, we have $(\hat W^{n,0}, \hat X^{n,0}, \hat R^{n,0})\Rightarrow(W,X,L\mathbf{e}_N)$. Finally, the fact $\hat Q_i^{n,s}=e^{n,s}$, $i\le N-1$, stated in Lemma 3.1, and the relations $\mathbf{1}\cdot\hat Q^{n,s} = (\mathbf{1}\cdot\hat X^{n,s})^+$, $\hat\Psi^{n,s} = \hat X^{n,s} - \hat Q^{n,s}$ yield the result by the continuous mapping theorem. □

4. Serve the longest queue.

In this section we carry out our analysis under the SLQ scheduling. The crucial property in this case is the state space collapse exhibited by the queue length processes. Recall the constants θ_i from (33), that determine the upper limit on the value attained by $\hat{Q}_i^{n,s}$. While in the previous section the threshold of the least priority class, θ_N, was significant, under the current service policy, the property that queue lengths remain equal makes the *minimal* threshold important. Thus, assume that the classes are labelled in such a way that

$$\theta_1 \geq \cdots \geq \theta_N,$$

and let $M = \min\{i\colon \theta_i = \theta_N\}$. We first treat the case $M = N$.

Lemma 4.1. *Assume $M = N$. Fix T.*

(i) *For $i = 1, 2, \ldots, N$ we have*

$$\sup_s \|\hat{Q}_i^{n,s} - N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s})^+\|_T \to 0, \quad \text{in probability, as } n \to \infty,$$

$$\sup_s \|\bar{\Psi}_i^{n,s} - \rho_i\|_T \to 0, \quad \text{in probability, as } n \to \infty.$$

(ii) *For $i = 1, 2, \ldots, N-1$, $\sup_s \hat{R}_i^{n,s}(T) \to 0$, in probability, as $n \to \infty$.*

Proof. Fix $\epsilon > 0$. Let $\varepsilon_1 = \varepsilon/(4(N-1))$ and consider the event

$$\Omega^n = \left\{\hat{Q}_i^n(0) \leq \frac{\epsilon}{8} \text{ and } |\bar{\Psi}^n(0) - \rho_i| \leq \frac{\epsilon_1}{2} \text{ for all } i = 1, \ldots, N\right\}.$$

Then it follows from the assumptions that $\mathbb{P}(\Omega^n) \to 1$. For $s \in \mathfrak{S}_0$ define

$$\begin{aligned}\tau_1^{n,s} = \inf\Big\{t \geq 0\colon\ & \min_i \hat{Q}_i^{n,s}(t) - N^{-1}(1\cdot\hat{X}^{n,s}(t))^+ \leq -\epsilon,\\ & \text{or } |\bar{\Psi}_i^{n,s}(t) - \rho_i| \geq \epsilon_1 \text{ for some } i = 1, \ldots, N-1,\\ & \text{or } |\bar{\Psi}_N^{n,s}(t) - \rho_N| \geq \epsilon\Big\}.\end{aligned}$$

Let $A^{n,s} = \{\tau_1^{n,s} \leq T\}$ and $A^n = \bigcup_s A^{n,s}$. Now let

$$\begin{aligned}A_1^{n,s,i} &= \{\omega \in A^{n,s}\colon \hat{Q}_i^{n,s}(\tau_1^{n,s}) - N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s}(\tau_1^{n,s}))^+ \leq -\epsilon\} \cap \Omega^n, \quad i = 1, \ldots, N,\\ A_2^{n,s,i} &= \Big\{\omega \in A^{n,s}\colon \min_j \hat{Q}_j^{n,s}(\tau_1^{n,s}) - N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s}(\tau_1^{n,s}))^+ > -\epsilon \text{ and } |\bar{\Psi}_i^{n,s}(\tau_1^{n,s}) - \rho_i| \geq \varepsilon_1\Big\} \cap \Omega^n,\\ &\qquad\qquad i = 1, \ldots, N-1,\\ A_3^{n,s} &= \Big\{\omega \in A^{n,s}\colon \min_j \hat{Q}_j^{n,s}(\tau_1^{n,s}) - N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s}(\tau_1^{n,s}))^+ > -\epsilon,\ \max_{j\leq N-1} |\bar{\Psi}_j^{n,s}(\tau_1^{n,s}) - \rho_j| < \varepsilon_1\\ &\qquad\qquad \text{and } |\bar{\Psi}_N^{n,s}(\tau_1^{n,s}) - \rho_N| \geq \epsilon\Big\} \cap \Omega^n.\end{aligned}$$

For $\omega \in A_1^{n,s,i}$, there exists $\eta_1^{n,s}$ such that

$$\hat{Q}_i^{n,s}(\eta_1^{n,s}) - N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s}(\eta_1^{n,s}))^+ > -\frac{\epsilon}{2} \quad \text{and on} \quad I_1^{n,s} := [\eta_1^{n,s}, \tau_1^{n,s}], \qquad \hat{Q}_i^{n,s} - N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s})^+ < 0. \tag{55}$$

Note that $\mathbf{1}\cdot\hat{X}^{n,s} = \mathbf{1}\cdot\hat{Q}^{n,s}$, hence, on the time interval $I_1^{n,s}$, the ith queue length is less than the average. Since the scheduling policy always chooses the longest queue and on this time interval, no customer from class i enters service. Therefore the class-i queue length can only increase during this period. Thus we have

$$N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s})^+[I_1^{n,s}] = N^{-1}(\mathbf{1}\cdot\hat{Q}^{n,s})[I_1^{n,s}] \geq \hat{Q}_i^{n,s}[I_1^{n,s}] + \frac{\epsilon}{2}. \tag{56}$$

Hence $N^{-1}\sum_{j\neq i}\hat{Q}_j^{n,s} \geq \epsilon/2$, and so by the balance equation for $Q^{n,s}$, (6),

$$\frac{\epsilon\sqrt{n}N}{2} \leq \sum_{j\neq i} Q_j^{n,s}[I_1^{n,s}] \leq \sum_{j\neq i} E_j^n[I_1^{n,s}] - \sum_{j\neq i} B_j^{n,s}[I_1^{n,s}]. \tag{57}$$

Since, as argued above, $B_i^{n,s}[I^{n,s}]=0$, it follows that the last term of (57) equals $\mathbf{1}\cdot B^{n,s}[I_1^{n,s}]$, and since $\mathbf{1}\cdot\Psi^{n,s}=n$ on this interval, it follows from (7) that the same term equals $\mathbf{1}\cdot D^{n,s}[I_1^{n,s}]$. The argument from Lemma 3.1 (following (30)) now shows that $\mathbb{P}(\bigcup_s A_1^{n,s,i})\to 0$.

Now we analyze the event $A_2^{n,s,i}$. By (7),

$$\begin{aligned}\bar\Psi_i^{n,s}(t)-\rho_i &= \bar\Psi_i^n(0)-\rho_i+(\bar B_i^{n,s}(t)-\lambda_i t)-\frac{1}{n}\bigg(S_i^n\bigg(\int_0^t n\bar\Psi_i^{n,s}(u)\,du\bigg)-\mu_i\int_0^t n\bar\Psi_i^{n,s}(u)\,du\bigg)\\ &\quad-\mu_i\int_0^t(\bar\Psi_i^{n,s}(u)-\rho_i)\,du\\ &=\bar\Psi_i^n(0)-\rho_i+(\bar E_i^n(t)-\lambda_i t)-\frac{1}{n}\bigg(S_i^n\bigg(\int_0^t n\bar\Psi_i^{n,s}(u)\,du\bigg)-\mu_i\int_0^t n\bar\Psi_i^{n,s}(u)\,du\bigg)\\ &\quad-\mu_i\int_0^t(\bar\Psi_i^{n,s}(u)-\rho_i)\,du-\bar Q_i^{n,s}(t)-\bar R_i^{n,s}(t).\end{aligned}\tag{58}$$

Thus, for $t\in[0,T]$,

$$\begin{aligned}|\bar\Psi_i^{n,s}(t)-\rho_i| &\le |\bar\Psi_i^n(0)-\rho_i|+\|\bar E_i^n-\lambda_i\cdot\|_T+n^{-1/2}\|\hat S_i^n\|_T+\|\bar Q_i^{n,s}\|_T\\ &\quad+\bar R_i^{n,s}(t)+\mu_i\int_0^t|\bar\Psi_i^{n,s}(u)-\rho_i|\,du.\end{aligned}$$

And so by Gronwall's lemma we have

$$|\bar\Psi_i^{n,s}(t)-\rho_i|\le(|\bar\Psi_i^n(0)-\rho_i|+\|\bar E_i^n-\lambda_i\cdot\|_T+n^{-1/2}\|\hat S_i^n\|_T+\|\bar Q_i^{n,s}\|_T+\bar R_i^{n,s}(t))e^{\mu_i T}.$$

Using the identity $(\mathbf{1}\cdot\hat X^{n,s})^+=\mathbf{1}\cdot\hat Q^{n,s}$, we have on $A_2^{n,s,i}$ that $\min_j\hat Q^{n,s}\ge N^{-1}\mathbf{1}\cdot\hat Q^{n,s}-\varepsilon$ up to the time $\tau_1^{n,s}$. As a result, $\max_j\hat Q_j^{n,s}\le N^{-1}\mathbf{1}\cdot\hat Q^{n,s}+N\varepsilon$. Using the fact that the queue length is limited by $\hat Q_N^{n,s}\le\theta_N+2n^{-1/2}$ at all times, it follows that for all large n, up to time $\tau_1^{n,s}$,

$$\max_j\hat Q_j^{n,s}\le\theta_N+(N+1)\epsilon.$$

Hence, if ε is sufficiently small then up to time $\tau_1^{n,s}$ there can be at most one reneging of class-j customers for $j\le N-1$. Thus, on $A_2^{n,s,i}$, we have

$$\varepsilon_1\le|\bar\Psi_i^{n,s}(\tau_1^{n,s})-\rho_i|\le(|\bar\Psi_i^n(0)-\rho_i|+\|\bar E_i^n-\lambda_i\cdot\|_T+n^{-1/2}\|\hat S_i^n\|_T+n^{-1/2}(\theta_i+1)+n^{-1})e^{\mu_i T}.$$

Using the convergence of $\hat E^n$ and $\hat S^n$ (28) and that of the initial condition (18), we therefore obtain $\mathbb{P}(\bigcup_s A_2^{n,s,i})\to 0$.

Finally we analyze $A_3^{n,s}$. We have

$$\bar\Psi_N^{n,s}(\tau_1^{n,s})\le 1-\sum_{i=1}^{N-1}\bar\Psi_i^{n,s}(\tau_1^{n,s})\le\rho_N+\frac{\epsilon}{4}.$$

Thus by the way $A_3^{n,s}$ is defined, we have $\bar\Psi_N^{n,s}(\tau_1^{n,s})\le\rho_N-\epsilon$. And so there exists $\eta_2^{n,s}$ such that

$$\bar\Psi_N^{n,s}(\eta_2^{n,s})>\rho_N-\frac{\epsilon}{2}\quad\text{and on}\quad[\eta_2^{n,s},\tau_1^{n,s}],\qquad\bar\Psi_N^{n,s}(t)<\rho_N-\frac{\epsilon}{4}.\tag{59}$$

Hence on $[\eta_2^{n,s},\tau_1^{n,s}]$, we have $\sum\bar\Psi_i^{n,s}(t)<\sum\rho_i+\epsilon/4-\epsilon/4=1$. Thus, on this interval we have $\mathbf{1}\cdot\hat Q^{n,s}=0$, and so, the argument provided in the last part of the proof of Lemma 3.2 shows $\mathbb{P}(\bigcup_s A_3^{n,s})\to 0$.

We have thus shown that $\mathbb{P}(A^n)\to 0$. The conclusion of item (i) now follows on using again the fact that $\min_j\hat Q_j^{n,s}\ge N^{-1}\mathbf{1}\cdot\hat Q^{n,s}-\varepsilon$ implies $\max_j\hat Q_j^{n,s}\le N^{-1}\mathbf{1}\cdot\hat Q^{n,s}+N\varepsilon$.

As for item (ii), recall that $\theta_N<\theta_i$ for all $i<M=N$. Hence the assertion is a direct consequence of (34) and item (i). □

Next, consider $M\in\{1,2,\dots,N\}$. Fix a sequence k_n, $n\in\mathbb{N}$, such that $\lim n^{-1/2}k_n=\infty$ and $\lim n^{-1}k_n=0$. Given $T<\infty$, define

$$T_{n,s}=\inf\{t\colon\mathbf{1}\cdot R^{n,s}(t)\ge k_n\}\wedge T.$$

We use the notation $U^{*,n,s}=U^{n,s}(\cdot\wedge T_{n,s})$ for any process $U^{n,s}$, and refer to these processes as *stopped versions* of the original processes. The following result states that Lemma 4.1 is valid for the stopped processes.

Lemma 4.2. *Consider general M.*

(i) *For $i = 1, 2, \dots, N$ we have*

$$\sup_s \|\hat{Q}_i^{*,n,s} - N^{-1}(\mathbf{1}\cdot\hat{X}^{*,n,s})^+\|_T \to 0, \quad \textit{in probability, as } n\to\infty,$$

$$\sup_s \|\bar{\Psi}_i^{*,n,s} - \rho_i\|_T \to 0, \quad \textit{in probability, as } n\to\infty.$$

(ii) *For $i = 1, 2, \dots, M-1$, $\sup_s \hat{R}_i^{*,n,s}(T) \to 0$, in probability, as $n\to\infty$.*

Proof. Note that, by definition, $\bar{R}^{*,n,s} = e^{n,s}$. Hence a use of (58) and again Gronwall's lemma immediately give $\bar{\Psi}^{*,n,s} = \rho + e^{n,s}$, proving the second part of item (i) on the lemma. With this at hand, the remaining assertions are proved as in Lemma 4.1. □

In the case where $M = N$, we provide a convergence result. We do not attempt such an analysis for $M < N$, where, as is shown in a work in progress (Atar and Saha [4]), the limiting behaviour may depend on properties that are finer than first and second order data. Thus, for $M < N$, we only obtain C-tightness of the processes, that however will suffice for the purpose of proving the main result.

To present the result regarding the case $M = N$, we consider an SDE of the form (39) with different domain G and drift b. Namely, we consider

$$G = \{y\in\mathbb{R}^N: \mathbf{1}\cdot y \le N\theta_N\},$$

and $b: \mathbb{R}^N \to \mathbb{R}^N$ given by

$$b(y) = -\big(\mu_1(y_1 - N^{-1}(\mathbf{1}\cdot y)^+), \dots, \mu_N(y_N - N^{-1}(\mathbf{1}\cdot y)^+)\big). \tag{60}$$

The process $W(t)$ is as in Section 3, and the SDE of interest is now

$$\begin{gathered} X(t) = X_0 + W(t) + \int_0^t b(X(u))\,du - L(t)\mathbf{e}_N, \quad t\ge 0,\\ \int_{[0,\infty)} 1_{\{\mathbf{1}\cdot X(t) < N\theta_N\}}\,dL(t) = 0, \end{gathered} \tag{61}$$

where a solution (X, L) is defined similarly. The map $\Gamma: \mathbb{D}_{\mathbb{R}^N}([0,T]) \to \mathbb{D}_{\mathbb{R}^N}([0,T])$ that is relevant for the present setting is given by

$$\Gamma(f)(t) = f(t) - g(t)\mathbf{e}_N, \quad g(t) = \sup_{0\le u\le t}(N\theta_N - (\mathbf{1}\cdot f(u)))^-.$$

Proposition 4.3. (i) *For general M, the processes $\hat{W}^{n,s}$, $\hat{X}^{n,s}$, $\hat{R}^{n,s}$, $\hat{Q}^{n,s}$ and $\hat{\Psi}^{n,s}$ are C-tight, uniformly in s.*

(ii) *In the case $M = N$, as $n\to\infty$, $(\hat{W}^{n,0}, \hat{X}^{n,0}, \hat{R}^{n,0}, \hat{Q}^{n,0}, \hat{\Psi}^{n,0})$ converges in distribution to $(W, X, L\mathbf{e}_N, Q, \Psi)$, where (X, L) form the solution to the SDE (61), and*

$$Q = N^{-1}(\mathbf{1}\cdot X)^+\sum_{i=1}^N \mathbf{e}_i, \quad \Psi = X - Q.$$

Proof. *Step* 1. In this and the next step we consider the case $M = N$. We have

$$\begin{aligned} \hat{X}_i^{n,s} &= \hat{X}_i^n(0) + \hat{W}_i^{n,s} - \mu_i\int_0^{\cdot} \hat{\Psi}_i^{n,s}(u)\,du - \hat{R}_i^{n,s}\\ &= \hat{X}_i^n(0) + \hat{W}_i^{n,s} - \mu_i\int_0^{\cdot} (\hat{X}_i^{n,s}(u) - \hat{Q}_i^{n,s}(u))\,du - \hat{R}_i^{n,s}\\ &= \hat{X}_i^n(0) + \hat{W}_i^{n,s} - \mu_i\int_0^{t} (\hat{X}_i^{n,s}(u) - N^{-1}(1\cdot\hat{X}^{n,s}(u))^+)\,du - \hat{R}_i^{n,s} + e_i^{n,s}, \end{aligned}$$

where we have used Lemma 4.1(i) on the last line. Next, by Lemma 4.1(ii),

$$\hat{X}^{n,s} = \hat{X}^n(0) + \hat{W}^{n,s} + \int_0^{\cdot} b(\hat{X}^{n,s}(u))\,du - \hat{R}_N^{n,s}\mathbf{e}_N + e^{n,s}, \tag{62}$$

with b as in (60). Define

$$Z_i^{n,s} = \hat{X}_i^{n,s} + \hat{Q}_N^{n,s} - N^{-1}(\mathbf{1}\cdot\hat{X}^{n,s})^+, \quad i = 1,\dots,N,$$

and note that $Z^{n,s} = \hat{X}^{n,s} + e^{n,s}$. Let

$$K^{n,s} = [N^{-1}(\mathbf{1}\cdot Z^{n,s})] \wedge \theta_N - N^{-1}(\mathbf{1}\cdot Z^{n,s}).$$

Since

$$\begin{aligned} N^{-1}(\mathbf{1}\cdot Z^{n,s}) &= N^{-1}(\mathbf{1}\cdot \hat{X}^{n,s}) + \hat{Q}_N^{n,s} - N^{-1}(\mathbf{1}\cdot \hat{X}^{n,s})^+ \\ &= N^{-1}(\mathbf{1}\cdot \hat{\Psi}^{n,s}) + \hat{Q}_N^{n,s} \\ &\le \hat{Q}_N^{n,s} \le \theta_N + 2n^{-1/2}, \end{aligned}$$

we have $K^{n,s} = e^{n,s}$. Define $\tilde{Z}_i^{n,s} = Z_i^{n,s} + K^{n,s}$, $i = 1, 2, \ldots, N$. Then

$$N^{-1}(\mathbf{1}\cdot \tilde{Z}^{n,s})(t) \le \theta_N, \quad t \ge 0. \tag{63}$$

Moreover, $\tilde{Z}^{n,s} = \hat{X}^{n,s} + e^{n,s}$, hence by the Lipschitz property of b and (62),

$$\tilde{Z}^{n,s} = \hat{X}^n(0) + \hat{W}^{n,s} + \int_0^{\cdot} b(\tilde{Z}^{n,s}(u))\,du - \hat{R}_N^{n,s}\mathbf{e}_N + e^{n,s}. \tag{64}$$

As in the case of FP, an argument based on the fact that under the reference scenario no class-i reneging occurs when $\hat{Q}_i^{n,0} < \theta_i$ shows that

$$\int 1_{\{N^{-1}(1\cdot\tilde{Z}^{n,s})<\theta_N\}}\, d\tilde{R}_N^{n,s} = 0, \tag{65}$$

for a nonnegative, nondecreasing process $\tilde{R}_N^{n,s}$ that is close to $\hat{R}_N^{n,s}$ in the sense

$$\tilde{R}_N^{n,s} = \hat{R}_N^{n,s} + e^{n,s}. \tag{66}$$

Step 2. To prove (i) (with $M = N$) and (ii), combine (63), (64) (with $\hat{R}_N^{n,s}$ replaced by $\tilde{R}_N^{n,s}$) and (65) to write

$$\tilde{Z}^{n,s} = \Gamma\left(\hat{X}^n(0) + \hat{W}^{n,s} + \int_0^{\cdot} b(\tilde{Z}^{n,s}(u))\,du + e^{n,s}\right), \tag{67}$$

$$\tilde{R}^{n,s}\mathbf{e}_N = (I - \Gamma)\left(\hat{X}^n(0) + \hat{W}^{n,s} + \int_0^{\cdot} b(\tilde{Z}^{n,s}(u))\,du + e^{n,s}\right). \tag{68}$$

The completion of the proof, based on the above, is precisely as in Proposition 3.3.

Step 3. It remains to prove (i) for $M < N$. We start by arguing that conclusions analogous to those obtained in Step 1 are valid here too, but for the stopped processes. Indeed, working as in Step 1 with Lemma 4.2 in place of Lemma 4.1 shows that

$$\hat{X}^{*,n,s} = \hat{X}^n(0) + \hat{W}^{*,n,s} + \int_0^{\cdot\wedge T_{n,s}} b(\hat{X}^{n,s}(u))\,du - \sum_{i=M}^{N} \hat{R}_i^{*,n,s}\mathbf{e}_i + e^{n,s}.$$

Define

$$Z_i^{*,n,s} = \hat{X}_i^{n,s} + \max_{i\in\{M,\ldots,N\}} \hat{Q}_N^{n,s} - N^{-1}(\mathbf{1}\cdot \hat{X}^{n,s})^+, \qquad i = 1, \ldots, N$$

and

$$K^{*,n,s} = [N^{-1}(\mathbf{1}\cdot Z^{*,n,s})] \wedge \theta_N - N^{-1}(\mathbf{1}\cdot Z^{*,n,s}).$$

Now with $\tilde{Z}^{*,n,s}$ defined as $\tilde{Z}_i^{*,n,s} = Z_i^{*,n,s} + K^{*,n,s}$, $i = 1, \ldots, N$, it can be argued as in step 1 that,

$$\tilde{Z}^{*,n,s} = \hat{X}^n(0) + \hat{W}^{*,n,s} + \int_0^{\cdot\wedge T_{n,s}} b(\tilde{Z}^{n,s}(u))\,du - \sum_{i=M}^{N} \hat{R}_i^{*,n,s}\mathbf{e}_i + e^{n,s}, \tag{69}$$

$$\tilde{Z}^{*,n,s} = \hat{X}^{*,n,s} + e^{n,s}, \tag{70}$$

$$\int 1_{\{N^{-1}(\mathbf{1}\cdot\tilde{Z}^{n,s})<\theta_N\}}\, d\tilde{R}_i^{n,s} = 0, \quad i = M, \ldots, N, \tag{71}$$

for nonnegative, nondecreasing processes $\tilde{R}_i^{n,s}$ that are close to $\hat{R}_i^{n,s}$ in the sense

$$\tilde{R}_i^{n,s} = \hat{R}_i^{n,s} + e^{n,s}, \quad i = M, \ldots, N, \tag{72}$$

(note that the above refers to the unstopped versions of the processes, because again (50) is valid).

Denote

$$\zeta^{n,s} = \mathbf{1}\cdot\tilde{Z}^{n,s}, \qquad \xi^{n,s} = \mathbf{1}\cdot\hat{X}^n(0) + \mathbf{1}\cdot\hat{W}^{n,s} + \int_0^{\cdot} \mathbf{1}\cdot b(\tilde{Z}^{n,s}(u))\,du, \qquad \rho^{n,s} = \sum_{i=M}^{N} \tilde{R}_i^{n,s}. \tag{73}$$

Then $\xi^{n,s}$ and $\rho^{n,s}$ have sample paths in $\mathbb{D}_{\mathbb{R}}$, where those of $\rho^{n,s}$ are nonnegative and nondecreasing, and moreover, as follows from (63), (69), (71), and (72),

$$\zeta^{*,n,s} = \xi^{*,n,s} + e^{n,s} - \rho^{*,n,s} \le N\theta_N, \qquad \int_{[0,\infty)} 1_{\{\zeta^{n,s} < N\theta_N\}}\, d\rho^{n,s} = 0.$$

It follows that $\rho^{*,n,s}$ is given by

$$\rho^{*,n,s}(t) = \sup_{0\le u\le t} (N\theta_N - \xi^{*,n,s}(u) + e^{n,s}(u))^-. \tag{74}$$

We now write c for generic constants and use the Lipschitz property of b. We have

$$\begin{aligned}\rho^{*,n,s}(t) &\le c + \|\xi^{*,n,s}\|_t + e^{n,s}(t)\\ &\le c + \|\hat{X}^n(0)\| + c\|\hat{W}^{*,n,s}\|_t + c\int_0^{t\wedge T_{n,s}} \|\tilde{Z}^{n,s}(u)\|\,du + e^{n,s}(t).\end{aligned}$$

Going back to (69) and recalling that $\rho^{n,s}$ has been defined as the sum of positive terms,

$$\|\tilde{Z}^{*,n,s}(t)\| \le c\|\hat{X}^n(0)\| + c\|\hat{W}^{*,n,s}\|_t + c\int_0^t \|\tilde{Z}^{*,n,s}(u)\|\,du + e^{n,s}(t).$$

A use of Gronwall's lemma now shows that for T fixed, $\|\tilde{Z}^{*,n,s}\|_T$, $n\in\mathbb{N}$, are tight, uniformly in s. Next, using (73) and the C-tightness of $\hat{W}^{*,n,s}$ shows that $\xi^{*,n,s}$ are C-tight, uniformly in s. In turn, using (74), shows that so are the processes $\rho^{*,n,s}$. In particular, for fixed T,

$$\rho^{*,n,s}(T) \quad \text{are tight uniformly in } s. \tag{75}$$

Now, note that

$$1\cdot\hat{R}^{*,n,s}(T) = \sum_{i=1}^{M-1} \hat{R}_i^{*,n,s}(T) + \rho^{*,n,s}(T) + e^{n,s} = \rho^{*,n,s}(T) + e^{n,s},$$

where we used Lemma 4.2(ii). Hence in view of (75), the definition of $T_{n,s}$, and the assumption $\lim n^{-1/2}k_n = \infty$, we have $\mathbb{P}(\text{for some } s,\ T_{n,s} < T) \to 0$ as $n\to\infty$. Thus all conclusions we have obtained for the stopped processes are valid for the unstopped versions. Namely, $\|\tilde{Z}^{n,s}\|_T$ are tight, uniformly in s, $\xi^{n,s}$ and $\rho^{n,s}$ are C-tight uniformly in s, and (69) and (70) hold without the asterisk sign.

Using the last part of (73) and the fact that each of the processes $\tilde{R}_i^{n,s}$, $n\in\mathbb{N}$, $i = M,\dots,N$, is nondecreasing shows that these processes are also C-tight, uniformly in s. Hence by (69), $\tilde{Z}^{n,s}$, and in turn, $\hat{X}^{n,s}$ are C-tight, uniformly in s. Finally, Lemma 4.2 is now valid for the processes without the asterisk sign. Thus the uniform C-tightness of $\hat{Q}^{n,s}$ follows from that of $\hat{X}^{n,s}$ upon using Lemma 4.2(i) and the continuous mapping theorem, and that of $\hat{\Psi}^{n,s}$ follows from the identity (27). □

5. Reiman's snapshot principle and proof of the main result.

We finally state and prove RSP and obtain the main result as an immediate consequence thereof. RSP is based on the C-tightness of the processes $B^{n,s}$, established as part of the limit results above. The two policies, namely, FP and SLQ, are addressed here simultaneously.

The proof uses the following identity, that holds regardless of the service policy:

$$\hat{Q}_i^{n,s}(\mathrm{JT}_i^{n,s}(t)) = \hat{B}_i^{n,s}(\mathrm{JT}_i^{n,s}(t) + \mathrm{WT}_i^{n,s}(t)) - \hat{B}_i^{n,s}(\mathrm{JT}_i^{n,s}(t)) + \lambda_i\widehat{\mathrm{WT}}_i^{n,s}(t), \tag{76}$$

and on properties of the processes involved in it. This identity follows from (15), and the definition of the scaled processes, (16) and (17). The main argument is that the l.h.s. and the last term on the r.h.s. must be asymptotically equal once one has that $\hat{B}^{n,s}$ are uniformly C-tight and the term $\widehat{\mathrm{WT}}^{n,s}$ is small.

Proposition 5.1. *We have for* $i = 1,\dots,N$,

$$\gamma_i^n(T) := \sup_s \sup_{t\in[0,T]} \big|\hat{Q}_i^{n,s}(\mathrm{JT}_i^{n,s}(t)) - \lambda_i\widehat{\mathrm{WT}}_i^{n,s}(t)\big| \to 0, \quad \textit{in probability, as } n\to\infty. \tag{77}$$

Proof. First we argue that the results of Sections 3 and 4 imply that $\hat{B}^{n,s}$ are C-tight, uniformly in s. Indeed, by (25),

$$\hat{B}_i^{n,s}(t) = \hat{Q}_i^n(0) + \hat{E}_i^n(t) + \hat{\lambda}_i t - \hat{Q}_i^{n,s}(t) - \hat{R}_i^{n,s}(t) + e^{n,s}(t).$$

By (18) and (28), the sum of the first two terms forms a C-tight sequence of processes. By Proposition 3.3, $\hat{Q}_i^{n,s}$ and $\hat{R}_i^{n,s}$ are C-tight, uniformly in s, under FP, and by Proposition 4.3, the same is true under SLQ. Thus follows the uniform C-tightness of $\hat{B}^{n,s}$, and in particular, for $i = 1, 2, \dots, N$ and $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \mathbb{P}\Big(\sup_s w_{T+2}(\hat{B}_i^{n,s}, \delta) > \epsilon\Big) \to 0. \tag{78}$$

Fix $\varepsilon \in (0,1)$ and define

$$\Omega_i^{n,s} = \Big\{\sup_{t\in[0,T]} |\mathrm{JT}_i^{n,s}(t) - t| > \epsilon\Big\}.$$

Then, on $\Omega_i^{n,s}$ there exists $t \in [0,T]$ such that $J_i^{n,s}(t+\epsilon) - J_i^{n,s}(t) = 0$, hence

$$J_i^{n,s}(t+\epsilon) - n\lambda_i(t+\epsilon) - [J_i^{n,s}(t) - n\lambda_i t] = -n\lambda_i\epsilon.$$

Hence, on $\bigcup_s \Omega_i^{n,s}$,

$$\sup_s \sup_{0\le t\le T+1} |J_i^{n,s}(t)/n - \lambda_i t| + \sup_s \sup_{0\le t\le T} |J_i^{n,s}(t)/n - \lambda_i t| \ge \lambda_i \epsilon.$$

By (5), $J^{n,s} = E^n - R^{n,s}$, and therefore by the tightness of $\|\hat{E}^n\|_{T+1}$ and $\|\hat{R}^{n,s}\|_{T+1}$, uniformly in s, we have that

$$\sup_s \sup_{0\le t\le T+1} \left|\frac{J_i^{n,s}(t) - n\lambda_i t}{\sqrt{n}}\right|$$

are tight. Hence

$$\mathbb{P}\Big(\sup_s \sup_{0\le t\le T} |\mathrm{JT}_i^{n,s}(t) - t| > \epsilon\Big) \to 0, \quad \text{as } n\to\infty. \tag{79}$$

Next we show

$$\mathbb{P}\Big(\sup_s \sup_{0\le t\le T} \mathrm{WT}_i^{n,s}(t) > 1\Big) \to 0, \quad \text{as } n\to\infty. \tag{80}$$

For every ω in the event under consideration there exist t and s such that $\mathrm{WT}_i^{n,s}(t) > 1$. Therefore, by (15),

$$\begin{aligned} Q_i^{n,s}(\mathrm{JT}_i^{n,s}(t)) &= B_i^{n,s}\big(\mathrm{JT}_i^{n,s}(t) + \mathrm{WT}_i^{n,s}(t)\big) - B_i^{n,s}(\mathrm{JT}_i^{n,s}(t)) \\ &\ge B_i^{n,s}(\mathrm{JT}_i^{n,s}(t) + 1) - B_i^{n,s}(\mathrm{JT}_i^{n,s}(t)), \end{aligned}$$

thus

$$\hat{Q}_i^{n,s}(\mathrm{JT}_i^{n,s}(t)) \ge \hat{B}_i^{n,s}(\mathrm{JT}_i^{n,s}(t)+1) - \hat{B}_i^{n,s}(\mathrm{JT}_i^{n,s}(t)) + \lambda_i\sqrt{n}.$$

The conclusion follows using (79) and the tightness of the r.v.s $\sup_s \|\hat{Q}^{n,s}\|_{T+1}$ and $\sup_s \|\hat{B}^{n,s}\|_{T+2}$, $n \in \mathbb{N}$.

Using (76), the tightness of the r.v.s $\sup_s \|\hat{Q}^{n,s}\|_{T+1}$ and $\sup_s \|\hat{B}^{n,s}\|_{T+2}$ and the facts (79) and (80), gives that of $\sup_s \|\widehat{\mathrm{WT}}^{n,s}\|_T$. As a result, $\mathrm{WT}^{n,s} = e^{n,s}$. Using (76) again shows that $\gamma_i^n(T)$ of (77) satisfies

$$\gamma_i^n(T) \le \sup_s w_{T+2}(\hat{B}^{n,s}, \delta)$$

on the event $\{\sup_s \sup_{t\le T} (\mathrm{JT}_i^{n,s}(t) + \mathrm{WT}_i^{n,s}(t)) \le T+2\} \cap \{\sup_s \mathrm{WT}_i^{n,s} < \delta\}$. Since we have just argued that the probability of this event converges to 1 as $n \to \infty$, the result follows from (78). □

Finally, we provide the proof of our main result, as a direct consequence of Proposition 5.1.

Proof of Theorem 2.1. Let $\tilde{\Omega}^n$ be the event defined by (23). Fix $(i,j) \in \mathfrak{S}$. Then if

$$C_{ij}^n(\sigma_{ij}^n, \sigma^{n,ij}) > C_{ij}^n(\bar{\sigma}_{ij}^n, \sigma^{n,ij}) + \epsilon, \tag{81}$$

we have by (20), that $\mathrm{AT}_{ij}^n \le \bar{T}$. Now, there can be two cases.

Case 1. $h_i(\lambda_i^{-1}Q_i^{n,0}(\mathrm{AT}_{ij}^n-)) < r_i$. Then by (22), $\Delta_i^n(j)=1$, hence by (20), $C_{ij}^n(\sigma_{ij}^n, \sigma^{n,ij}) = h_i(\widehat{\mathrm{WT}}_{ij}^{n,0})$, whereas $C_{ij}^n(\bar{\sigma}_{ij}^n, \sigma^{n,ij}) = r_i$. Thus $h_i(\widehat{\mathrm{WT}}_{ij}^{n,0}) > r_i + \epsilon$, and so

$$h_i(\widehat{\mathrm{WT}}_{ij}^{n,0}) - \epsilon > r_i > h_i\left(\frac{\hat{Q}_i^{n,0}(\mathrm{AT}_{ij}^n-)}{\lambda_i}\right).$$

Since $\hat{Q}_i^{n,0}$ is bounded by $\theta_i + 1$, it follows that

$$\begin{aligned}\sum_{k=1}^N \frac{\gamma_k^n(T)}{\lambda_k} + \frac{1}{\sqrt{n}} &\geq \widehat{\mathrm{WT}}_i^{n,0}(\mathrm{AT}_{ij}^n) - \frac{\hat{Q}_i^{n,0}(\mathrm{JT}_i^{n,s}(\mathrm{AT}_{ij}^n))}{\lambda_i} + \frac{1}{\sqrt{n}} = \widehat{\mathrm{WT}}_{ij}^{n,0} - \frac{\hat{Q}_i^{n,0}(\mathrm{AT}_{ij}^n-)}{\lambda_i} \\ &\geq \inf\{b-a\colon h(b)-h(a) > \varepsilon,\ a \in [0, \lambda_i^{-1}(\theta_i+1)], b \geq 0\} > 0, \end{aligned} \tag{82}$$

by the continuity of h.

Case 2. $h_i(\lambda_i^{-1}\hat{Q}_i^{n,0}(\mathrm{AT}_{ij}^n-)) \geq r_i$. In this case, by (22) $\Delta_i^n(j)=0$, by (20), $C_{ij}^n(\sigma_{ij}^n, \sigma^{n,ij}) = r_i$ and $C_{ij}^n(\bar{\sigma}_{ij}^n, \sigma^{n,ij}) = h_i(\widehat{\mathrm{WT}}_{ij}^{n,s})$. Hence $h_i(\widehat{\mathrm{WT}}_{ij}^{n,s}) < r_i - \epsilon$, and so

$$h_i(\widehat{\mathrm{WT}}_{ij}^{n,s}) + \epsilon < r_i \leq h_i\left(\frac{\hat{Q}_i^{n,0}(\mathrm{AT}_{ij}^n-)}{\lambda_i}\right).$$

As a result,

$$\begin{aligned}\sum_{k=1}^N \frac{\gamma_k^n(T)}{\lambda_k} &\geq \frac{\hat{Q}_i^{n,0}(\mathrm{JT}_i^{n,s}(\mathrm{AT}_{ij}^n))}{\lambda_i} - \frac{1}{\sqrt{n}} - \widehat{\mathrm{WT}}_i^{n,s}(\mathrm{AT}_{ij}^n) = \frac{\hat{Q}_i^{n,s}(\mathrm{AT}_{ij}^n-)}{\lambda_i} - \widehat{\mathrm{WT}}_{ij}^{n,s} \\ &\geq \inf\left\{b-a\colon h(b)-h(a) > \varepsilon,\ b \in [0, \lambda_i^{-1}(\theta_i+1)], a \geq 0\right\} > 0, \end{aligned} \tag{83}$$

by the continuity and strict monotonicity of h.

Combining (82) and (83) shows that if (81) holds for *some* $(i, j) \in \mathfrak{S}$, then

$$\sum_{k=1}^N \frac{\gamma_k^n(T)}{\lambda_k} \geq c > 0,$$

where c is a constant that does not depend on n. Using Proposition 5.1 shows that $\mathbb{P}((\tilde{\Omega}^n)^c) \to 0$ as $n \to \infty$. This completes the proof. □

6. Concluding remarks.

This paper combines game theoretic analysis with heavy traffic theory. It addresses a particular queueing model and two families of scheduling disciplines, leaving room for various extensions. Let us briefly mention some.

As already mentioned, the proof of the main result provided above is based solely on Proposition 5.1. This proof shows that the sequence $\{\sigma^n\}$ of (22) is an ε-Nash equilibrium w.h.p. under *any scheduling policy* for which RSP holds. At the same time, as stated in Remark 2.4, RSP does not hold for arbitrary policies. It is therefore of interest to ask to what degree RSP (and consequently our main result) can be extended to cover a larger collection of scheduling policies. Another problem, already mentioned in Remark 2.3, is to obtain our main result without assuming that customers have access to any of the system parameters. One may assume, for example, that customers know how long other customers have waited.

While the latter problem is relevant for the scheduling policies treated in this paper, it is of interest to study both problems with regard to a larger class of scheduling policies. Specific ones that we find natural are the following.

1. When a server becomes available it chooses a customer class uniformly at random from those with nonempty buffers. Like FP, this policy does not use queue length information (beyond the information of which buffers are nonempty), and in a sense is at the other extreme as far as fairness is concerned.

2. When a server becomes available it chooses a customer of class i at random with probability $Q_i/(\sum_j Q_j)$. This can be thought of as a randomized version of SLQ.

Beyond the desired extensions alluded to above for the queueing model under consideration, it would be of interest to implement an approach that combines game theory and heavy traffic theory, such as the one proposed here, in the setting of more general queueing networks.

Acknowledgments. The authors are grateful to two referees for their helpful comments and to the associate editor for referring them to the article Allon and Gurvich [2]. Research was supported in part by the Israel Science Foundation [Grant 1315/12].

References

[1] Aksin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

[2] Allon G, Gurvich I (2010) Pricing and dimensioning competing large-scale service providers. *Manufacturing Service Oper. Management* 12(3):449–469.

[3] Anderson R, Orey S (1976) Small random perturbations of dynamical systems with reflecting boundary. *Nagoya Math. J.* 60:189–216.

[4] Atar R, Saha S (2016) A note on non-existence of diffusion limits for serve-the-longest-queue when the buffers are equal in size. *Electron. Commun. Probab.* 21(2):article 2.

[5] Atar R, Shifrin M (2014) An asymptotic optimality result for the multiclass queue with finite buffers in heavy traffic. *Stochastic Systems* 4:556–603.

[6] Atar R, Cidon I, Shifrin M (2014) MDP based optimal pricing for a cloud computing queueing model. *Performance Eval.* 78:1–6.

[7] Billingsley P (1999) *Convergence of Probability Measures*, Wiley Series in Probability and Statistics: Probability and Statistics, 2nd ed. (John Wiley & Sons, New York).

[8] Dupuis P, Ishii H (1991) On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. *Stochastics Stochastics Rep.* 35:31–62.

[9] Gopalakrishnan R, Doroudi S, Ward A, Wierman A (2016) Routing and staffing when servers are strategic. *Oper. Res.* 64(4):1033–1050.

[10] Guo P, Hassin R (2011) Strategic behavior and social optimization in Markovian vacation queues. *Oper. Res.* 59(4):986–997.

[11] Gurvich I, Whitt W (2009) Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* 34(2):363–396.

[12] Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.

[13] Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, International Series in Operations Research and Management Science, Vol. 59 (Kluwer Academic Publishers, Boston).

[14] Jacod J, Shiryaev AN (1987) *Limit Theorems for Stochastic Processes, Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*], Vol. 288 (Springer, Berlin).

[15] Manou A, Economou A, Karaesmen F (2014) Strategic customers in a transportation station: When is it optimal to wait? *Oper. Res.* 62(4):910–925.

[16] Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

[17] Reiman MI (1982) The heavy traffic diffusion approximation for sojourn times in Jackson networks. Disney RL, Ott TJ, eds. *Applied Probability—Computer Science: The Interface*. Progress in Computer Science, Vol. 3 (Birkhäuser, Boston), 409–421.

[18] Zhan D, Ward A (2015) Incentive based service system design: Staffing and compensation to trade off speed and quality. SSRN: https://ssrn.com/abstract=2568007.