MATHEMATICS OF OPERATIONS RESEARCH
Vol. 33, No. 3, August 2008, pp. 561–586
ISSN 0364-765X | EISSN 1526-5471 | 08 | 3303 | 0561

DOI 10.1287/moor.1070.0310

Queues with Many Servers: The Virtual Waiting-Time Process in the QED Regime

Avishai Mandelbaum
Industrial Engineering and Management, Technion, Haifa 3200, Israel,
avim@tx.technion.ac.il

Petar Momčilović
Department of Electrical Engineering and Computer Science, University of Michigan,
Ann Arbor, Michigan 48109, petar@eecs.umich.edu

We consider a first-come first-served multiserver queue in the Quality- and Efficiency-Driven (QED) regime. In this regime, which was first formalized by Halfin and Whitt, the number of servers N is not small, servers' utilization is $1-O(1/\sqrt{N})$ (Efficiency-Driven) while waiting time is $O(1/\sqrt{N})$ (Quality-Driven). This is equivalent to having the number of servers N being approximately equal to $R+\beta\sqrt{R}$, where R is the offered load and β is a positive constant.

For the G/D$_K$/N queue in the QED regime, we analyze the virtual waiting time $V_N(t)$, as N increases indefinitely. Assuming that the service-time distribution has a finite support (hence the D$_K$ in G/D$_K$/N), it is shown that, in the limit, the scaled virtual waiting time $\widehat{V}_N(t)=\sqrt{N}V_N(t)/\mathbb{E}S$ is representable as a supremum over a random weighted tree (S denotes a service time). Informally, it is then argued that, for large N,

$$\widehat{V}_N(t)\approx(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N(t-S)]+\widehat{X}(t)-\beta)^+,\quad t\in\mathbb{R};$$

here $\mathbb{E}_{\langle S\rangle}[\widehat{V}_N(t-S)]$ is the averaging of $\widehat{V}_N(t-S)$ over S, and the process $\widehat{X}(t)$ is zero-mean Gaussian that summarizes all relevant information about arrivals and service times ($\widehat{X}$ arises as a limit of an *infinite*-server (G/D$_K$/∞) process, appropriately scaled). The results are obtained by using both combinatorial and probabilistic arguments. Possible applications of our approximations include fast simulation of queues and estimation/prediction of customer waiting times in the QED regime.

Key words: multiserver queue; heavy-traffic; quality- and efficiency-driven (QED) or Halfin–Whitt regime; economies of scale; telephone call or contact centers
MSC2000 subject classification: Primary: 60K25; secondary: 90B22
OR/MS subject classification: Primary: queues, limit theorems, multichannel; secondary: probability, stochastic model applications
History: Received August 30, 2005; revised January 19, 2007, and August 29, 2007.

1. Introduction. The interrelation between service efficiency (as manifested through resources' utilization) and service quality (as perceived by users) is the key trade-off in dimensioning queueing systems. Very often high utilization is achieved at the cost of low service quality (e.g., frequent long delays). When considering a single resource, this trade-off is inherent and essentially cannot be avoided. However, economies of scale come to the rescue here: large-scale service systems can operate in a regime where both objectives, efficiency and quality, do coexist. This regime, which we describe as Quality and Efficiency Driven (QED), is the subject of our paper.

QED Example. The advantage of multiserver systems over single-server systems is well illustrated in Figure 1. For demonstration purposes, we are considering the M/M/N system, which is the only multiserver queue for which explicit formulas for its performance measures are available. On the left, we plot a typical relationship between the probability of delay and utilization in the M/M/N system, for $N=1$, 4, 16, and 64. On the right, we plot the expected delay (given delay) as a function of utilization for the same sequence of systems. As seen from the figure, increasing the number of servers enables one to operate in the QED regime, which corresponds to the lower-right corner of the plots: here, high efficiency (servers' utilization) and high service quality (low probability of delay, as well as short delays) do indeed coexist. To be concrete, with $N=1$, at least 50% of the customers will be served without delay if the server's utilization does not exceed 50%; with $N=64$, on the other hand, this same service level is achieved with servers' utilization even exceeding 93%.

Our Results. In the present paper, we analyze the virtual waiting-time process $V_N=\{V_N(t),t\geq 0\}$, where $V_N(t)$ is the amount of time (beyond t) required until one of the N servers becomes idle. By convention, we take V_N to be right-continuous. Our focus is the behavior of V_N in the QED regime. To this end, we consider a sequence of single-class, first-come-first-served (FCFS) queues that is indexed by $N\uparrow\infty$. The queues operate in the QED regime, or, formally, the traffic intensity in the Nth system is $1-\beta/\sqrt{N}+o(1/\sqrt{N})$, for some

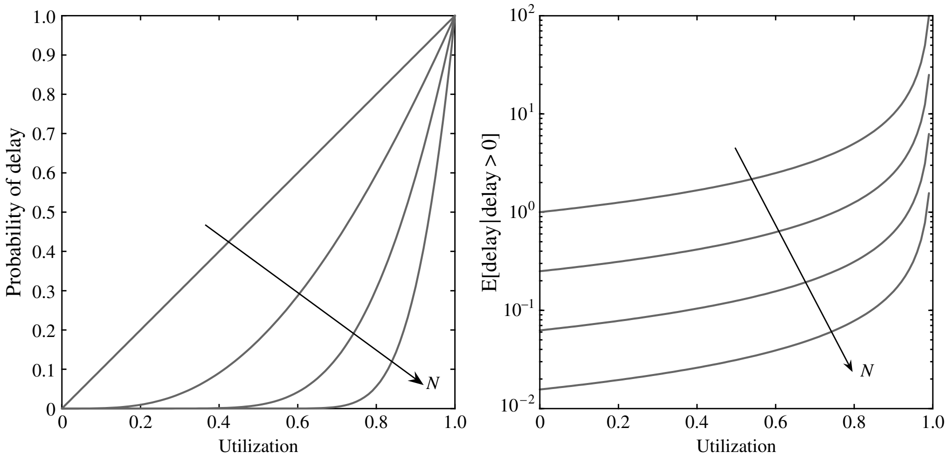

FIGURE 1. Performance of a single-server system (M/M/1) and multiple servers (M/M/N, for $N = 4$, 16, and 64), as a function of servers' utilization: Probability of delay is in the left plot and expected delay (given delay) is on the right. The plots depict economies of scale: given a quality-of-service requirement, one can achieve it with an increasingly higher server utilization, as the number of severs N increases; equivalently, fixing servers' utilization and increasing N improved service level.

constant $\beta > 0$. This is equivalent to having the number of servers N being equal to $R_N + \beta\sqrt{R_N} + o(\sqrt{R_N})$, where R_N is the offered load of the Nth system (arrival rate multiplied by mean service time). Because $N > R_N$ ($\beta > 0$) must hold to ensure stability, $N \approx R_N + \beta\sqrt{R_N}$ is referred to as square-root *safety* staffing.

We fix a service-time distribution and assume that its *support is a set of finite cardinality*. Let S represent a generic service time. Our main result is Theorem 3.1, where it is shown that, for the above sequence of queues, as $N \uparrow \infty$, the scaled virtual waiting time $\hat{V}_N(t) := \sqrt{N} V_N(t)/\mathbb{E}S$ can be represented as a supremum over a random weighted tree. The weights of nodes in the tree are defined in (24), by the values of a zero-mean Gaussian process $\hat{X}$ (see (10)), jointly with a specification of the virtual waiting time process on some initial time interval (the length of which is at least the largest value that S can take). The process $\hat{X}$ summarizes all the information about arrivals and services that is asymptotically relevant. This summary is carried out through a limit of a sequence of *infinite*-server queues, the Nth element of which has the same arrivals and services as our original queue, for all $N = 1, 2, \ldots$: indeed, $\hat{X}$ is the weak limit, as $N \uparrow \infty$, of the number of customers $\hat{X}_N$ in the Nth infinite-server processes (8), appropriately scaled. Informally, it is then argued that

$$\hat{V}_N(t) \approx (\mathbb{E}_{\langle S\rangle}[\hat{V}_N(t-S)] + \hat{X}(t) - \beta)^+$$

for large N; here $\mathbb{E}_{\langle S\rangle}[\hat{V}_N(t-S)]$ is the averaging of $\hat{V}_N(t-S)$ over the distribution of S (see (20) for a precise definition). Possible applications of this last representation include fast simulation of queues and estimation/prediction of customer waiting times in the QED regime (§4). The representation prevails, in fact, for exponential service times. Hence, it is conjectured to hold also for infinite-support services (§3.8).

QED vs. conventional heavy traffic. The subtlety of the QED regime can be demonstrated by examining an M/D$_2$/100/FCFS queue with two-valued service times (Mandelbaum and Schwartz [28]). Let S take values $(1-\varepsilon)$ and $(1+\varepsilon^{-1})$ with such probabilities that both its mean and standard deviation are equal to 1. As demonstrated in Mandelbaum and Schwartz [28] (via simulation), for the QED regime (servers' utilization, say, around 95%), the M/D$_2$/100 queue with small ε (such that the service times are $\approx$1 and $\approx$100) performs like M/D$_1$/100 with service times being approximately 1. The reason is that the one server typically busy with the very long service time is no obstacle for the many other servers to perform as a system with the short (deterministic) service time. This is in stark contrast with conventional heavy-traffic approximations, under which the M/D$_2$/100 queue with the described two-valued service times is expected to perform like M/M/100 rather than M/D$_1$/100. (Indeed, M/D$_2$/1 with our two-valued service times does, at 95% server's utilization, perform like a corresponding M/M/1, and similarly M/D$_2$/N for small N.)

QED relevance. Current interest in the QED regime stems from the expansion of telephone-based services, provided by telephone call centers: well run call centers *are* QED (Gans et al. [13]). The management of such centers is a challenging task, with staffing being one of its primary components (Borst et al. [9]). Staffing involves quantifying the trade-off between the number of agents (cost) and users-perceived service quality (e.g., probability of delay). Hence, call centers have been naturally and usefully modeled as queueing systems, with agents and calls being the servers and customers, respectively. The most prevalent practical model has been the FCFS M/M/N queue; e.g., see Borst et al. [9]. In practice, however, call arrivals need not be simply Poissonian, and service durations need not be exponentially distributed (Brown et al. [10]), which underscores the importance of the QED G/GI/N queue.

QED research. The literature on multiserver queues is extensive (Whitt [34]). We restrict our attention here to relevant QED papers. The QED regime can be analyzed in either steady state, which entails convergence of steady-state distributions, or transient state, which entails functional limit theorems. As mentioned, convergence takes place as the number of servers $N \uparrow \infty$, so that service capacity and offered load are carefully balanced, for example via the square-root staffing rule mentioned earlier.

The QED regime was first introduced by Erlang [11] for both M/M/N and M/M/N/N in steady state. Erlang derived square-root staffing rules via marginal and numerical analysis, without analytical proofs—these were later provided by the editors of Erlang [11]. The QED steady-state M/M/N/N queue was treated in Jagerman [21] as part of an overall asymptotic analysis. But the prevalent characterization of the QED regime in steady state—in terms of the equivalence between square-root staffing and a limiting nonnegligible delay probability—had to await Halfin and Whitt [18], who studied a system with exponential service times (GI/M/N) in both transient and steady states. A more general model with *phase-type* service times, covering also multiple customer classes with static priorities, was examined in Puhalskii and Reiman [31]. There the authors studied the transient behavior of properly scaled queue-length and waiting-time processes, establishing convergence to a particular finite-dimensional diffusion process. Finite-dimensionality of the limit is inherited from that of the prelimit, the latter being due to the finite number of exponential phases of the service times. With deterministic service times, however, such finite-dimensionality is lacking, which introduces further challenges. These were circumvented in Jelenković et al. [22], where it was shown that the steady-state performance of a system with deterministic service times is related to a negative-drift Gaussian random walk. The present paper adds transient analysis to Jelenković et al. [22] and, in fact, generalizes it to finite-support service times. Note that both Puhalskii and Reiman [31] and Jelenković et al. [22] fall short of covering steady state.

Current research of QED queues focuses mainly on enriching modeling scope. Models that take into account customer impatience, important for call center applications, can be found in Fleming et al. [12], Garnett et al. [14], and Zeltyn and Mandelbaum [38]. A diffusion approximation for a finite buffer queue with some nonexponential service times was studied in Whitt [35, 37]. Revenue maximization and constraint satisfaction were discussed in Armony and Maglaras [2, 3], Borst et al. [9], Maglaras and Zeevi [26], and Mandelbaum and Zeltyn [29]. Optimal stochastic control of QED systems with multiclass customers and multiskilled servers was considered in Atar [5, 6], Harrison and Zeevi [19], and Tezcan [32], while optimal *joint* control and staffing were achieved in Atar et al. [7] for a single customer class and Gurvich et al. [17] for a homogeneous single-server pool.

Contents. The paper is organized as follows. In the next section we present two basic lemmas on multiserver queues, formally define the QED regime, and state some weak-convergence facts for the arrivals and related infinite-server processes. Section 3 contains the main results of our paper (Theorem 3.1), as well as two key lemmas (Lemmas 3.2 and 3.3) that are instrumental in establishing these results. Our approach is based on the analysis of a sorting operator (Lemma 2.1) that is used to describe the evolution of customers' waiting times. The limiting waiting time is represented in terms of a supremum over random weighted trees introduced in §3.3. The initial conditions of the system are discussed in §3.5. The special cases of GI/D/N and GI/M/N are examined in §§3.7 and 3.8, respectively, giving rise to conjectures on G/GI/N QED queues in steady state that accommodate also infinite-support service times. Possible applications of the proposed approximation are outlined in §4. We conclude with some proofs in §5.

Notations and conventions. Denote by $D(I, \mathbb{R}^k)$ the space of all $\mathbb{R}^k$-valued functions on an interval $I \subseteq [0, \infty)$, which are right-continuous with left-hand limits (r.c.l.l.) everywhere in I, endowed with the usual Skorohod J_1 topology (Whitt [34, §3.3]); let $d_I(\cdot, \cdot)$ be the J_1 metric on $D(I, \mathbb{R}^k)$, $I \subseteq [0, \infty)$. Interval I is of either the form $[a, b]$ or $[a, c)$, where $0 \le a < b < c \le \infty$. Let $C(I, \mathbb{R})$ be the space of all continuous $\mathbb{R}$-valued functions on I. From this point on we adopt the convention that *all stochastic processes, unless stated otherwise, are extended to* $(-\infty, 0)$ *by setting them identically to* 0 *on* $(-\infty, 0)$. Let $\Rightarrow$ denote convergence in distribution—for stochastic processes in $D(I, \mathbb{R}^k)$ and for random variables in $\mathbb{R}^k$. Denote by $\mathrm{Disc}(x) = \{t \in I: x(t-) \neq x(t)\}$ the set of discontinuity points of $x \in D(I, \mathbb{R})$; whenever we write Disc(x), the domain of x will be clear from the context; hence, it will not appear explicitly. Finally, let $1_{\{\cdot\}}$ be the usual indicator function and θ_τ: $D([0,T],\mathbb{R}) \to D([0,T-\tau],\mathbb{R})$ be the shift operator for $\tau \in \mathbb{R}$, defined by $\theta_\tau(x)(t) = x((t+\tau)^+)1_{\{t\in[\tau^-,\,T-\tau]\}}$, $0 \le t \le T-\tau$ (x^+ and x^- denote the positive and negative parts of x, respectively; that is, $x^+ = \max\{x,0\}$ and $x^- = \max\{-x,0\}$).

2. Model and preliminaries.

2.1. Combinatorics. We start this section by considering an N-server queue with customers being served in the order of arrival (FCFS service discipline). Here we omit a detailed description of the multiserver queue; such descriptions are standard (using the Kiefer-Wolfowitz vector [24]) and can be found in, e.g., Asmussen [4] and Baccelli and Bremaud [8]. Customers, labeled with integers in the order of their arrival, arrive to the system at times $0 < t_1 < t_2 \cdots$ and require services $\{S_n\}$. The number of arrivals in the time interval $[0,t]$ is finite for all $t \ge 0$, almost surely. We allow that $l < \infty$ customers are present in the system at time $t=0$ (either receiving service or waiting). These customers are labeled by $-1,-2,\ldots,-l$, and the initial state of the system (at time $t=0$) is specified by their respective departure times: $d_{-1}, d_{-2}, \ldots, d_{-l}$. The waiting and departure times of the nth ($n \ge 1$) customer are denoted by w_n and d_n, respectively; clearly $d_n = t_n + w_n + S_n$. In addition, we define $v(t)$ to be the virtual waiting time at time $t \ge 0$, namely the amount of time that a (possibly hypothetical) customer would have to wait for service had it arrived at time t. Equivalently, $v(t)$ is the amount of time (beyond t) required until one of the servers becomes idle; thus, $v(t)=0$ if at least one server is idle at time t. By definition, $v(t)$ is right-continuous so that $v(t_n-) = w_n$ (hence, $v(t_n) = w_n + S_n$). Finally, let $\mathcal{O}$ denote the sorting (ordering) operator defined on finite sequences of reals, and let $\mathcal{O}_k\{x_i\}$ be the kth largest element in the sequence $\{x_i\}$. We set $\mathcal{O}_k\{x_i\} = 0$ when $\{x_i\}$ contains fewer than k elements.

Lemma 2.1. *Consider an N-server FCFS queue. Then the waiting times satisfy*

$$w_n = \mathcal{O}_N\{(d_i - t_n)^+,\ i<n\}, \quad n \ge 1, \tag{1}$$

and the virtual waiting time adheres to

$$v(t) = \mathcal{O}_N\{(d_i - t)^+,\ i \le \max\{j\colon\ t_j \le t\}\}, \quad t \ge 0. \tag{2}$$

Remark 2.1. Observe that, if $N=1$, then departures are in the order of arrivals and, thus, $\mathcal{O}_j\{d_i,\ i<n\} = d_{n-j}$. This relationship is not necessarily true for $N>1$ because the system's ability to process more than one customer simultaneously can lead to customers departing out of order of their arrival.

Remark 2.2. The relation (1) is a constructive recursive way to generate the sequence of waiting times $\{w_n\}$. It is thus a multiserver analogue of the classical Lindley's equations. Indeed, $\{d_i\}$ is monotone increasing when $N=1$, in which case

$$w_n = (d_{n-1} - t_n)^+ = (w_{n-1} + S_{n-1} - (t_n - t_{n-1}))^+.$$

Proof. Consider the waiting time of the nth customer. Because the queue operates in a FCFS fashion, one can ignore all customers with indices higher than n. Observe that right after time $\mathcal{O}_i\{d_j,\ j<n\}$, $i>1$, there are at most $(i-1)$ customers in the system (waiting or in service) that arrived before the nth customer. This follows from the fact that at most $(i-1)$ departure times are larger than $\mathcal{O}_i\{d_j,\ j<n\}$. However, given that there are only N servers in the system, the nth customer can get into service only if there are no more than $(N-1)$ customers in the system with lower indices. Namely, the nth customer can potentially start service at $\mathcal{O}_N\{d_i,\ i<n\}$. Yet service cannot start until the customer arrives, i.e., until t_n. If $d_i - t_n > 0$ for N or more i's, with $i<n$, then the customer with index n starts waiting upon arrival at time t_n and must wait for a time equal to the Nth largest of these positive times. Otherwise, the nth customer enters service immediately upon arrival at time t_n. Thus, w_n satisfies (1).

The arguments justifying (2) are similar. The quantity $\mathcal{O}_N\{d_i,\ i \le \max\{j\colon\ t_j \le t\}\}$ represents the time at which a hypothetical customer arriving at time t could potentially start service, and, therefore, (2) holds. □

Next, we state a simple sample-path lemma for FCFS systems. Effectively, the lemma formalizes the notion that customers with equal service requirements depart in the order of their arrival.

Lemma 2.2. *A customer with service requirement S arriving to the system no earlier* (*no later*) *than time t departs no earlier* (*no later*) *than $t + v(t) + S$.*

Proof. Due to the FCFS service discipline, the customer arriving at time t can start service no earlier (no later) than $t+v(t)$. This implies that the departure occurs no earlier (no later) than $t+v(t)+S$. □

Applying the lemma, with t being the arrival time of a customer whose service requirement is S, yields that a customer arriving prior to time t, whose service requirement is also S, departs no later than $t+v(t)+S$, the latter being the departure time of the customer arriving at time t. Equivalently, customers with equal service requirements depart in the order of their arrival.

2.2. Our G/D$_K$/N model. The D$_K$ in G/D$_K$/N stands for service times that take a finite number of values (have finite support). Formally, with S denoting a generic service time, let S take a finite number of K possible values, $0 < s_1 < s_2 < \cdots < s_K$, such that $\mathbb{P}[S = s_i] = p_i > 0$. In the remainder of the paper we consider a sequence of G/D$_K$/N queues, indexed by the number of servers N, with long-run arrival rates $\lambda_N \to \infty$, as $N \to \infty$. (Quantities referring to the Nth system are indexed by N.) We assume that service requirements of customers do not vary with N, that they have *finite support*, that they are independent and identically distributed (i.i.d.) across customers, and that they are independent of the arrivals. Let μ denote the service rate. Then

$$\mu^{-1} := \mathbb{E}S = \sum_{i=1}^{K} p_i s_i < \infty.$$

The offered load to the Nth system is then $R_N := \lambda_N \mathbb{E}S$, and the traffic intensity (servers' utilization) is $\rho_N := R_N/N$.

2.3. Arrival processes. Let $A_N(t)$, $t \geq 0$, be the number of arrivals in the Nth system during time interval $[0, t]$. The process $\{A_N(t),\ t \geq 0\}$ is r.c.l.l., nondecreasing, nonnegative, and integer-valued with jumps of size 1 such that $A_N(0) = 0$ and $A_N(t) < \infty$ for all $t \geq 0$, almost surely. That is, $\{A_N(t),\ t \geq 0\}$ is a counting process with the added assumptions of jumps of size 1 (customers arrive one at a time) and $A_N(t) < \infty$ for all $t \geq 0$ (finite number of arrivals in any finite time interval). It is assumed that the sequence of arrival processes obeys the functional strong law of large numbers (FSLLN):

$$\sup_{0 \leq t \leq T} \left| \frac{A_N(t)}{\lambda_N} - t \right| \to 0, \tag{3}$$

with probability 1, as $N \to \infty$, for every fixed $0 < T < \infty$, and

$$\lambda_N/N \to \lambda,$$

as $N \to \infty$, for some finite $\lambda > 0$. In addition, it is assumed that the same sequence satisfies a form of the functional central limit theorem (FCLT):

$$\left\{ \frac{A_N(t) - \lambda_N t}{\sqrt{N}},\ t \geq 0 \right\} \ \Rightarrow\ Z \tag{4}$$

in $D([0, \infty), \mathbb{R})$, as $N \to \infty$, where $Z \equiv \{Z(t),\ t \geq 0\}$ is a zero-mean Gaussian process with continuous sample paths. For example, suppose that the arrival process in the Nth system is a renewal process with interarrival times that are i.i.d., with mean $1/\lambda_N$ such that $\lambda_N/N \to \lambda$, and coefficient of variation v_N such that $v_N \to v$, as $N \to \infty$; then, by the CLT for renewal processes, the process Z is a driftless Brownian motion with the variance parameter λv^2. The preceding two limits constitute our two main assumptions on the sequence of arrival processes.

Next we discuss the implications of (3) and (4) on the arrival processes of customers with the same service requirement. Define $A_{i,N}(t)$, $t \geq 0$, to be the number of arrivals of customers with service requirement s_i, in the Nth system, during the time interval $[0, t]$; $A_N = \sum_{i=1}^{K} A_{i,N}$. Then (3), the i.i.d. assumption on the service requirements, and the independence between arrivals and services imply that the processes $\{A_{i,N}(t),\ t \geq 0\}$ also adhere to a FSLLN:

$$\sup_{0 \leq t \leq T} \left| \frac{A_{i,N}(t)}{p_i \lambda_N} - t \right| \to 0$$

with probability 1, as $N \to \infty$, for every fixed $0 < T < \infty$. The scaled-centered version of $A_{i,N}(t)$ is defined as

$$\hat{A}_{i,N}(t) := \frac{A_{i,N}(t) - p_i \lambda_N t}{\sqrt{N}}, \quad t \geq 0; \tag{5}$$

$\hat{A}_N = \sum_{i=1}^{K} \hat{A}_{i,N}$. Throughout the paper we use the hat symbol to indicate such scaled (and centered, when relevant) processes; underlined variables are used to denote vectors. The following lemma states that the vector process $\underline{\hat{A}}_N \equiv \{(\hat{A}_{1,N}(t), \ldots, \hat{A}_{K,N}(t)),\ t \geq 0\}$ converges weakly to a process that is related to the process Z. Let the vector process $\underline{B} := \{(B_1(t), \ldots, B_K(t)),\ t \geq 0\}$ be a K-dimensional zero-drift Brownian motion, independent of Z, with the covariance matrix $\Sigma = [\Sigma_{ij}]$ defined by its elements $\Sigma_{ii} = p_i(1 - p_i)$ and $\Sigma_{ij} = -p_i p_j$ for $i \neq j$.

LEMMA 2.3. *We have, as $N \to \infty$,*

$$\underline{\hat{A}}_N \;\Rightarrow\; \underline{\hat{A}}$$

in $D([0,\infty),\mathbb{R}^K)$, where $\underline{\hat{A}} := \{(\hat{A}_1(t),\dots,\hat{A}_K(t)),\ t \ge 0\}$ with

$$\hat{A}_i(t) = p_i Z(t) + B_i(\lambda t).$$

REMARK 2.3. As a consequence of $Z = \sum_{i=1}^K \hat{A}_i$, one must have $\sum_{i=1}^K B_i \equiv 0$, which prevails in view of $\mathrm{Var}(\sum_{i=1}^K B_i) = (\sum_{i=1}^K p_i) - (\sum_{i=1}^K p_i)^2 = 1 - 1 = 0$.

PROOF. The lemma follows from the continuous mapping theorem, applied to composition and addition. (The continuity of addition is due to continuity of the sample paths of Z and $\underline{B}$ in this case.) See Whitt [34, §9.5] for details. □

2.4. Infinite-server processes. Associated with the Nth element in our sequence of finite-server G/D$_K$/N queues, $N = 1, 2, \dots$, there is a naturally *corresponding* infinite-server G/D$_K$/∞ queue: its arrival process is A_N and the service times are S-distributed independently of N. The sequence of corresponding infinite-server queues plays an important role in the analysis of its originating finite-server sequence. It will now be discussed and then analyzed asymptotically, as $N \to \infty$.

Infinite-server systems are typically more tractable than their finite-server counterparts. This has motivated approximations of multiserver systems by corresponding infinite-server systems. For a review of results on infinite-server systems we refer the reader to Whitt [34, Ch. 10].

The assumption of the service time taking only a finite number of values was found advantageous also in the analysis of infinite-server systems (Glynn and Whitt [15]). There it was observed that, in that case, the number of customers in the system can be expressed in a simple form. Applying that observation specifically to our setting, for $t \ge 0$,

$$X_{i,N}(t) := A_{i,N}(t) - A_{i,N}(t - s_i)$$

is the number of customers, at time t, in the corresponding G/D$_K$/∞ system (arrivals A_N and services S), whose service requirement is s_i, and with the additional assumption that the system is empty at time $t = 0$. *The assumption that $X_{i,N}(0) = 0$, for all i and all N, will be maintained throughout the paper.*

The total number of customers in the corresponding Nth infinite-server system, at time $t \ge 0$, is given by

$$X_N(t) := \sum_{i=1}^K X_{i,N}(t).$$

The next lemma indicates that the number of customers in the Nth infinite-server system does not reach the number of servers N, over the time interval $[0, s_K - \delta)$, $\delta > 0$.

LEMMA 2.4. *If $\lambda_N \mathbb{E}S/N \le 1$ then for every fixed $\delta \in (0, s_K)$ there exists a fixed $N_\delta < \infty$ such that, for all $N \ge N_\delta$, condition (3) implies*

$$\sup_{t\in[0,s_K-\delta)} X_N(t) < N \quad \textit{with probability } 1. \tag{6}$$

PROOF. See §5. □

We now study the limiting behavior of a sequence of G/D$_K$/∞, with $\lambda_N \to \infty$ as $N \to \infty$ under (3) and (4). We introduce scaled versions of the related infinite-server processes:

$$\hat{X}_{i,N}(t) := \frac{X_{i,N}(t) - p_i\lambda_N(s_i \wedge t)}{\sqrt{N}},$$

$$\hat{X}_N(t) := \frac{X_N(t) - \lambda_N\sum_{i=1}^K p_i(s_i \wedge t)}{\sqrt{N}},$$

for $i = 1, \dots, K$ and $t \ge 0$, where $\wedge$ denotes the minimum operator, and note that by the preceding and (5),

$$\hat{X}_{i,N}(t) = \hat{A}_{i,N}(t) - \hat{A}_{i,N}(t - s_i), \tag{7}$$

$$\hat{X}_N(t) = \sum_{i=1}^K \hat{X}_{i,N}(t). \tag{8}$$

Furthermore, the following processes are of interest in the next section:

$$\hat{X}_{i,N}^{\uparrow\varepsilon}(t) := \sup_{u\in[t-\varepsilon,\,t+\varepsilon]} \hat{A}_{i,N}(u) - \inf_{u\in[t-\varepsilon,\,t+\varepsilon]} \hat{A}_{i,N}(u-s_i), \qquad \hat{X}_N^{\uparrow\varepsilon}(t) := \sum_{i=1}^{K} X_{i,N}^{\uparrow\varepsilon}(t),$$

$$\hat{X}_{i,N}^{\downarrow\varepsilon}(t) := \inf_{u\in[t-\varepsilon,\,t+\varepsilon]} \hat{A}_{i,N}(u) - \sup_{u\in[t-\varepsilon,\,t+\varepsilon]} \hat{A}_{i,N}(u-s_i), \qquad \hat{X}_N^{\downarrow\varepsilon}(t) := \sum_{i=1}^{K} X_{i,N}^{\downarrow\varepsilon}(t),$$

where $t \ge 0$ and $\varepsilon \ge 0$; recall that $A_{i,N}(t) = 0$ for $t < 0$ by definition. In general, we use the symbols $\uparrow$ and $\downarrow$ as superscripts to indicate sup and inf operators, respectively. However, for convenience we slightly abuse this notation in the above definitions.

Next, let $\hat{A}_i^{\uparrow\varepsilon}(t)$ and $\hat{A}_i^{\downarrow\varepsilon}(t)$ be defined as $\sup_{u\in[t-\varepsilon,\,t+\varepsilon]} \hat{A}_i(u)$ and $\inf_{u\in[t-\varepsilon,\,t+\varepsilon]} \hat{A}_i(u)$, respectively; the processes $\{\hat{A}_i(t),\, t \ge 0\}$ are defined in Lemma 2.3. It is worth mentioning that, by the modulus of continuity of the Brownian path (e.g., see Karatzas and Shreve [23, p. 114]) and the assumption (4), both $\hat{A}_i^{\uparrow\varepsilon}(t)$ and $\hat{A}_i^{\downarrow\varepsilon}(t)$ converge to $\hat{A}_i(t)$ with probability 1 as $\varepsilon \downarrow 0$ on any finite interval (see Lemma 2.3). The following lemma characterizes the infinite-server process in the limit, as $N \to \infty$. Set $\underline{\hat{X}}_N \equiv \{(\hat{X}_{1,N}(t), \ldots, \hat{X}_{K,N}(t)),\, t \ge 0\}$.

Lemma 2.5. *We have, as* $N \to \infty$,

$$(\underline{\hat{X}}_N, \hat{X}_{1,N}^{\uparrow\varepsilon}, \ldots, \hat{X}_{K,N}^{\uparrow\varepsilon}, \hat{X}_{1,N}^{\downarrow\varepsilon}, \ldots, \hat{X}_{K,N}^{\downarrow\varepsilon}) \;\Rightarrow\; (\underline{\hat{X}}, \hat{X}_1^{\uparrow\varepsilon}, \ldots, \hat{X}_K^{\uparrow\varepsilon}, \hat{X}_1^{\downarrow\varepsilon}, \ldots, \hat{X}_K^{\downarrow\varepsilon})$$

in $D([0,\infty), \mathbb{R}^{3K})$, *where*

$$\underline{\hat{X}} \equiv \{(\hat{X}_1(t), \ldots, \hat{X}_K(t)),\, t \ge 0\},$$

and

$$\hat{X}_i(t) = \hat{A}_i(t) - \hat{A}_i(t-s_i),$$
$$\hat{X}_i^{\uparrow\varepsilon}(t) = \hat{A}_i^{\uparrow\varepsilon}(t) - \hat{A}_i^{\downarrow\varepsilon}(t-s_i), \qquad \hat{X}_i^{\downarrow\varepsilon}(t) = \hat{A}_i^{\downarrow\varepsilon}(t) - \hat{A}_i^{\uparrow\varepsilon}(t-s_i).$$

Proof. The lemma follows from standard continuous-mapping arguments (e.g., see Whitt [34, §3.4]). To this end, we need to verify that all of the mappings involved are continuous in the J_1 topology. However, that follows from the continuity of the sup and inf operators and from (7). In particular, $d_{[0,T]}(x^{\uparrow\varepsilon}, y^{\uparrow\varepsilon}) \le d_{[0,T+\varepsilon]}(x,y)$ and $d_{[0,T]}(x^{\downarrow\varepsilon}, y^{\downarrow\varepsilon}) \le d_{[0,T+\varepsilon]}(x,y)$, for $x, y \in D([0, T+\varepsilon], \mathbb{R})$ by the definition of the J_1 metric; here $x^{\uparrow\varepsilon} := \{x^{\uparrow\varepsilon}(t) := \sup_{u\in[(t-\varepsilon)^+,\, t+\varepsilon]} x(u),\, t \in [0,T]\}$ and $x^{\downarrow\varepsilon} := \{x^{\downarrow\varepsilon}(t) := \inf_{u\in[(t-\varepsilon)^+,\, t+\varepsilon]} x(u),\, t \in [0,T]\}$. Addition is continuous in this case [33] because $\mathrm{Disc}(\hat{A}_i) = \mathrm{Disc}(\hat{A}_i^{\uparrow\varepsilon}) = \mathrm{Disc}(\hat{A}_i^{\downarrow\varepsilon}) = \varnothing$ with probability 1 for all i (see Lemma 2.3). Finally, the convergence in the product space is due to Proposition VI.2.2 in [20]. □

Suppose now that the arrival process in the Nth system is renewal, with interarrival times as described previously: mean $1/\lambda_N$ ($\lambda_N/N \to \lambda$) and coefficient of variation $v_N \to v$, as $N \to \infty$. Then, Lemmas 2.3 and 2.5 imply that the limiting vector process $\underline{\hat{X}}$ is a driftless K-dimensional Gaussion process, characterized by its covariance function

$$\mathrm{cov}(\hat{X}_i(t), \hat{X}_j(t+r)) = \begin{cases} -\lambda p_i p_j (1-v^2)(t - (t-s_i)^+ \vee (t+r-s_j)^+)^+, & i \ne j, \\[2ex] \lambda p_i (1 - p_i(1-v^2))(t - (t+r-s_i)^+)^+, & i = j, \end{cases} \tag{9}$$

where $t, r \ge 0$ and $\vee$ denotes the maximum operator. When the arrival process is not renewal, a similar but more cumbersome expression can be derived in terms of the covariance function of Z.

We conclude this section with some remarks on the weak limit of $\hat{X}_N$, as $N \to \infty$. This infinite-server process limit $\hat{X} \equiv \{\hat{X}(t),\, t \ge 0\}$ is given by

$$\hat{X}(t) := \sum_{i=1}^{K} \hat{X}_i(t), \tag{10}$$

with the $\hat{X}_i$'s described in Lemma 2.5 above. If the arrival process in the Nth system is renewal, then the process $\hat{X}$ is zero-mean Gaussian with the following covariance function (Whitt [34, p. 353]; see also [25]):

$$\mathrm{cov}(\hat{X}(t), \hat{X}(t+r)) = \lambda \int_0^t H(u)\bar{H}(u+r)\,du + \lambda v^2 \int_0^t \bar{H}(u)\bar{H}(u+r)\,du, \qquad t, r \ge 0, \tag{11}$$

where $H(u) = \mathbb{P}[S \le u] = \sum p_i 1_{\{s_i \le u\}}$ and $\bar{H}(u) = 1 - H(u)$; recall that, by our convention, $\hat{X}(t) \equiv 0$ for $t < 0$. (Equality (11) holds in fact for general service-time distributions (Whitt [34, p. 354]).) In the special case when the sequence of arrival processes is Poisson, the elements of the vector process $\underline{\hat{A}}$ are independent (due to the thinning property of the Poisson process) and $v = 1$. The expression for the covariance function then simplifies to

$$\operatorname{cov}(\hat{X}(t), \hat{X}(t+r)) = \lambda \int_0^t \bar{H}(u+r)\, du, \qquad t, r \geq 0; \tag{12}$$

in particular, for our finite support services (see also (9)),

$$\operatorname{cov}(\hat{X}(t), \hat{X}(t+r)) = \lambda \sum_{i=1}^{K} p_i (t - (t+r-s_i)^+)^+, \qquad t, r \geq 0.$$

REMARK 2.4. Lemmas 2.5 and (10) indicate that the process $\hat{X}$ reaches its steady-state regime at time $t = s_K$. Furthermore, for any $\delta > 0$ and all $\varepsilon \leq \delta/2$, the processes $\hat{X}^{\uparrow\varepsilon} \equiv \{\hat{X}^{\uparrow\varepsilon}(t),\ t \geq 0\}$ and $\hat{X}^{\uparrow\varepsilon} \equiv \{\hat{X}^{\downarrow\varepsilon}(t),\ t \geq 0\}$, defined by $\hat{X}^{\uparrow\varepsilon}(t) = \sum_{i=1}^{K} \hat{X}_i^{\uparrow\varepsilon}(t)$ and $\hat{X}^{\downarrow\varepsilon}(t) = \sum_{i=1}^{K} \hat{X}_i^{\downarrow\varepsilon}(t)$, respectively, reach steady-state by time $t = s_K + \delta$.

3. Main results.

3.1. QED scaling. Our main objective is to characterize the (asymptotic) behavior of the virtual waiting time $V_N(t)$ in the QED regime. This regime is defined for a sequence of $G/D_K/N$ queues, $N = 1, 2, \ldots$, in the following manner. The sequence of arrival rates $\{\lambda_N\}$ increase indefinitely so that the number of servers N and utilizations ρ_N are related, in the limit as $N \uparrow \infty$, via

$$\lim_{N \to \infty} \sqrt{N}(1 - \rho_N) = \beta,$$

for some $0 < \beta < \infty$. Equivalently, the number of servers exceeds the offered load by a square-root term:

$$N = R_N + \beta\sqrt{R_N} + o(\sqrt{R_N}),$$

as $N \uparrow \infty$. In the QED regime, the arrival rate and the number of servers are proportional in the limit, i.e., $\lambda_N / N \to \mu$. This implies that the arrival rate λ, appearing, for example, in Lemma 2.3 and in (11), in fact equals the service rate μ; hence *only μ will be used from now on*. For notational simplicity, introduce

$$\beta_N := \sqrt{N} \frac{N - R_N}{R_N}, \tag{13}$$

and note that $\beta_N \to \beta$, as $N \uparrow \infty$.

3.2. Preparatory lemmata. Define $\{D_i^t\}_{i \geq 1}$ to be the ordered (in a decreasing order, with ties broken arbitrarily) sequence of departure times of customers that arrived to the system no later than time t; in particular, D_1^t is the departure time of the last customer to leave the system among those with arrival times at most t. Then the virtual waiting time $V_N(t)$ is determined by D_N^t (see Lemma 2.1), or more precisely $V_N(t) = (D_N^t - t)^+$. If the number of arrivals prior to and including time t is less than j we set $D_i^t = -\infty$ for $i \geq j$, so that $V_N(t) = 0$ if that number of arrivals is less than N. It is useful to define $z_t(r)$ as the number of customers that arrive to the system no later than time t and depart strictly after time r (i.e., the number of departures among $\{D_i^t\}_{i \geq 1}$ that occur after time r):

$$z_t(r) := \sum_{i=1}^{\infty} 1_{\{D_i^t > r\}},$$

for $r, t \geq 0$. Note that $z_t(r)$ relates to D_N^t via

$$\{z_t(r) < N\} \subseteq \{D_N^t \leq r\} \qquad \text{and} \qquad \{z_t(r) > N\} \subseteq \{D_N^t > r\}. \tag{14}$$

It suffices to consider the virtual waiting time after N customers have started to receive service because no waiting is possible prior to that time. That is, if fewer than N departures occur in the future, then the waiting time is equal to 0 because at least one of the servers is idle in that case (if a customer enters service then it must depart from the system as some point in time in the future).

If $\{T_i\}$ is a sequence of arrival times, then $V_N(T_i-)$ is the waiting time of the customer with arrival time T_i. The quantity $V_N(T_i-)$ is well defined because customers arrive one at a time. Let $T_N^*(t)$ and $S_N^*(t)$ be the arrival and service time of the customer with the departure time D_N^t, respectively. Note that, by definition,

$$T_N^*(t) + S_N^*(t) + V_N(T_N^*(t)-) = D_N^t. \tag{15}$$

One key feature of the QED regime, the one that pertains to its "QD" (quality-driven) aspect, is that the virtual waiting time $V_N(t)$ vanishes in the limit as the number of servers N increases. Formalizing this, the following lemma states that, in the QED regime, the virtual waiting time $V_N(t)$ at time t is determined by customers arriving to the system during the time period $\bigcup_{i=1}^{K}(t-s_i-\varepsilon, t-s_i+\varepsilon)$, for some small $\varepsilon>0$, with N large enough. The implication of the lemma is that $V_N(t)$ in the QED regime is determined (in addition to the arrival process) by the virtual waiting time in the neighborhoods of time instances $t-s_i$, $i=1,\ldots,K$. The proof of Lemma 3.1 is based on an analysis of the function $z_t(r)$ and uses the fact that the sequence of arrival processes satisfies a FSLLN.

Lemma 3.1. *If for fixed $T>s_K$ and some $\varepsilon\in(0,\min\{s_1,(T-s_K)/2\})$, as $N\to\infty$,*

$$\mathbb{P}\Big[\sup_{t\in[T-s_K-2\varepsilon,\,T]} V_N(t)\le\varepsilon\Big]\to 1, \tag{16}$$

then, as $N\to\infty$,

$$\mathbb{P}\Big[\sup_{t\in(T,\,T+s_1-\varepsilon]} V_N(t)<2\varepsilon\Big]\to 1$$

and

$$\mathbb{P}\big[T_N^*(t)+S_N^*(t)\in(t-2\varepsilon,t+2\varepsilon),\ \forall\, t\in(T,T+s_1-\varepsilon]\big]\to 1.$$

Proof. See §5. □

The following corollary states that, if the waiting time is bounded on an interval of length strictly larger than the maximum service requirement, then the waiting time remains bounded on an arbitrary interval of finite length. The proof is obtained through successive use of the first statement of Lemma 3.1.

Corollary 3.1. *If for all $\varepsilon>0$ small enough, as $N\to\infty$,*

$$\mathbb{P}\Big[\sup_{t\in[0,\,s_K+2\varepsilon]} V_N(t)\le\varepsilon\Big]\to 1,$$

then, for each $T<\infty$ there exists $c\equiv c(T)<\infty$ such that for all $\varepsilon>0$ small enough, as $N\to\infty$,

$$\mathbb{P}\Big[\sup_{t\in[0,\,T]} V_N(t)\le c\varepsilon\Big]\to 1.$$

Proof. See §5. □

At this point, we introduce two additional processes, $\{V_N^{\uparrow\varepsilon}(t),\ t\ge 0\}$ and $\{V_N^{\downarrow\varepsilon}(t),\ t\ge 0\}$, that are defined as functions of the virtual waiting time:

$$V_N^{\uparrow\varepsilon}(t):=\sup_{u\in[(t-\varepsilon)^+,\,t+\varepsilon]} V_N(u),\qquad V_N^{\downarrow\varepsilon}(t):=\inf_{u\in[(t-\varepsilon)^+,\,t+\varepsilon]} V_N(u), \tag{17}$$

where $t\ge 0$ and $\varepsilon\ge 0$. These processes will play an important role in bounding the virtual waiting time. According to Lemma 2.1 and (15), the virtual waiting time at time t is determined by $V_N(T_N^*(t)-)$. However, Lemma 3.1 establishes only lower and upper bounds for $T_N^*(t)$. Hence, we use $V_N^{\uparrow\varepsilon}(t)$ and $V_N^{\downarrow\varepsilon}(t)$ to estimate $V_N(T_N^*(t)-)$.

Note that by (17) and Lemma 2.2 we have

$$0\le V_N^{\downarrow\varepsilon}(t)\le V_N^{\uparrow\varepsilon}(t)\le 2\varepsilon+V_N(t+\varepsilon). \tag{18}$$

We also introduce corresponding scaled processes $\{\widehat{V}_N(t),\ t\ge 0\}$, $\{\widehat{V}_N^{\uparrow\varepsilon}(t),\ t\ge 0\}$ and $\{\widehat{V}_N^{\downarrow\varepsilon}(t),\ t\ge 0\}$ by

$$\widehat{V}_N(t):=\sqrt{N}\,\frac{V_N(t)}{\mathbb{E}S} \tag{19}$$

and

$$\widehat{V}_N^{\uparrow\varepsilon}(t):=\sqrt{N}\,\frac{V_N^{\uparrow\varepsilon}(t)}{\mathbb{E}S},\qquad \widehat{V}_N^{\downarrow\varepsilon}(t):=\sqrt{N}\,\frac{V_N^{\downarrow\varepsilon}(t)}{\mathbb{E}S}.$$

The scaling (19) of the (virtual) waiting time is standard in the context of the QED regime, e.g., see Halfin and Whitt [18], Jelenković et al. [22], and Puhalskii and Reiman [31]. Next, for notational simplicity, given a random process $f(\cdot)$ let

$$\mathbb{E}_{\langle S\rangle}[f(t-S)]=\sum_{i=1}^{K}p_i f(t-s_i), \tag{20}$$

in which $\langle S\rangle$ stands for averaging over the distribution of the random variable S. Formally, $\mathbb{E}_{\langle S\rangle}[f(\cdot)]$ is the conditional expectation over a nonnegative discrete random variable S independent of $f(\cdot)$, given the sigma field generated by $f(\cdot)$ up to time $(t-s_1)$.

The next lemma is the first of two technical results that play an instrumental role in establishing our main results. Using the preceding lemma, it quantifies the relationship between $V_N(t)$ and $V_N^{\uparrow\varepsilon}(t-s_i)$, $V_N^{\downarrow\varepsilon}(t-s_i)$ and, effectively, strengthens Lemma 3.1.

Lemma 3.2. *Let $\delta_N = K/(\rho_N\sqrt{N})$. If for fixed $T > s_K$ and all sufficiently small $\varepsilon > 0$, as $N\to\infty$,*

$$\mathbb{P}\Big[\sup_{t\in[T-s_K-2\varepsilon,\,T]} V_N(t)\le\varepsilon\Big]\to 1, \tag{21}$$

then, as $N\to\infty$,

$$\mathbb{P}\big[\widehat{V}_N(t)\le \big(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 2\varepsilon}(t-S)]+\rho_N^{-1}\widehat{X}_N^{\uparrow 2\varepsilon}(t)-\beta_N+\delta_N\big)^+,\ \forall\, t\in(T,T+s_1-2\varepsilon]\big]\to 1$$

and

$$\mathbb{P}\big[\widehat{V}_N(t)\ge \big(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\downarrow 2\varepsilon}(t-S)]+\rho_N^{-1}\widehat{X}_N^{\downarrow 2\varepsilon}(t)-\beta_N\big)^+,\ \forall\, t\in(T,T+s_1-2\varepsilon]\big]\to 1.$$

Proof. See §5. □

Remark 3.1. Note that the assumption of the lemma is on the virtual waiting time, *not* its scaled version.

Remark 3.2. Recall from §3.1 that $\rho_N\to 1$, $\delta_N\to 0$, and $\beta_N\to\beta$ as $N\to\infty$.

The following lemma is similar to the preceding one. However, instead of bounding the virtual waiting time $\widehat{V}_N(t)$ this lemma bounds the supremum and infimum of the virtual waiting time over some small neighborhood, i.e., $\widehat{V}_N^{\uparrow\varepsilon}(t)$ and $\widehat{V}_N^{\downarrow\varepsilon}(t)$. The asymptotic inequalities established below will be used to estimate terms that appear in the statement of Lemma 3.2.

Lemma 3.3. *Let $\delta_N = K/(\rho_N\sqrt{N})$. If for $T > s_K$ and all sufficiently small $\varepsilon > 0$, as $N\to\infty$,*

$$\mathbb{P}\Big[\sup_{t\in[T-s_K-3\varepsilon,\,T]} V_N(t)\le\varepsilon\Big]\to 1, \tag{22}$$

then, as $N\to\infty$,

$$\mathbb{P}\big[\widehat{V}_N^{\uparrow\varepsilon}(t)\le \big(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 3\varepsilon}(t-S)]+\rho_N^{-1}\widehat{X}_N^{\uparrow 3\varepsilon}(t)-\beta_N+\delta_N\big)^+,\ \forall\, t\in(T,T+s_1-3\varepsilon]\big]\to 1,$$

and

$$\mathbb{P}\big[\widehat{V}_N^{\downarrow\varepsilon}(t)\ge \big(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\downarrow 3\varepsilon}(t-S)]+\rho_N^{-1}\widehat{X}_N^{\downarrow 3\varepsilon}(t)-\beta_N\big)^+,\ \forall\, t\in(T,T+s_1-3\varepsilon]\big]\to 1.$$

Proof. See §5. □

Having established bounds on $\widehat{V}_N(t)$, $\widehat{V}_N^{\uparrow\varepsilon}(t)$, and $\widehat{V}_N^{\downarrow\varepsilon}(t)$, we shall consider the distribution of $\widehat{V}_N(t)$. The basic idea is to use Lemma 3.2 and Lemma 3.3 iteratively in order to relate $\widehat{V}_N(t)$ to $\{\widehat{X}_N(u),\ u\in[0,t]\}$ ($\widehat{X}_N(u)$ is defined in (8)) and the values of $\widehat{V}_N(\cdot)$ on some "initial" finite interval. Informally, Lemmas 3.2 and 3.3 will provide (in the limit) a Lindey-type recursion that relates $\widehat{V}_N(t)$ to $\widehat{V}_N(t-s_i)$, $i=1,\dots,K$.

3.3. Random trees.

Before stating the main result of our paper, we need to introduce some additional notation. Let $\mathcal{T}[t,\delta]$, $t,\delta>0$, be a full K-ary tree of finitely many nodes rooted at $r(\mathcal{T}[t,\delta])$; that is, each node has either 0 or K children. Each nonroot vertex in the tree can be *uniquely* identified with a vector $\underline{d} := (d_1,\dots,d_l)\in\{1,\dots,K\}^l$, where $|\underline{d}| := l$ is the distance from the vertex to the root (depth). The node corresponding to $(d_1,\dots,d_l,d_{l+1})$ is a child of the node corresponding to $(d_1,\dots,d_l)$. In particular, the K children of a nonleaf node $(d_1,\dots,d_l)$ correspond to $(d_1,\dots,d_l,1),(d_1,\dots,d_l,2),\dots,(d_1,\dots,d_l,K)$. The children of a nonleaf root are identified with unique scalars $d\in\{1,\dots,K\}$. For notational simplicity we associate the root with $d=0$, and, for $\underline{d}=(d_1,\dots,d_l)$, let

$$p_{\underline{d}} := \prod_{i=1}^{l} p_{d_i} \qquad\text{and}\qquad s_{\underline{d}} := \sum_{i=1}^{l} s_{d_i},$$

with the understanding that $p_{\underline{d}} := 1$ and $s_{\underline{d}} := 0$ for the root node; recall that p_i is the probability that a customer's service requirement equals s_i, where $s_1 < s_2 < \cdots < s_K$.

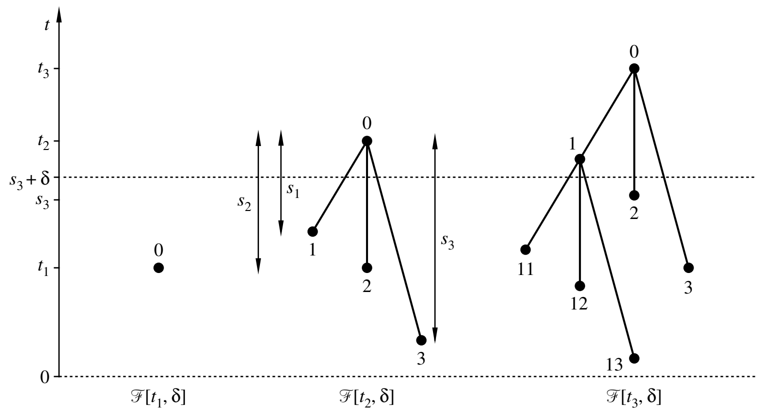
t
t_3
t_2
$s_3+\delta$
s_3
t_1
0
0
s_2
s_1
0
1
2
3
s_3
0
1
2
11
12
3
13
$\mathcal{T}[t_1,\delta]$
$\mathcal{T}[t_2,\delta]$
$\mathcal{T}[t_3,\delta]$

FIGURE 2. Illustration of the tree $\mathcal{T}[t,\delta]$ for $K=3$ and three different values of t: $t_1 \in [0, s_K+\delta)$, $t_2 \in [s_K+\delta, s_K+\delta+s_1)$, and $t_3 \in [s_K+\delta+s_1, (s_K+\delta+2s_1)\wedge(s_K+\delta+s_2))$. By definition we have $s_1 < s_2 < s_3$. Nodes are labeled with the corresponding $\underline{d}$'s.

To complete the description of the tree $\mathcal{T}[t,\delta]$ we need to specify its leaf nodes. The leaves of $\mathcal{T}[t,\delta]$ correspond to vectors $\underline{d}=(d_1,\ldots,d_l)$ that satisfy

$$s_K+\delta-s_{d_l} \le t-s_{\underline{d}} < s_K+\delta, \tag{23}$$

where s_K is the largest possible service requirement. If $t < s_K+\delta$ then the tree consists of a single node, the root. Note that $\mathcal{T}[t,\delta]$ depends only on the values of t, δ, and the service times $s_1 < s_2 < \cdots < s_K$. See Figure 2 for some examples. Note that $\mathcal{T}[t,\delta]$ is indeed a *full* K-ary tree as follows from $s_1 < s_2 < \cdots < s_K$ and (23): if a node has a child then it must have all K children. Let $\mathcal{L}[t,\delta]$ be the the set of vectors $\underline{d}$ that correspond to leaf nodes of $\mathcal{T}[t,\delta]$.

Next, let $\mathcal{T}[t,\delta,w]$ be a weighted full K-ary tree of finitely many nodes rooted at $r(\mathcal{T}[t,\delta,w])$. The nodes of $\mathcal{T}[t,\delta,w]$ and the relationships between these nodes (or, equivalently, the corresponding vectors) are the same as the nodes of $\mathcal{T}[t,\delta]$ and the relationships between those. However, in $\mathcal{T}[t,\delta,w]$ each node (or equivalently the corresponding vector $\underline{d}$) is assigned a "weight," a real number. The weights are characterized by a *weight-function* $w\colon [0,\infty)\to\mathbb{R}$; the weight of a node corresponding to the vector $\underline{d}$ in $\mathcal{T}[t,\delta,w]$ is defined as $p_{\underline{d}}w(t-s_{\underline{d}})$. In particular, the weight of the root $r(\mathcal{T}[t,\delta,w])$ is equal to $w(t)$.

One specific weight function will be of interest in the next section:

$$\chi(t) = Y(t)1_{\{t<s_K+\delta\}} + (\hat{X}(t)-\beta)1_{\{t\ge s_K+\delta\}}, \qquad t\ge 0, \tag{24}$$

for some function $Y\in D([0,s_K+\delta),[0,\infty))$, where $\hat{X}(t)$ is a sample path of the scaled infinite-server process defined in (10). (The function Y will serve as an initial condition for the limiting virtual waiting-time process—see Theorem 3.1 in the next section.)

The weight $\mathcal{W}_{\mathcal{T}}$ of a *tree* $\mathcal{T}$ rooted at $r(\mathcal{T}[t,\delta,w])$, such that $\mathcal{T}\subseteq\mathcal{T}[t,\delta,w]$, is defined as the sum of weights of nodes that belong to $\mathcal{T}$:

$$\mathcal{W}_{\mathcal{T}} := \sum_{\underline{d}\in\mathcal{T}} p_{\underline{d}} w(t-s_{\underline{d}}); \tag{25}$$

the summation $\underline{d}\in\mathcal{T}$ is over all nodes that belong to $\mathcal{T}$; the empty tree has weight zero. (The tree $\mathcal{T}$ need not be a full K-ary tree.) Due to the recursive nature of trees, the weight of a tree can be expressed as the weight of the root plus the sum of the weights of the subtrees that are rooted at the children of the root. Formally, if $\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_K$ are the K possible subtrees of a nonempty $\mathcal{T}$, rooted at the K possible children of $r(\mathcal{T})$ (see Figure 3), with some or all $\mathcal{T}_i$'s possibly empty, then

$$\mathcal{W}_{\mathcal{T}} := w(t) + \sum_{i=1}^{K} \mathcal{W}_{\mathcal{T}_i}. \tag{26}$$

Moreover, if $\mathcal{T}_1,\ldots,\mathcal{T}_K$ are the K subtrees of the full K-ary tree $\mathcal{T}[t,\delta,w]$, $t\ge s_K+\delta$, rooted at the K children of $r(\mathcal{T}[t,\delta,w])$, then from the structure of our weights we have

$$\mathcal{T}_i = \mathcal{T}[t-s_i,\delta,p_i w].$$

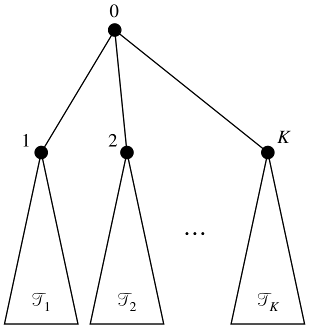

FIGURE 3. The weight of a tree is equal to the weight of the root (0) and the sum of weights of subtrees ($\mathcal{T}_1, \dots, \mathcal{T}_K$) rooted at the children of the root ($1, \dots, K$).

Unless specified otherwise, whenever we write $\mathcal{T} \subseteq \mathcal{F}$ for two trees $\mathcal{T}$ and $\mathcal{F}$, it is implicitly assumed that the trees have a common root, i.e., $r(\mathcal{T}) = r(\mathcal{F})$.

The limiting virtual waiting-time process will be described in terms of a supremum operator over weights of subtrees, which we now introduce. Combining (26) with the proceeding displayed equality yields, for all $t \geq s_K + \delta$,

$$\begin{aligned}\psi(t, \delta, w) &:= \sup_{\mathcal{T} \subseteq \mathcal{F}[t, \delta, w]} (\mathcal{W}_{\mathcal{T}})^+ \\ &= \sup_{\mathcal{T}_i \subseteq \mathcal{F}[t-s_i, \delta, w]} \left(w(t) + \sum_{i=1}^{K} (\mathcal{W}_{\mathcal{T}_i})^+ \right)^+ \\ &= \left(w(t) + \sum_{i=1}^{K} \sup_{\mathcal{T}_i \subseteq \mathcal{F}[t-s_i, \delta, w]} (\mathcal{W}_{\mathcal{T}_i})^+ \right)^+, \qquad (27)\end{aligned}$$

where the second relationship follows from $(x^+ \vee (x+y)^+)^+ = (x + y^+)^+$, for all $x, y \in \mathbb{R}$. (The sup operator in (27) can be replaced by the max operator because we consider only trees with a finite number of nodes ($t < \infty$), i.e., the number of trees $\mathcal{T} \subseteq \mathcal{F}[t, \delta, w]$ is finite.) Therefore,

$$\psi(t, \delta, w) = \left(w(t) + \sum_{i=1}^{K} p_i \psi(t - s_i, \delta, w) \right)^+, \qquad (28)$$

because, by (27), $p_i \psi(t - s_i, \delta, w)$ is the positive part of the supremum of tree weights over all trees that belong to the subtree of $\mathcal{F}[t, \delta, w]$ rooted at the ith child of the root. When $t \in [0, s_K + \delta)$, then $\mathcal{F}[t, \delta, w]$ consists of only a single node (root), and

$$\psi(t, \delta, w) = (w(t))^+, \quad t \in [0, s_K + \delta). \qquad (29)$$

In the following lemma, which concludes this section, we analyze the continuity of $\psi_\delta(w) := \{\psi(t, \delta, w),\ t \in [0, T]\}$, for $w \in D([0, T], \mathbb{R})$. To this end, introduce

$$E_{[a, T]} := \begin{cases} \{x \in D([0, T], \mathbb{R})\colon \{x(t),\ t \geq a\} \in C([a, T]), \mathbb{R})\}, & a \in (0, T), \\ D([0, T], \mathbb{R}), & a \geq T, \end{cases}$$

where $C([a, T], \mathbb{R})$ is the space of all continuous $\mathbb{R}$-valued functions on $[a, T]$; then let

$$F_{[a, T]} := \{w \in E_{[a, T]}\colon \psi(T, \delta, w) = \psi(T-, \delta, w)\}.$$

As stated earlier in this section, we consider weight functions of the form (24). Given that the sample paths of $\hat{X}$ are continuous with probability 1, it is sufficient to consider weight functions that have no discontinuity points in the time interval $[s_K + \delta, \infty)$, i.e., functions whose restrictions to $[0, T]$ belong to the set $E_{[s_K + \delta, T]}$, for $T < \infty$. Furthermore, we shall be interested in establishing convergence of $\{\psi(t, \delta, w_N),\ t \in [0, \infty)\}$ in $D([0, \infty), \mathbb{R})$, as $N \to \infty$, i.e., convergence of $\{\psi_\delta(w_N),\ t \in [0, T]\}$ in $D([0, T], \mathbb{R})$, as $N \to \infty$, for all $T < \infty$ such that $\psi(w)(T) = \psi(w)(T-)$, where $w_N \to w$ (Whitt [34, p. 83]). Hence, it is sufficient to consider the even smaller set of weight functions $F_{[s_K + \delta, T]}$.

Recall from §1 that $\theta_\tau\colon D([0, T], \mathbb{R}) \to D([0, T - \tau], \mathbb{R})$ is the shift operator. It is known that θ_τ is a continuous operator for $\tau < 0$ (Whitt [34, p. 351]).

LEMMA 3.4. *The function ψ_δ: $F_{[s_K+\delta,T]} \to D([0,T],\mathbb{R})$ is continuous for $K=1$ and all $T>0$; and for $K>1$ and all $T>0$, at those w such that for all $t\in[s_K+\delta,T]$ and all $\underline{d}$, $\underline{b}\in\mathscr{L}[t,\delta]$ such that $s_{\underline{d}}\neq s_{\underline{b}}$,*

$$\{\mathrm{Disc}(\theta_{-s_{\underline{d}}}(w))\cup\{s_K+\delta+s_{\underline{d}}\}\}\cap\{\mathrm{Disc}(\theta_{-s_{\underline{b}}}(w))\cup\{s_K+\delta+s_{\underline{b}}\}\}=\varnothing. \tag{30}$$

REMARK 3.3. Note that if $w\in F_{[s_K+\delta,T]}$ then $\mathrm{Disc}(w)\subset[0,s_K+\delta]$. The set of potential discontinuity points of the $\mathbb{R}$-valued function $\psi_\delta(w)(t)$, $t\geq 0$, can be established based on $\mathrm{Disc}(w)$ (whether a discontinuity point exists or not depends also on the value of $\psi_\delta(w)(t)$; the operator $(\cdot)^+$ can eliminate some of the potential discontinuity points—for example, if $K=1$ and $-w((x+s_1)-)=-w(x+s_1)>\psi(x-,\delta,w)>\psi(x,\delta,w)>0$ for some $x>s_1+\delta$, then $w((x+s_1)-)+\psi(x-,\delta,w)>w(x+s_1)+\psi(x,\delta,w)$ and yet $(w((x+s_1)-)+\psi(x-,\delta,w))^+=(w(x+s_1)+\psi(x,\delta,w))^+$, rendering $\psi((x+s_1)-,\delta,w)=\psi(x+s_1,\delta,w)=0$). Let $\mathscr{D}_w=\mathrm{Disc}(w)\backslash\{s_K+\delta\}$. Potential discontinuity points of the $\mathbb{R}$-valued $\psi_\delta(w)(t)$ are of the form $x+s_{\underline{d}}\in\mathbb{R}$, where $x\in\mathscr{D}_w$ and $\underline{d}\in\mathscr{L}[t,\delta]$. This is due to the fact that $\psi_\delta(w)(t)$ depends on the weights of leaf nodes and these weights experience discontinuities at points in $\mathscr{D}_w$. Namely, if $\mathscr{W}_{\mathscr{T}}(w)(t)$ is a weight of $\mathscr{T}\subseteq\mathscr{F}[t,\delta,w]$, for fixed δ and w, then from (25) it is evident that discontinuities of $\mathscr{W}_{\mathscr{T}}(w)(t)$, $t\geq 0$, can occur only at t satisfying $t-s_{\underline{d}}\in\mathscr{D}_w$ for some $\underline{d}\in\mathscr{T}$. However, $\psi_\delta(w)(\cdot)$ is a maximum of $\mathscr{W}_{\mathscr{T}}(w)(\cdot)$ over a finite set of trees $\mathscr{T}$, and, hence, any discontinuity of $\psi_\delta(w)(\cdot)$ must also be a discontinuity of some $\mathscr{W}_{\mathscr{T}}(w)(\cdot)$. The point $\{s_K+\delta\}$ is special (see (28) and (29)). Consider a node $\underline{d}$ in the tree $\mathscr{F}[t,\delta,w]$ and the weight of a subtree (a full K-ary tree) rooted at that specific node. Then this weight (as a function of t) can experience a discontinuity at $t=s_{\underline{d}}+s_K+\delta$ even if $\{s_K+\delta\}\in\mathrm{Disc}(w)$. In particular, no discontinuity occurs only if $w((s_K+\delta)-)=(w(s_K+\delta)-\beta+\sum_{i=1}^K p_i w(s_K+\delta-s_i))^+$.

REMARK 3.4. Condition (30) ensures that the summation in (28) is continuous in $D([0,\infty),\mathbb{R})$ because the set of common discontinuity points of the summands is an empty set [33]. The need for (30) is illustrated by the following example. Let $s_2=2s_1=2$, $p_1=p_2=1/2$, and $\delta=1/4$. Suppose that $w_n\in D([0,13/4],\mathbb{R})$ is defined by $w_n(t):=1_{\{t\in[1/2,3/4-1/n)\}}+1_{\{t\in[7/4+1/n,2)\}}$. Then $w_n\to w$ in $D([0,13/4],\mathbb{R})$, as $n\to\infty$, where $w(t)=1_{\{t\in[1/2,3/4)\}}+1_{\{t\in[7/4,2)\}}$. However, due to the nonnegativity of w_n and w it follows that $\psi_\delta(w_n)=\{\mathscr{W}_{\mathscr{F}[t,\delta,w_n]},\ t\in[0,13/4]\}=\{1_{\{t\in[1/2,3/4-1/n)\}}+1_{\{t\in[7/4+1/n,2)\}}+1_{\{t\in[5/2,11/4-1/n)\}}/2+1_{\{[11/4+1/n,3)]\}}/2,\ t\in[0,13/4]\}$ and $\psi_\delta(w)=\{\mathscr{W}_{\mathscr{F}[t,\delta,w]},\ t\in[0,13/4]\}=\{1_{\{t\in[1/2,3/4)\}}+1_{\{t\in[7/4,2)\}}+1_{\{t\in[5/2,3)\}}/2,\ t\in[0,13/4]\}$, yielding that $\psi_\delta(w_n)$ does not converge to $\psi_\delta(w)$, as $n\to\infty$. Note that (30) is violated in this example because

$$\mathrm{Disc}(\theta_{-s_1}(w))\cap\mathrm{Disc}(\theta_{-s_2}(w))=\{3/2,7/4,11/4,3\}\cap\{5/2,11/4,15/4,4\}=\{11/4\}\neq\varnothing.$$

PROOF. The lemma holds trivially for $T\leq s_K+\delta$ due to (29), and, hence, we examine only $T>s_K+\delta$.
($K>1$). First, we consider $T\in[s_K+\delta,s_K+s_1+\delta)$. For these values of T, the operator $\psi_\delta(w)$ is given by

$$\psi(t,\delta,w)=\left(w(t)+1_{\{t\in[s_K+\delta,\,s_K+s_1+\delta)\}}\sum_{i=1}^K p_i w((t-s_i)^+)\right)^+,\quad 0\leq t\leq T.$$

The J_1 continuity of $\psi_\delta(w)$ follows from the continuity of the $(\cdot)^+$ operator (by the definition of the J_1 metric), continuity of the time-shift operator (Glynn and Whitt [15]), continuity of addition when the set of common discontinuity points of the summands is an empty set [33], and the assumptions of the lemma on the discontinuity set of w (noting that w is continuous at $t=s_K+\delta-s_i$ for all i).

The proof for general values of T is by induction. The case $0<T<s_K+s_1+\delta$ serves as the base of the induction. Next, we describe the inductive step. Suppose that the lemma holds for some $T_0>s_K+\delta+s_1/2$ and consider $w\in F_{[s_K+\delta,T]}$ with $T\in(T_0,T_0+s_1/2]$. Because $\psi_\delta(w)\in D([0,T],\mathbb{R})$, the number of discontinuity points of $\psi_\delta(w)$ is either finite or infinitely countable (Whitt [34, p. 393]). This implies that there exists $\tau\in(T_0-s_1/2,T_0]$ that is a point of continuity of $\psi_\delta(w)$. Element $\psi_\delta(w)\in D([0,T],\mathbb{R})$ can be defined as a function of $\psi_\delta(w)\in D([0,\tau],\mathbb{R})$ and $\{w(t),\ t\in[\tau,T]\}\in C([\tau,T],\mathbb{R})$. Namely, if $\varphi_{\tau,T}$: $D([0,\tau],\mathbb{R})\times C([\tau,T],\mathbb{R})\to D([0,T],\mathbb{R})$ is defined by

$$\begin{aligned}\varphi_{\tau,T}(y,x)(t) &:= y(t)1_{\{t\in[0,\tau)\}}+\left(x(t)+\sum_{i=1}^K p_i y(t-s_i)\right)^+1_{\{t\in[\tau,T]\}}\\ &:= y(t)1_{\{t\in[0,\tau)\}}+\gamma(t)1_{\{t\in[\tau,T]\}},\end{aligned} \tag{31}$$

then

$$\{\psi(t,\delta,w),t\in[0,T]\}=\varphi_{\tau,T}(\{\psi(t,\delta,w),\ t\in[0,\tau]\},\{w(t),\ t\in[\tau,T]\}).$$

The lemma thus holds for $T\in(T_0,T_0+s_1/2]$ due to the inductive assumption and the continuity of the operator $\varphi_{\tau,T}$. The latter continuity of $\varphi_{\tau,T}$ at w, which satisfies the assumptions of the lemma, follows from the continuity of the $(\cdot)^+$ operator, continuity of the shift operator (Glynn and Whitt [15]), continuity of addition when the set of common discontinuity points of the summands is an empty set [33], and

$$d_{[0,T]}(\varphi_{\tau,\Delta}(y_1, x_1), \varphi_{\tau,\Delta}(y_2, x_2)) \leq d_{[0,\tau]}(y_1, y_2) + d_{[\tau,T]}(\gamma_1, \gamma_2),$$

where $\gamma_1, \gamma_2 \in D([\tau, T], \mathbb{R})$ are defined as in (31). This concludes the proof for $K > 1$.

$(K = 1)$. Condition (30) is not needed in this case because (28) reduces to $\psi(t, \delta, w) = (w(t) + \psi(t-s, \delta, w))^+$ (s is the service time) and w has no discontinuity points on $[s_K + \delta, \infty)$ by the definition of set $E_{[s+\delta, T]}$. □

3.4. The limit of $\widehat{V}_N(t)$. We are now ready to formulate our main result—Theorem 3.1 below. The theorem is proved in §3.6, and it enables one to approximate the virtual waiting time over $t > s_K$. Some nonnegative stochastic process Y with sample paths in $D([0, s_K + \delta), [0, \infty))$, as indicated in (24), provides the initial condition for the (asymptotic) virtual waiting time. This condition is specified over the time interval $[0, s_K + \delta)$, with $\delta > 0$ arbitrarily small, as elaborated on in §3.5.

THEOREM 3.1. *Consider a sequence of $G/D_K/N$ queues* (§2.2) *in the QED regime* (§3.1). *Recall that S denotes a generic service time, the distribution of which is assumed to have a support $0 < s_1 < \cdots < s_K < \infty$ of finite cardinality. Let $\underline{\widehat{X}}_N$ be the corresponding sequence of infinite-server vector processes* (*scaled and centered, vanishing at $t = 0$*), *with $\underline{\widehat{X}}$ being its Gaussian limit as in Lemma* 2.5, *and consider $\widehat{V}_N = \{\widehat{V}_N(t),\ t \geq 0\}$, the sequence of scaled virtual waiting-time processes defined in* (19).

Suppose that for some arbitrary $\delta > 0$, as $N \to \infty$,

$$(\underline{\widehat{X}}_N, \widehat{V}_N) \;\Rightarrow\; (\underline{\widehat{X}}, Y) \tag{32}$$

in $D([0, s_K + \delta), \mathbb{R}^{K+1})$. Then, as $N \to \infty$,

$$\widehat{V}_N \;\Rightarrow\; \psi_\delta(\chi) \tag{33}$$

in $D([0, \infty), \mathbb{R})$, assuming

$$\mathbb{P}[\exists t \in [s_K + \delta, \infty)\colon \mathscr{D}(\underline{d}) \cap \mathscr{D}(\underline{b}) \neq \varnothing,\ s_{\underline{b}} \neq s_{\underline{d}},\ \underline{d}, \underline{b} \in \mathscr{L}[t, \delta]] = 0, \tag{34}$$

where $\mathscr{D}(\underline{c}) := \mathrm{Disc}(\theta_{-s_{\underline{c}}}(Y)) \cup \{s_K + \delta + s_{\underline{c}}\}$, $\underline{c} \in \mathscr{L}[t, \delta]$ (*Lemma* 3.4). *In* (33), *the process $\psi_\delta(w) := \{\psi(t, \delta, w),\ t \geq 0\}$ is defined by* (28) *and* (29), *with its weight function χ given by*

$$\chi(t) = Y(t)1_{\{t < s_K + \delta\}} + (\widehat{X}(t) - \beta)1_{\{t \geq s_K + \delta\}},$$

where $\widehat{X}$ is the scalar Gaussian process defined in (10). (*Recall that $\underline{\widehat{X}}$ and hence $\widehat{X}$ both vanish at $t = 0$.*)

REMARK 3.5. The appropriateness of the J_1 topology (Whitt [34, §3.3]) can be illustrated on the following simple example. Consider an initially empty system with deterministic service times $S = s$. We demonstrate possibly a discontinuity of the virtual waiting time at $t = s$, in the limit as $N \to \infty$. To this end, let τ_1 and τ_N be the time of the first and Nth arrival to the Nth system. Then $V_N(t) = 0$ for all $t < \tau_N$; i.e., the virtual waiting time can become positive the earliest at $t = \tau_N$. Indeed, at time $t = \tau_N$, the virtual waiting time $V_N(\tau_N)$ is given by $(\tau_1 + s - \tau_N)^+$, because $(\tau_1 + s)$ is the time of the first departure from the system; see Figure 4 for details. Now, in the limit, as $N \to \infty$, we have $\tau_N \to s$ with probability 1. On the other hand, the size of the jump $\widehat{V}_N(\tau_N) - \widehat{V}_N(\tau_N-) = \widehat{V}_N(\tau_N) - 0 = \widehat{V}_N(\tau_N)$ satisfies, as $N \to \infty$,

$$\widehat{V}_N(\tau_N) = \frac{\sqrt{N}}{s}(\tau_1 + s - \tau_N)^+ \;\Rightarrow\; (\widehat{X}(s) - \beta)^+;$$

the limit follows from $\sqrt{N}\tau_1 \Rightarrow 0$, as $N \to \infty$, the fact that

$$\{\sqrt{N}(s - \tau_N) \geq sx\} = \{\hat{A}_N(s(1 - o(1))) - \beta(1 + o(1)) \geq x\},$$

the FCLT (4) for A_N, and $\hat{A}(s) = \widehat{X}(s)$ by Lemma 2.5. The fact that both the time of the jump τ_N and its size $\widehat{V}_N(\tau_N)$ converge implies that the J_1 topology is appropriate to handle this discontinuity—see Whitt [34, p. 79]. (We continue this example in the following section, under the heading "Empty system.")

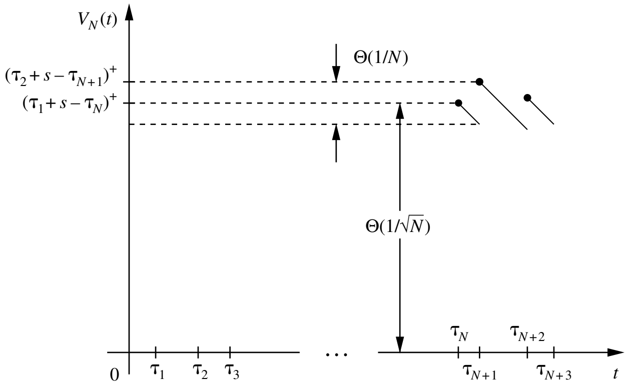

FIGURE 4. The behavior of the virtual waiting time for a system with a large (but finite) number of servers when the system is empty at $t=0$. Let τ_i be the arrival time of the ith customer. The first N arriving customers do not experience waiting, and, thus, the virtual waiting time satisfies $V_N(t)=0$ for all $t\in[0,\tau_N)$. At $t=\tau_N$, the function V_N experiences a jump of size $(\tau_1+s-\tau_N)^+=\Theta(1/\sqrt{N})$. The next jump of V_N occurs at $t=\tau_{N+1}$. However, the size of that jump is asymptotically negligible in comparison to the first one, because it is $((\tau_2+s-\tau_{N+1})^+-((\tau_1+s-\tau_N)^+-\tau_{N+1}+\tau_N)^+)^+\le(\tau_2-\tau_1)+(\tau_{N+1}-\tau_N)=\Theta(1/N)$.

REMARK 3.6. Condition (34) is a stochastic version of (30), which is the subject of Lemma 3.4 and the two remarks subsequent to it. Note that condition (34) is trivially satisfied whenever Y is continuous with probability 1. (This is the case in the "Gradual-departures example" of the following section.) More generally, for given values of service times, condition (34) places a restriction on where the discontinuities of Y can occur. For example, suppose that S is lattice-valued, i.e., $s_i=i\Delta$, $i=1,\dots,K$, for some $\Delta>0$. Then (34) prevails if one has $|x-y|\neq k\Delta$ for all $k\in\mathbb{N}$ and all $x,y\in\mathrm{Disc}(Y)\cup\{s_K+\delta\}$ such that $x,y\ge\delta$, with probability 1. (In particular, the condition is satisfied for the example of the previous remark by choosing $\delta<s$.) In words, due to the tree structure, each discontinuity point x of Y can, with positive probability, result in discontinuities of $\widehat{V}$ at points $x+i\Delta$, where i is an integer satisfying $i\ge(s_K+\delta-x)/\Delta$. Then, the above condition ensures that two different discontinuity points of Y do not produce a discontinuity of $\widehat{V}$ at the same time instance.

3.5. Initial conditions. Assumption (32) of Theorem 3.1 is one on the "initial" state of the system. The pair $(\underline{\widehat{X}},Y)\in D([0,s_K+\delta),\mathbb{R}^{K+1})$ specifies the initial conditions of the system, in the limit as $N\to\infty$; i.e., it implicitly defines the *state* of the queue (the number of customers in the system and their residual service times) at time $t=0$. Equivalently, the state of the queue at time $t=0$, jointly with the sequence of arrivals and service requirements on $[0,s_K+\delta)$ (summarized by $\underline{\widehat{X}}$), determine the asymptotic virtual waiting time Y on that same time interval $[0,s_K+\delta)$. Condition (32) is a convenient way to capture the initial conditions because it involves weak convergence of processes with dimensionality independent of the system size N. In addition, note that $\widehat{V}_N$ is the primary (scalar) process of interest while X_N depends on arrivals after $t=0$ only ($\widehat{X}_N(0)=0$, while the system need not be empty at time $t=0$). The prevalent approach in the literature for summarizing initial conditions has become a random measure that describes the state of the system at $t=0$, e.g., see [16]. Evaluating conditions on such random measures that lead to the QED regime is beyond the scope of this paper.

A parallel should be drawn with initial conditions that appear in the asymptotic analysis of systems where *finite-dimensional* diffusion processes arise in the limit, e.g., in the classical heavy-traffic regime or the QED regime with Markovian structure. There, typically, a simple (finite-dimensional) initial condition is specified at $t=0$, e.g., see Halfin and Whitt [18] and Puhalskii and Reiman [31]. However, due to the nonMarkovian structure of the number of customers in our present system, its initial condition must be specified as an evolution over a time interval the length of which is related to the largest service requirement. Effectively, the "memory" of the system at time t (in the limit as $N\to\infty$) is its history over $[t-s_K,t)$, s_K being the largest service requirement, and, hence, initial conditions must be specified over at least $[0,s_K]$.

Next, we provide examples that illustrate how Y arises (or cannot arise) from the queue state at $t=0$.

- *Empty system.* Consider an empty system at $t=0$ and observe that the event $\{X_N(t)\le N,\ \forall\, t\in[0,T]\}$ implies $\{V_N(t)=0,\ \forall\, t\in[0,T]\}$. This observation and Lemma 2.4 yield $V_N(t)=0$, for all $t\in[0,s_K-\delta)$, δ small enough and N large enough. Furthermore, for any $\varepsilon>0$, (3) also renders $V_N(t)<\varepsilon$, for all $t\in[0,s_K+\delta)$, δ small enough and N large enough. This, and the arguments used in the proof of Lemma 3.2, yield, for $t\in[s_K,s_K+\delta)$,

$$(\rho_N^{-1}\widehat{X}_N^{\downarrow\varepsilon}(t)-\beta_N)^+\le\widehat{V}_N(t)\le(\rho_N^{-1}\widehat{X}_N^{\uparrow\varepsilon}(t)-\beta_N+\delta_N)^+ \tag{35}$$

where $\varepsilon > 0$ is arbitrary and N is large enough; β_N and δ_N are as in Lemma 3.2; in order to avoid repetition we omit the details. Using (35) and the arguments used in the proof of Theorem 3.1 one can show that (32) holds with

$$Y(t) = (\hat{X}(t) - \beta)^+ 1_{\{s_K \le t < s_K + \delta\}}. \tag{36}$$

The derivation of (36) requires proving a version of Theorem 3.1 for the special case when the depth of the tree is at most 1. Furthermore, there exists a $\delta > 0$ such that (34) is also satisfied.

- *Overloaded system.* Consider service times that are deterministic: $S = s$. Let the number of customers in the queue at $t = 0$ be equal to N with *all* the remaining service requirements equal to s. It is then straightforward to verify that $V_N(t) = s - t + s\lfloor A_N(t)/N \rfloor$, over $t \in [0, s)$. As will be shown momentarily, the scaled $\hat{V}_N$ does not converge in $D([0, s), \mathbb{R})$; hence, no suitable Y exists. Moreover, $\hat{V}_N$ does not converge on any time interval $[a, b)$, $0 \le a < b < \infty$. The system does not operate in the QED regime in this case.

The system, in fact, is in the ED (efficiency-driven) regime [36], where the unscaled virtual waiting-time process V_N converges by itself. To see that, consider a finite $T \ne s \cdot i$, $i = 1, 2, \ldots$. Due to (3), for any $\varepsilon > 0$ and all N large enough,

$$\sup_{t \in [0, T]} |sN^{-1} A_N(t) - t| \le \varepsilon.$$

Hence, for all $t \in [0, T]$ one has $(t - \varepsilon)^+ \le sN^{-1} A_N(t) \le t + \varepsilon$, implying

$$\sum_{i=0}^{\infty} \theta_{-is}((s-t) 1_{\{t \in [2\varepsilon, s)\}}) \le V_N(t) \le \sum_{i=0}^{\infty} \theta_{-is}((s-t) 1_{\{t \in [0, s)\}} + 1_{\{t \in [s - 2\varepsilon, s)\}}),$$

where $\theta_\cdot(\cdot)$ is the shift operator defined in §1. However, the two bounding expressions in the preceding inequality are independent of N and converge to the same limit (as $\varepsilon \to 0$):

$$\sum_{i=0}^{\infty} \theta_{-is}((s-t) 1_{\{t \in [0, s)\}}) = s - t + \lfloor t/s \rfloor$$

for $t \ge 0$. Consequently, $V_N \to \{s - t + \lfloor t/s \rfloor,\ t \ge 0\}$ in $D([0, \infty), \mathbb{R})$ with probability 1, as $N \to \infty$.

- *Gradual-departures system.* Suppose that $S = s$ and the number of customers in the system at $t = 0$ is equal to $\lfloor \lambda_N s \rfloor$, for the Nth system. Let the residual service times of these customers at $t = 0$ be given by $si/\lfloor \lambda_N s \rfloor$, $i = 1, \ldots, \lfloor \lambda_N s \rfloor$; i.e., they depart in a deterministic fashion; a departure occurs every $s/\lfloor \lambda_N s \rfloor$ units of time. From Lemma 2.1 it is immediate that, as $N \to \infty$,

$$\{\hat{V}_N(t),\ t \in [0, s - \delta]\} \;\Rightarrow\; \{(\hat{X}(t) - \beta)^+,\ t \in [0, s - \delta]\}, \tag{37}$$

in $D([0, s - \delta], \mathbb{R})$, for any $\delta \in (0, s)$. The limit (37), Lemma 3.2, and the fact that sample paths of $\hat{X}$ are continuous with probability 1 yield that (32) holds for any $\delta \in (0, s)$, with Y given by

$$Y(t) = (\hat{X}(t) - \beta + (\hat{X}(t - s) - \beta)^+)^+,$$

where $\hat{X}(t) \equiv 0$, for all $t < 0$, as assumed throughout the paper. Note that, in this case, Y has continuous sample paths with probability 1 so (34) holds.

The prevalent characterization of the QED regime is based on steady-state behavior. In this paper, however, we analyze transient behavior, which strongly depends on the initial conditions. Therefore, we extend the prevalent QED definition and say that the limiting system is in the QED regime, at a given time $t > 0$, if the limiting probability of delay at time t is *strictly* in $(0, 1)$. The three examples above illustrate that the system need not start in the QED regime at time $t = 0$ in order to eventually get there and that the time to reach this regime depends on the initial conditions. Specifically, in the case of the overloaded system, the QED regime is never reached and the system operates in the ED (efficiency-driven) regime. On the other hand, the gradual-departures example represents the case when the system is in the QED regime at all $t > 0$. Finally, when the system is initially empty, it starts in the QD (quality-driven) regime and then reaches the QED regime at $t = s_K$. The system then remains in the QED regime at all $t > s_K$. Indeed, the assumption of $\delta > 0$ in Theorem 3.1 guarantees that the delay probability is positive for all $t \ge s_K + \delta$ for all N large enough, when the system starts empty at $t = 0$, because, in the limit, as $N \to \infty$, the delay probability at time $t \ge s_K + \delta$ is lower bounded by $\mathbb{P}[(\hat{X}(t) - \beta)^+ > 0]$ (see (36) and Theorem 3.1). In this example the system undergoes a change in the operating regime from QD to QED. (Regime changes of a somewhat similar nature were considered in the context of time-varying queues in [27].)

Finally, we point out that, although there do exist infinitely many δ's that can be applicable in Theorem 3.1, the resulting $\widehat{V}$ does not change with δ as we now explain. For $Y \in D([0, s_K+\delta), [0,\infty))$, $\delta > 0$, satisfying (34) and $\widehat{X} \in D([0,\infty), \mathbb{R})$, Theorem 3.1 implies a virtual waiting-time process $\widehat{V}$. Now, consider the restriction of $\widehat{V}$ to $[0, s_K+\delta+\varepsilon)$, where $\varepsilon > 0$ is such that (34) is satisfied with $(\delta+\varepsilon)$ instead of δ; denote this restriction by Y'. Then, the pair Y' and $\widehat{X}$ imply, via Theorem 3.1, the same virtual waiting-time process $\widehat{V}$. This consistency is due to the fact that all weights of the corresponding nonleaf nodes are identical in the trees $\mathcal{T}[t,\delta,\chi(Y,\widehat{X})]$ and $\mathcal{T}[t,\delta,\chi(Y',\widehat{X})]$, $t \ge s_K+\delta+\varepsilon$—hence, the equivalence of the two suprema over these trees and, consequently, that of the two virtual waiting-time processes.

3.6. Proof of Theorem 3.1.

Proof. First, the definition of the scaled process $\widehat{V}_N$ yields the following equality for all δ and ε:

$$\mathbb{P}\Big[\sup_{t\in[0,\,s_K+\delta)} V_N(t) \le \varepsilon\Big] = \mathbb{P}\Big[\sup_{t\in[0,\,s_K+\delta)} \widehat{V}_N(t) \le \sqrt{N}\varepsilon/\mathbb{E}S\Big].$$

This equality, the fact that $\widehat{V}_N \Rightarrow Y$ in $D([0, s_K+\delta), \mathbb{R})$, and Corollary 3.1 imply that for each $T < \infty$ and all $\varepsilon > 0$ small enough, as $N \to \infty$,

$$\mathbb{P}\Big[\sup_{t\in[0,\,T]} V_N(t) \le \varepsilon\Big] \to 1. \tag{38}$$

This limit will enable us to invoke Lemma 3.2 and Lemma 3.3.

Second, let $\mathcal{T}[t,\delta,\chi_N^{\uparrow\varepsilon}]$ be a weighted tree (as defined in §3.3) with a weight function

$$\chi_N^{\uparrow\varepsilon}(t) := \sup_{u\in[(t-\varepsilon)^+,\,t+\varepsilon]} \xi^{\uparrow}_{\varepsilon,N}(u) 1_{\{t<s_K+\delta\}} + (\rho_N^{-1}\widehat{X}_N^{\uparrow\varepsilon}(t) - \beta_N + \delta_N) 1_{\{t\ge s_K+\delta\}},$$

for $t \ge 0$, where

$$\xi^{\uparrow}_{\varepsilon,N}(t) = \widehat{V}_N(t) 1_{\{t<s_K+\delta\}} + (\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow\varepsilon}(t-S)] + \rho_N^{-1}\widehat{X}_N^{\uparrow\varepsilon}(t) - \beta_N + \delta_N)^+ 1_{\{t\ge s_K+\delta\}},$$

$t \ge 0$. Note that $\{\chi_N^{\uparrow\varepsilon}(t),\, t \ge \varepsilon\}$ depends on the values of the process $\{\widehat{V}_N(t),\, t \ge 0\}$. However, only values of $\widehat{V}_N(t)$ on the interval $[0, s_K+\delta)$ are relevant, i.e., the interval on which the initial condition is set (see (32)).

Next, in several steps, we show that, under the assumptions of the theorem, for each $T < \infty$ there exists a finite constant $c \equiv c_T$ such that for all $\varepsilon > 0$ small enough, as $N \to \infty$,

$$\mathbb{P}[\widehat{V}_N^{\uparrow\varepsilon}(t) \le \psi(t,\delta,\chi_N^{\uparrow c\varepsilon}),\ \forall\, t \in [0,T]] \to 1. \tag{39}$$

The proof of (39) is by induction. In order to verify the base of the induction, we will consider $T < s_K+\delta$. For $0 \le t < T < s_K+\delta$, $\mathcal{T}[t,\delta,\chi_N^{\uparrow c\varepsilon}]$ consists of a single node (root) with the weight

$$\sup_{u\in[(t-\varepsilon)^+,\,t+\varepsilon]} \xi^{\uparrow}_{c\varepsilon,N}(u).$$

However, $\xi^{\uparrow}_{c\varepsilon,N}(t) = \widehat{V}_N(t)$ for $t \in [0, s_K+\delta)$ by definition, and Lemma 3.3 with (38) implies that for $c = 2$ and all $\varepsilon > 0$ small enough

$$\begin{aligned}
&\mathbb{P}[\widehat{V}_N(t) \le \xi^{\uparrow}_{c\varepsilon,N}(t),\ \forall\, t \in [s_K+\delta, s_K+\delta+\varepsilon]]\\
&\quad = \mathbb{P}\big[\widehat{V}_N(t) \le \big(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow c\varepsilon}(t-S)] + \rho_N^{-1}\widehat{X}_N^{\uparrow c\varepsilon}(t) - \beta_N + \delta_N\big)^+,\ \forall\, t \in [s_K+\delta, s_K+\delta+\varepsilon]\big] \to 1,
\end{aligned}$$

as $N \to \infty$. Therefore, for $c = 2$ and all $\varepsilon > 0$ sufficiently small we have

$$\mathbb{P}[\widehat{V}_N^{\uparrow\varepsilon}(t) \le \psi(t,\delta,\chi_N^{\uparrow c\varepsilon}),\ \forall\, t \in [0,\ s_K+\delta)] \to 1,$$

as $N \to \infty$. This provides a base for the induction.

Assume now that (39) holds for some $s_K+\delta \le T = T_0 < \infty$, $2 \le c = c_0 < \infty$ and all $\varepsilon > 0$ small enough. Next, we demonstrate that (39) holds for some $T > T_0$, such that $(T-T_0)$ is independent of the value of T_0. To this end, (38) ensures that Lemma 3.3 is applicable:

$$\mathbb{P}\big[\widehat{V}_N^{\uparrow\varepsilon}(t) \le \big(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 3\varepsilon}(t-S)] + \chi_N^{\uparrow 3\varepsilon}(t)\big)^+,\ \forall\, t \in (T_0, T_0+s_1-3\varepsilon]\big] \to 1, \tag{40}$$

as $N \to \infty$. The monotonicity of the sup operator and (28) result in

$$\chi_N^{\uparrow 3\varepsilon}(t) + \sum_{i=1}^{K} p_i \psi(t - s_i, \delta, \chi_N^{\uparrow 3c_0\varepsilon}) \leq \chi_N^{\uparrow 3c_0\varepsilon}(t) + \sum_{i=1}^{K} p_i \psi(t - s_i, \delta, \chi_N^{\uparrow 3c_0\varepsilon}) = \psi(t, \delta, \chi_N^{\uparrow 3c_0\varepsilon}). \tag{41}$$

Then, (40), the inductive assumption, and (41) yield that (39) holds for $T < T_0 + s_1 - 3\varepsilon$ with $c = 3c_0$. This proves that (39) holds for every $T < \infty$.

By repeating the preceding steps of this proof it is straightforward to argue that, under the conditions of the theorem, for each $T < \infty$ and all $\varepsilon > 0$ small enough, a lower bound holds as well; i.e., as $N \to \infty$,

$$\mathbb{P}[\psi(t, \delta, \chi_N^{\downarrow\varepsilon}) \leq \hat{V}_N(t) \leq \psi(t, \delta, \chi_N^{\uparrow\varepsilon}), \ \forall\, t \in [0, T]] \to 1, \tag{42}$$

where the weight function $\chi_N^{\downarrow\varepsilon}(t)$ is defined by

$$\chi_N^{\downarrow\varepsilon}(t) := \inf_{u \in [(t-\varepsilon)^+, t+\varepsilon]} \xi_{\varepsilon, N}^{\downarrow}(u)\, 1_{\{t < s_K + \delta\}} + (\rho_N^{-1} \hat{X}_N^{\downarrow\varepsilon}(t) - \beta_N) 1_{\{t \geq s_K + \delta\}},$$

with

$$\xi_{\varepsilon, N}^{\downarrow}(t) = \hat{V}_N(t) 1_{\{t < s_K + \delta\}} + (\mathbb{E}_{\langle S \rangle}[\hat{V}_N^{\downarrow\varepsilon}(t - S)] + \rho_N^{-1} \hat{X}_N^{\downarrow\varepsilon}(t) - \beta_N)^+ 1_{\{t \geq s_K + \delta\}}.$$

Third, Lemma 2.5, (32), Lemma 3.4, the continuity of sample paths of $\hat{X}$ (with probability 1) and the continuous mapping theorem yield

$$(\psi_\delta(\chi_N), \psi_\delta(\chi_N^{\uparrow\varepsilon}), \psi_\delta(\chi_N^{\downarrow\varepsilon})) \ \Rightarrow\ (\psi_\delta(\chi), \psi_\delta(\chi^{\uparrow\varepsilon}), \psi_\delta(\chi^{\downarrow\varepsilon})) \tag{43}$$

in $D([0, \infty), \mathbb{R}^3)$, as $N \to \infty$, where

$$\chi^{\uparrow\varepsilon}(t) = \sup_{u \in [(t-\varepsilon)^+, t+\varepsilon]} \xi_\varepsilon^{\uparrow}(u)\, 1_{\{t < s_K + \delta\}} + (\hat{X}^{\uparrow\varepsilon}(t) - \beta) 1_{\{t \geq s_K + \delta\}},$$

$$\chi^{\downarrow\varepsilon}(t) = \inf_{u \in [(t-\varepsilon)^+, t+\varepsilon]} \xi_\varepsilon^{\downarrow}(u)\, 1_{\{t < s_K + \delta\}} + (\hat{X}^{\downarrow\varepsilon}(t) - \beta) 1_{\{t \geq s_K + \delta\}},$$

$$\xi_\varepsilon^{\uparrow}(t) = Y(t) 1_{\{t < s_K + \delta\}} + (\mathbb{E}_{\langle S \rangle}[Y^{\uparrow\varepsilon}(t - S)] + \hat{X}^{\uparrow\varepsilon}(t) - \beta)^+ 1_{\{t \geq s_K + \delta\}},$$

$$\xi_\varepsilon^{\downarrow}(t) = Y(t) 1_{\{t < s_K + \delta\}} + (\mathbb{E}_{\langle S \rangle}[Y^{\downarrow\varepsilon}(t - S)] + \hat{X}^{\downarrow\varepsilon}(t) - \beta)^+ 1_{\{t \geq s_K + \delta\}}.$$

Furthermore, Lemma 3.4 and the continuous-mapping theorem result in, as $\varepsilon \downarrow 0$,

$$(\psi_\delta(\chi^{\uparrow\varepsilon}), \psi_\delta(\chi^{\downarrow\varepsilon})) \ \Rightarrow\ (\psi_\delta(\chi), \psi_\delta(\chi)) \tag{44}$$

in $D([0, \infty), \mathbb{R}^2)$. Finally, the statement of the theorem is due to (42), (43), and (44). Namely, let f be an arbitrary bounded (say, $|f| \leq c_f$ for some $c_f \in [0, \infty)$), continuous, real-valued function on $D([0, T], \mathbb{R})$ (see Whitt [34, p. 83]). Then, $\mathbb{E} f(\hat{V}_N)$ can be bounded as follows

$$\mathbb{E}\Big[\inf_{G \in \mathscr{G}_N^\varepsilon} f(G)\Big] - c_f \mathbb{P}[\overline{\mathscr{E}}_{\varepsilon, T}] \leq \mathbb{E} f(\hat{V}_N) \leq \mathbb{E}\Big[\sup_{G \in \mathscr{G}_N^\varepsilon} f(G)\Big] + c_f \mathbb{P}[\overline{\mathscr{E}}_{\varepsilon, T}], \tag{45}$$

where $\mathscr{G}_N^\varepsilon := \{G \in D([0, T], \mathbb{R})\colon d_{[0, T]}(G, \chi_N) \leq d_{[0, T]}(\chi_N^{\downarrow\varepsilon}, \chi_N^{\uparrow\varepsilon})\}$ and event $\overline{\mathscr{E}}_{\varepsilon, T}$ is the complement of event

$$\mathscr{E}_{\varepsilon, T} := \{\psi(t, \delta, \chi_N^{\downarrow\varepsilon}) \leq \hat{V}_N(t) \leq \psi(t, \delta, \chi_N^{\uparrow\varepsilon}), \ \forall\, t \in [0, T]\};$$

note that $\{G \in D([0, T], \mathbb{R})\colon \psi(t, \delta, \chi_N^{\downarrow\varepsilon}) \leq G(t) \leq \psi(t, \delta, \chi_N^{\uparrow\varepsilon}), \ \forall\, t \in [0, T]\} \subseteq \mathscr{G}_N^\varepsilon$. Passing $N \to \infty$ in (45) and making use of (42), (43) yields

$$\liminf_{N \to \infty} \mathbb{E} f(\hat{V}_N) \leq \mathbb{E}\Big[\sup_{G \in \mathscr{G}^\varepsilon} f(G)\Big],$$

$$\limsup_{N \to \infty} \mathbb{E} f(\hat{V}_N) \geq \mathbb{E}\Big[\inf_{G \in \mathscr{G}^\varepsilon} f(G)\Big],$$

where $\mathscr{G}^\varepsilon := \{G \in D([0, T], \mathbb{R})\colon d_{[0, T]}(G, \chi) \leq d_{[0, T]}(\chi^{\downarrow\varepsilon}, \chi^{\uparrow\varepsilon})\}$. Letting $\varepsilon \downarrow 0$ in the preceding two inequalities and recalling (44) renders

$$\mathbb{E} f(\hat{V}_N) \to \mathbb{E} f(\psi_\delta(\chi)),$$

as $N \to \infty$. This completes the proof. □

3.7. Deterministic service time. The distribution of $\sup(\mathscr{W}_{\mathcal{T}})^+$, which appears implicitly in the statement of Theorem 3.1 (see (27)), can be difficult to evaluate analytically, in general. However, in the special case of deterministic service times, the expression for $\sup(\mathscr{W}_{\mathcal{T}})^+$ simplifies significantly. To wit, recall from §2.3 that Z is the process that characterizes the arrivals in the limit as $N \to \infty$. The following corollary describes the virtual waiting time in a system with deterministic service times. (We remark that, in steady state, the GI/D/N system in the QED regime was analyzed in Jelenković et al. [22].)

Corollary 3.2 (Deterministic Service). *Let the service time be deterministic with $S = s$. If, as $N \to \infty$,*

$$(\hat{X}_N, \hat{V}_N) \;\Rightarrow\; (\hat{X}, Y),$$

in $D([0, s+\delta), \mathbb{R}^2)$, for some arbitrary $\delta > 0$, then, as $N \to \infty$,

$$\hat{V}_N \;\Rightarrow\; \Big\{ \sup_{0 \le n \le n_t} [Z(t) - Z(t-ns) - n\beta + Y(t-ns)1_{\{n=n_t\}}],\ t \ge 0 \Big\},$$

in $D([0,\infty), \mathbb{R})$, where $n_t := \inf\{n\colon t - ns < s + \delta\}$.

Remark 3.7. The set $\mathrm{Disc}(\hat{V}) \subset [0,\infty)$ when $S \equiv s$ can be characterized as follows. If $x \in \mathrm{Disc}(Y) \subset [0, s+\delta)$ and $x \in [\delta, s+\delta)$ then $\mathbb{P}[x + is \in \mathrm{Disc}(\hat{V})] > 0$ for all finite $i = 1, 2, \ldots$. In addition, if $\mathbb{P}[Y((s+\delta)-) = (\hat{X}(s+\delta) - \beta + Y(\delta))^+] < 1$ then also $\mathbb{P}[is + \delta \in \mathrm{Disc}(\hat{V})] > 0$ for all finite $i = 1, 2, \ldots$. Finally, $\mathbb{P}[x \in \mathrm{Disc}(\hat{V}), \forall x \notin \mathcal{J}] = 0$, where $\mathcal{J} := \{x + is > s + \delta\colon x \in \mathrm{Disc}(Y) \cup \{s+\delta\}, i = 1, 2, \ldots\}$.

Proof. When the service times are deterministic, the tree $\mathscr{F}[t, \delta, \chi]$ reduces to a single path. Thus, function ψ in the statement of Theorem 3.1 reduces to

$$\sup_{0 \le n \le n_t} \left(Y(t-ns)1_{\{n=n_t\}} + \sum_{i=0}^{n \wedge (n_t - 1)} (X(t - is) - \beta) \right).$$

The proof is completed with the observation that $X(t) = Z(t) - Z(t-s)$ (see §2.4). Condition (34), which appears in the statement of Theorem 3.1, is not needed when service times are deterministic because the function $\psi_\delta(\cdot)$ is continuous for relevant weight functions (with probability 1)—see Lemma 3.4 and the proof of Theorem 3.1. □

3.8. Conjectures in steady state. Consider a sequence of GI/D/N queues with constant service times equal to s and renewal arrival processes with interarrival times with the coefficient of variation $v_N \to v$ as $N \to \infty$. In that case, the distribution of $\hat{V}(t)$ tends, as $t \to \infty$, to the distribution of the supremum of a Gaussian random walk with negative drift, i.e.,

$$\hat{V}(t) \;\Rightarrow\; \sup_{n \ge 0} \sum_{i=1}^{n} \zeta_i \quad \text{in } \mathbb{R}, \quad \text{as } t \to \infty,$$

where $\{\zeta_i\}$ is a sequence of i.i.d. normal random variables with mean $-\beta$ and standard deviation σ (see §2.3). It is interesting to note that, in Jelenković et al. [22], it was formally shown that the scaled *steady-state* waiting time tends in distribution to the quantity on the right-hand side of the preceding equation. This opens up the possibility that, for our GI/GI/N model, the following conjecture holds.

Conjecture 3.1. *The scaled limiting* stationary *virtual waiting time process $\tilde{V} = \{\tilde{V}(t),\ t \in \mathbb{R}\}$ can be expressed in terms of a supremum over a weighted tree*:

$$\tilde{V}(t) = \sup_{\mathcal{T} \subseteq \tilde{\mathscr{F}}[t]} (\mathscr{W}_{\mathcal{T}})^+, \qquad t \in \mathbb{R}, \tag{46}$$

where $\tilde{\mathscr{F}}[t]$ is an infinite weighted full K-ary tree. Each node in the tree has exactly K children. The weight of a node associated with a vector $\underline{d}$ is defined to be $p_{\underline{d}}(\tilde{X}(t - s_{\underline{d}}) - \beta)$, where $\{\tilde{X}(t),\ t \in \mathbb{R}\}$ is a stationary *process on $\mathbb{R}$ obtained by extending the arrival process to $(-\infty, 0)$.*

We note that $\{\tilde{X}(t),\ t \in \mathbb{R}\}$ is a stationary zero-mean Gaussian process with a known covariance function (Whitt [34, p. 353])

$$\mathrm{cov}(\tilde{X}(t), \tilde{X}(t+r)) = \mu \int_0^\infty H(u)\bar{H}(u+r)\,du + \mu v^2 \int_0^\infty \bar{H}(u)\bar{H}(u+r)\,du, \tag{47}$$

where $H(u) := \mathbb{P}[S \le u]$ and $\bar{H}(u) := 1 - H(u)$. The expressions for the covariance functions of $\{\hat{X}(t),\ t \ge 0\}$ and $\{\tilde{X}(t),\ t \ge 0\}$, i.e., (11) and (47), differ in the limits of integration. Given the following lemma, it is straightforward to conclude that, under conjecture (46), $\tilde{V}(t)$ satisfies

$$\tilde{V}(t) = (\mathbb{E}_{\langle S\rangle}[\tilde{V}(t-S)] + \tilde{X}(t) - \beta)^+, \quad t \in \mathbb{R}. \tag{48}$$

Lemma 3.5. *Let the process $h = \{h(t),\ t \in \mathbb{R}\}$ be defined by*

$$h(t) = \sup_{\mathcal{T} \subseteq \tilde{\mathcal{T}}[t]} (\mathcal{W}_{\mathcal{T}})^+.$$

Then h satisfies

$$h(t) = (\mathbb{E}_{\langle S\rangle}[h(t-S)] + \tilde{X}(t) - \beta)^+.$$

Remark 3.8. Identifying conditions under which $\tilde{V}$ in (46) is the unique solution of (48) remains an interesting open problem. Equation (48) is related to a class of max-type recursive distributional equations surveyed in [1]. The framework in [1] requires the independence of $\mathbb{E}_{\langle S\rangle}[\tilde{V}(t-S)]$ and $\tilde{X}(t)$; hence, only the case $S = s$ fits the framework (see (47)).

Proof. The weight of the root of $\tilde{\mathcal{T}}[t]$ is equal to $(\tilde{X}(t) - \beta)$ according to the definition. The recursive nature of $\tilde{\mathcal{T}}[t]$ leads to (see also (28))

$$\begin{aligned} h(t) &= (\tilde{X}(t) - \beta)^+ \vee \left\{ \tilde{X}(t) - \beta + \sum_{i=1}^{K} p_i \sup_{\mathcal{T} \subseteq \tilde{\mathcal{T}}[t-s_i]} (\mathcal{W}_{\mathcal{T}})^+ \right\} \\ &= \left(\tilde{X}(t) - \beta + \sum_{i=1}^{K} p_i \sup_{\mathcal{T} \subseteq \tilde{\mathcal{T}}[t-s_i]} (\mathcal{W}_{\mathcal{T}})^+ \right)^+ \\ &= (\tilde{X}(t) - \beta + \mathbb{E}_{\langle S\rangle}[h(t-S)])^+, \end{aligned}$$

where $\vee$ denotes the maximum operator. We conclude the proof with the observation that the quantities $h(t)$ and $h(t - s_i)$ are equal in distribution due to the structure of the node weights, i.e., the stationarity of $\{\tilde{X}(t),\ t \in \mathbb{R}\}$. □

Moreover, it is tempting to conjecture that (48) holds not only for QED systems with service times belonging to a set of finite cardinality, but in fact for a larger class of system. This conjecture is further supported by the following example, in which service time is exponential. Consider a GI/M/N system with exponential service times with mean $1/\mu$ and interarrival times with the coefficient of variation equal to v. Suppose that the limiting virtual waiting time satisfies (48), with $\{\tilde{X}(t),\ t \in \mathbb{R}\}$ being the corresponding limiting scaled GI/M/∞ process. The definition of the conditional expectation $\mathbb{E}_{\langle S\rangle}[\tilde{V}(t-S)]$ yields

$$\mathbb{E}_{\langle S\rangle}[\tilde{V}(t-S)] = \mu \int_{-\infty}^{t} e^{-\mu(t-u)} \tilde{V}(u)\, du. \tag{49}$$

It is well known that in this case $\{\tilde{X}(t),\ t \in \mathbb{R}\}$ is an Ornstein–Uhlenbeck process with infinitesimal drift $m(x) = -\mu x$ and constant infinitesimal variance $\mu(1+v^2)$, e.g., see Whitt [34, p. 354]. Equivalently, the process $\tilde{X}(t)$ satisfies the following stochastic differential equation (Karatzas and Shreve [23, p. 358]):

$$d\tilde{X}(t) = -\mu \tilde{X}(t)dt + \sqrt{\mu(1+v^2)}dB(t), \tag{50}$$

where $\{B(t),\ t \in \mathbb{R}\}$ is the standard Brownian motion. Now, for positive values of $\tilde{V}(t)$, Equations (48) and (49) yield

$$d\tilde{V}(t) = \mu(\tilde{X}(t) - \beta)dt + d\tilde{X}(t),$$

and, thus, by (50) for positive values of $\tilde{V}(t)$

$$d\tilde{V}(t) = -\mu\beta dt + \sqrt{\mu(1+v^2)}dB(t).$$

However, a result in Halfin and Whitt [18] (together with [30]) formally yields that the preceding equation holds, and, thus, (48) is indeed valid for the GI/M/N system.

4. Concluding remarks. In this section, we briefly discuss two possible applications of the approximation

$$\widetilde{V}(t) \approx (\mathbb{E}_{\langle S\rangle}[\widetilde{V}(t-S)] + \widetilde{X}(t) - \beta)^+. \tag{51}$$

First, we mention the possibility of using (51) for simulation-based evaluation of the parameter $\beta > 0$ that results from some required quality of service (e.g., specified by $\mathbb{P}[\widetilde{V}(t) > 0]$ or $\mathbb{E}\widetilde{V}(t)$). Having obtained a desired value of β, the required number of servers (e.g., staffing level in a call center) can be estimated by $R + \beta\sqrt{R}$, where R is the system's offered load. However, simulation of large multiserver queues under high load is computationally intensive. Thus, having fast algorithms that evaluate statistics of the virtual waiting time with reasonable accuracy are of interest. In particular, if $S \in \{\Delta, 2\Delta, \dots, K\Delta\}$ and $p_i = \mathbb{P}[S = i\Delta]$, for some $\Delta > 0$ and $K < \infty$, then simulation via (51) can be more efficient than a direct simulation of a large multiserver queue. In this case, simulation of the steady-state characteristics of $\widetilde{V}(t)$ reduces to iterations of

$$\widetilde{V}(j\Delta) = \left(\sum_{i=1}^{K} p_i \widetilde{V}((j-i)\Delta) + G(j\Delta) - \beta\right)^+$$

over j, where $G(j\Delta)$ is a Gaussian process with covariance structure that is determined by the distribution of S as well as the arrival process. For example, in the case of Poisson arrivals, the covariance function simplifies to (see §2.4)

$$\operatorname{cov}(G(j\Delta), G(j\Delta + k\Delta)) = \lambda\Delta \sum_{i=1}^{K} p_i (i-k)^+.$$

Note that, in this case, the amount of memory required to keep the state of the system during the simulation is $O(K)$ because one needs to keep track of the last K values of the processes $\widetilde{V}(j\Delta)$ and $G(j\Delta)$. This follows from the fact that the covariance function has bounded support when S is bounded (Whitt [34, p. 353]).

The second possible application of (51) is waiting-time prediction/estimation. The prospect of using (51) might be appealing in cases when information on the residual service times of customers in service is not available to the entity that provides newly arrived customers with estimates on their waiting times. Suppose that the value of $\widehat{V}_N(t+\tau)$ must be predicted at time t, and assume that $\tau < s_1$ (in case of $\tau \geq s_1$, the prediction can be obtained by iterating back over time). Observe that, at time $(t+\tau)$, the quantity $\mathbb{E}_{\langle S\rangle}[\widetilde{V}_N(t+\tau-S)]$ is simply a weighted average of past (before t) waiting times and, thus, given the distribution function of S, it is straightforward to evaluate. To make use of (51) one must also estimate the value at $(t+\tau)$ of the infinite-server process $\{\widetilde{X}(t),\ t \in \mathbb{R}\}$, a zero-mean stationary Gaussian process with the covariance function given by (47). However, it is straightforward to estimate $\widetilde{X}(t+\tau)$ based on $\widetilde{X}(t)$ (if the latter is known—otherwise, set $\widetilde{X}(t+\tau)$ to a value of a generated normal random variable $\mathcal{N}(0, \gamma^2)$, where γ^2 is equal to the right-hand side of (47) evaluated for $r = 0$; with Poisson arrivals, $\gamma^2 = \mu\mathbb{E}S$). Finally, the parameter β can be calculated (off-line) either as $(N - R_N)/\sqrt{R_N}$ or as $\sqrt{N}(1-\rho)$, where ρ is the servers' utilization.

5. Proofs. This final section contains the proofs of Lemma 2.4, Lemma 3.1, Corollary 3.1, Lemma 3.2, and Lemma 3.3.

Proof of Lemma 2.4. It is sufficient to demonstrate that, for $\delta \in (0, s_K)$ and all $N \geq N_\delta$, there exists a fixed $\varepsilon > 0$ such that

$$\sup_{t\in[0,\, s_K-\delta)} \frac{X_N(t)}{N} \leq 1 - \varepsilon \quad \text{with probability 1.}$$

To this end, the definition of the infinite-server process renders

$$\begin{aligned}\frac{X_N(t)}{N} &= \sum_{i=1}^{K} \frac{A_{i,N}(t) - A_{i,N}(t-s_i)}{N}\\ &\leq \sum_{i=1}^{K}\left[\left|\frac{A_{i,N}(t)}{p_i\lambda_N} - t\right| + \left|\frac{A_{i,N}(t-s_i)}{p_i\lambda_N} - (t-s_i)^+\right|\right]\frac{p_i\lambda_N}{N} + \sum_{i=1}^{K}\frac{p_i\lambda_N}{N}(t-(t-s_i)^+)\end{aligned}$$

Due to $\lambda_N/N \to \lambda$ as $N \to \infty$ and (3) for every fixed $\varepsilon > 0$ there exists a fixed N_ε such that for all $N \geq N_\varepsilon$ the supremum over $t \in [0, s_K - \delta)$ of the first sum in the preceding inequality is bounded from above by ε with probability 1. On the other hand, the second sum is upper bounded by $1 - \delta p_K \lambda_N/N$. Choosing ε appropriately and setting $N_\delta = N_\varepsilon$ completes the proof. □

Proof of Lemma 3.1. Fix $T > s_K$ and $\varepsilon > 0$ such that (16) holds. On the event

$$\mathscr{E} := \Big\{ \sup_{u \in [T - s_K - 2\varepsilon,\, T]} V_N(u) \le \varepsilon \Big\},$$

by Lemma 2.2, one has that customers with service requirement s_i arriving prior to time $(t - s_i + \varepsilon)$ depart from the system not later than time $(t + 2\varepsilon)$, for all $t \in (T,\, T + s_1 - \varepsilon]$. Hence, on the event $\mathscr{E}$ customers that depart (strictly) after time $(t + 2\varepsilon)$ arrive to the system (strictly) after time $(t - s_i + \varepsilon)$ if their service requirement is equal to s_i. This leads to, on event $\mathscr{E}$,

$$z_t(t + 2\varepsilon) \le \sum_{i=1}^{K} [A_{i,N}(t) - A_{i,N}(t - s_i + \varepsilon)],$$

$\forall\, t \in (T,\, T + s_1 - \varepsilon]$. Recalling from §3.1 that in the QED regime $N = R_N + \beta\sqrt{R_N} + o(\sqrt{R_N})$, the preceding inequality, the FSLLN, and assumption (16) of the lemma ($\mathbb{P}[\mathscr{E}] \to 1$ as $N \to \infty$) yield

$$\mathbb{P}\Big[\sup_{t \in (T,\, T + s_1 - \varepsilon]} z_t(t + 2\varepsilon) < N - \delta \Big] \to 1,$$

for some $\delta > 0$, as $N \to \infty$. Then (14) implies

$$\mathbb{P}[D_N^t < t + 2\varepsilon,\ \forall\, t \in (T, T + s_1 - \varepsilon]] \to 1, \tag{52}$$

as $N \to \infty$. Because $V_N(t) = (D_N^t - t)^+$ by Lemma 2.1 for all t in the interval of interest, the first statement of the lemma follows from (52). Moreover, the waiting time of each customer is a nonnegative quantity, and, thus, $T_N^*(t) + S_N^*(t) \le D_N^t$ (see (15)) resulting in

$$\mathbb{P}[T_N^*(t) + S_N^*(t) < t + 2\varepsilon,\ \forall\, t \in (T,\, T + s_1 - \varepsilon]] \to 1, \tag{53}$$

as $N \to \infty$.

On the other hand, the nonnegativity of the waiting time also renders that a customer with service requirement s_i arriving after time $(t - s_i - \varepsilon)$ departs from the system not earlier than $(t - \varepsilon)$. Equivalently, all customers with service requirement s_i that arrive (strictly) after time $(t - s_i - \varepsilon)$ depart (strictly) after time $(t - \varepsilon)$, i.e.,

$$z_t(t - \varepsilon) \ge \sum_{i=1}^{K} [A_{i,N}(t) - A_{i,N}(t - s_i - \varepsilon)],$$

$\forall\, t \in (T, T + s_1 - \varepsilon]$. The preceding inequality and the FSLLN yield

$$\mathbb{P}\Big[\sup_{t \in (T,\, T + s_1 - \varepsilon]} z_t(t - \varepsilon) > N \Big] \to 1$$

as $N \to \infty$, i.e.,

$$\mathbb{P}[D_N^t > t - \varepsilon,\ \forall\, t \in (T, T + s_1 - \varepsilon]] \to 1, \tag{54}$$

as $N \to \infty$, due to (14). Furthermore, by Lemma 2.2, on the event $\mathscr{E}$ a customer with service time s_i arriving prior to time $(t - s_i - 2\varepsilon)$ departs from the system not later than time $(t - \varepsilon)$. Formally, by considering the customer with arrival time $T_N^*(t)$ and service time $S_N^*(t)$ (see (15)) we get

$$\{T_N^*(t) + S_N^*(t) \le t - 2\varepsilon\} \cap \mathscr{E} \subseteq \{D_N^t \le t - \varepsilon\},$$

for all $t \in (T,\, T + s_1 - \varepsilon]$, or equivalently

$$\{T_N^*(t) + S_N^*(t) > t - 2\varepsilon,\ \forall\, t \in (T,\, T + s_1 - \varepsilon]\} \cup \overline{\mathscr{E}} \supseteq \{D_N^t > t - \varepsilon,\ \forall\, t \in (T,\, T + s_1 - \varepsilon]\},$$

where $\overline{\mathscr{E}}$ is the complement of $\mathscr{E}$. Hence, given (54), the preceding relationship, and $\mathbb{P}[\mathscr{E}] \to 1$ as $N \to \infty$ (assumption (16)), we have

$$\mathbb{P}[T_N^*(t) + S_N^*(t) > t - 2\varepsilon,\ \forall\, t \in (T,\, T + s_1 - \varepsilon]] \to 1, \tag{55}$$

as $N \to \infty$. The limits (53) and (55) yield the second statement of the lemma. □

Proof of Corollary 3.1. In order to avoid ambiguity, let T and ε in the statement of Lemma 3.1 be denoted by T' and ε', respectively. It is sufficient to consider $\varepsilon \le s_1/2$ in the statement of the corollary. Invoking Lemma 3.1 with $T' = s_K + 2\varepsilon$ yields, as $N \to \infty$,

$$\mathbb{P}\Big[\sup_{t\in[0,\, s_K+2\varepsilon+s_1/2]} V_N(t) \le 2\varepsilon\Big] \to 1.$$

Furthermore, after iteratively applying Lemma 3.1 for k times, one has, as $N \to \infty$,

$$\mathbb{P}\Big[\sup_{t\in[0,\, s_K+2\varepsilon+ks_1/2]} V_N(t) \le 2^k\varepsilon\Big] \to 1;$$

on the kth iteration, T' and ε' are set to $s_K + 2\varepsilon + (k-1)s_1/2$ and $2^{k-1}\varepsilon$, respectively (sufficiently small ε guarantee that the lemma is applicable). Therefore, for a fixed T, Lemma 3.1 is applied $k_T = \lceil 2(T - s_K - 2\varepsilon)^+/s_1\rceil$ times and the resulting c is given by 2^{k_T}. □

Proof of Lemma 3.2. Let $\{T^t_{i,j}\}_{j=1}^{A_{i,N}(t)-A_{i,N}(t-s_i-2\varepsilon)}$, $i = 1, \ldots, K$, be the arrival times (sorted in the increasing order) of customers with service times s_i, arriving during the time interval $(t - s_i - 2\varepsilon, t]$. The family of sequences $\{T^t_{i,j}\}$ is defined for all t in the interval $(T,\ T + s_1 - 2\varepsilon]$. Define event

$$\mathscr{E}_* := \{T^*_N(t) + S^*_N(t) \in (t - 2\varepsilon, t + 2\varepsilon),\ \forall\, t \in (T,\ T + s_1 - 2\varepsilon]\},$$

and recall the definitions of $T^*_N(t)$ and $S^*_N(t)$ from (15). It should be noted that, as $N \to \infty$,

$$\mathbb{P}[\mathscr{E}_*] \to 1 \tag{56}$$

due to (21) and the second statement of Lemma 3.1. In the definition of $\mathscr{E}^*$ we set $t \in (T, T + s_1 - 2\varepsilon]$ so that $\widehat{V}^{\uparrow 2\varepsilon}_N(t - S)$ and $\widehat{V}^{\downarrow 2\varepsilon}_N(t - S)$ depend on the values of $\widehat{V}_N(\cdot)$ on the interval $[T - s_K - 2\varepsilon, T]$ only (see the statement of the lemma).

For every sufficiently small $\varepsilon > 0$, on event $\mathscr{E}_*$ we have $T^*_N(t) \in \{T^t_{i,j}\}$ and according to (15)

$$T^*_N(t) + S^*_N(t) + V_N(T^*_N(t)-) = \mathscr{O}_N\{T^t_{i,j} + s_i + V_N(T^t_{i,j}-),\ j \ge 1,\ 1 \le i \le K\}, \tag{57}$$

$\forall\, t \in (T,\ T + s_1 - 2\varepsilon]$, where $\mathscr{O}$ is the sorting operator defined in §2. Note that the Nth element $\mathscr{O}_N$ is well defined, for all N large enough and $t \in (T,\ T + s_1 - 2\varepsilon]$, because the number of arrivals $\sum_{i=1}^K [A_{i,N}(t) - A_{i,N}(t - s_i - 2\varepsilon)]$ is strictly larger than N by the FSLLN (ε is a positive quantity independent of N). Then, on the event $\mathscr{E}_*$, Lemma 2.1 and (17) yield

$$\begin{aligned} V_N(t) &= \mathscr{O}_N\{(T^t_{i,j} + s_i + V_N(T^t_{i,j}-) - t)^+,\ j \ge 1,\ 1 \le i \le K\} \\ &\le (\mathscr{O}_N\{T^t_{i,j} + s_i + V^{\uparrow 2\varepsilon}_N(t - s_i) - t,\ j \ge 1,\ 1 \le i \le K\})^+ \\ &:= (\mathscr{O}^{\uparrow 2\varepsilon}_N(t))^+, \end{aligned} \tag{58}$$

for all $t \in (T,\ T + s_1 - 2\varepsilon]$; the inequality follows from the fact that if $T^*_N(t) = T^t_{i,j}$ for some i and j, then $V_N(T^t_{i,j}-) \le V^{\uparrow 2\varepsilon}_N(t - s_i)$ on the event $\mathscr{E}_*$. Note that the quantity $\mathscr{O}^{\uparrow 2\varepsilon}_N(t)$ defined in (58) depends on the number of arrivals up to time t ($T^t_{i,j} \le t$ by definition), or more specifically one has the following for all $t \in (T,\ T + s_1 - 2\varepsilon]$ and $x \ge 0$:

$$\{\mathscr{O}^{\uparrow 2\varepsilon}_N(t) \ge x\} \subseteq \left\{\sum_{i=1}^K [A_{i,N}(t) - A_{i,N}(t + x - s_i - V^{\uparrow 2\varepsilon}_N(t - s_i))] \ge N - K\right\}. \tag{59}$$

Relationship (59) is due to the fact that (see (58))

$$\sum_{i=1}^K \sum_{j=1}^{A_{i,N}(t)-A_{i,N}(t-s_i-2\varepsilon)} 1_{\{T^t_{i,j}+s_i+V^{\uparrow 2\varepsilon}_N(t-s_i)-t \ge x\}} \ge N \tag{60}$$

only if

$$\sum_{i=1}^K [A_{i,N}(t) - A_{i,N}(t + x - s_i - V^{\uparrow 2\varepsilon}_N(t - s_i))] \ge N - K. \tag{61}$$

In other words, customers with service requirement s_i arriving to the system during the time interval $(t - s_i - 2\varepsilon, t - s_i + 2\varepsilon)$ are delayed (wait for service) at most $V^{\uparrow 2\varepsilon}_N(t - s_i)$ time units on event $\mathscr{E}_*$. Hence, all customers with service requirement s_i arriving prior to time $t+x-s_i-V_N^{\uparrow 2\varepsilon}(t-s_i)$ depart from the system before time $t+x$. The term $(-K)$ on the right-hand side of (61) is due to the nonstrict inequality inside the indicator function in (60) and the right continuity of $\{A_{i,N}(t),\ t\geq 0\}$ for $i=1,2,\ldots,K$. This term is based on the fact that customers arrive one at a time; i.e., at most one customer can arrive at any time instant.

Next, we consider a scaled and centered version of the sum in the preceding equation. Namely, recalling that the number of servers can be written as $N=R_N+\beta_N\sqrt{\rho_N R_N}$ (see (13)), straightforward algebra yields

$$\begin{aligned}&\frac{\sum_{i=1}^K\big[A_{i,N}(t)-A_{i,N}(t+x-s_i-V_N^{\uparrow 2\varepsilon}(t-s_i))\big]-N+K}{\sqrt{N}}\\&\qquad=\tilde{X}_N^{\uparrow 2\varepsilon}(x)-\rho_N\beta_N+\frac{\lambda_N}{\sqrt{N}}(\mathbb{E}_{\langle S\rangle}[V_N^{\uparrow 2\varepsilon}(t-S)]-x)+\frac{K}{\sqrt{N}},\end{aligned}\tag{62}$$

where

$$\begin{aligned}\tilde{X}_N^{\uparrow 2\varepsilon}(x)&:=\sum_{i=1}^K\frac{A_{i,N}(t)-A_{i,N}\big(t+x-s_i-V_N^{\uparrow 2\varepsilon}(t-s_i)\big)-\lambda_N p_i\big(s_i+V_N^{\uparrow 2\varepsilon}(t-s_i)-x\big)}{\sqrt{N}}\\&=\sum_{i=1}^K\big[\hat{A}_{i,N}(t)-\hat{A}_{i,N}(t+x-s_i-V_N^{\uparrow 2\varepsilon}(t-s_i))\big].\end{aligned}\tag{63}$$

Then, combining (59), (62), and (19) results in

$$\begin{aligned}&\{V_N(t)\leq(\mathscr{O}_N^{\uparrow 2\varepsilon}(t))^+,\ \forall\, t\in(T,\ T+s_1-2\varepsilon]\}\\&\quad\subseteq\{\widehat{V}_N(t)\leq(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 2\varepsilon}(t-S)]+\rho_N^{-1}\tilde{X}_N^{\uparrow 2\varepsilon}(V_N(t))-\beta_N+\delta_N)^+,\ \forall\, t\in(T,\ T+s_1-2\varepsilon]\}.\end{aligned}\tag{64}$$

In addition, on the event

$$\mathscr{E}_V:=\Big\{\sup_{t\in(T,\,T+s_1-2\varepsilon]}V_N(t)\leq 2\varepsilon\Big\}\cap\Big\{\sup_{t\in[T-s_K-2\varepsilon,\,T]}V_N(t)\leq\varepsilon\Big\}\tag{65}$$

we have, for $t\in(T,T+s_1-2\varepsilon]$,

$$t-s_i-\varepsilon\leq t+V_N(t)-s_i-V_N^{\uparrow 2\varepsilon}(t-s_i)\leq t-s_i+2\varepsilon.$$

Thus, (63) implies that the term $\tilde{X}_N^{\uparrow 2\varepsilon}(V_N(t))$ in (64) can be bounded for $t\in(T,T+s_1-2\varepsilon]$ on event $\mathscr{E}_V$ as follows:

$$\begin{aligned}\tilde{X}_N^{\uparrow 2\varepsilon}(V_N(t))&\leq\sum_{i=1}^K\hat{X}_{i,N}^{\uparrow 2\varepsilon}(t)\\&=\hat{X}_N^{\uparrow 2\varepsilon}(t),\end{aligned}\tag{66}$$

where $\hat{X}_{i,N}^{\uparrow 2\varepsilon}(t)$ and $\hat{X}_N^{\uparrow 2\varepsilon}(t)$ are defined in §2.4.

Now, observe that $\mathbb{P}[\mathscr{E}_V]\to 1$ as $N\to\infty$ because by (21) the condition of Lemma 3.1 is satisfied. This fact and (56) render

$$\lim_{N\to\infty}\mathbb{P}[\mathscr{E}_*\cap\mathscr{E}_V]=1.\tag{67}$$

Finally, the preceding limit, (58), (64), and (66) yield

$$\begin{aligned}1&=\lim_{N\to\infty}\mathbb{P}[\mathscr{E}_*\cap\mathscr{E}_V]\\&\leq\lim_{N\to\infty}\mathbb{P}[\mathscr{E}_V\cap\{V_N(t)\leq(\mathscr{O}_N^{\uparrow 2\varepsilon}(t))^+,\ \forall\, t\in(T,T+s_1-2\varepsilon]\}]\\&\leq\lim_{N\to\infty}\mathbb{P}[\widehat{V}_N(t)\leq(\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 2\varepsilon}(t-S)]+\rho_N^{-1}\hat{X}_N^{\uparrow 2\varepsilon}(t)-\beta_N+\delta_N)^+,\ \forall\, t\in(T,T+s_1-2\varepsilon]]\leq 1.\end{aligned}$$

This concludes the proof of the statement concerning the upper bound on the virtual waiting time. The lower bound can be obtain by using the same arguments; for completeness, the proof is provided below. The starting point of the lower bound proof is (57). On the event $\mathscr{E}_*$, Lemma 2.1, and (17) yield

$$\begin{aligned}V_N(t)&=\mathscr{O}_N\{(T_{i,j}^t+s_i+V_N(T_{i,j}^t-)-t)^+,\ j\geq 1,\ 1\leq i\leq K\}\\&\geq(\mathscr{O}_N\{T_{i,j}^t+s_i+V_N^{\downarrow 2\varepsilon}(t-s_i)-t,\ j\geq 1,\ 1\leq i\leq K\})^+\\&:=(\mathscr{O}_N^{\downarrow 2\varepsilon}(t))^+,\end{aligned}\tag{68}$$

$\forall\, t \in (T, T+s_1-2\varepsilon]$; the inequality follows from the fact that, if $T_N^*(t) = T_{i,j}^t$ for some i and j, then $V(T_{i,j}^t-) \geq V^{\downarrow 2\varepsilon}(t-s_i)$ on the event $\mathscr{E}_*$. Note that the quantity $\mathscr{O}_N^{\downarrow 2\varepsilon}(t)$, defined in (68), depends on the number of arrivals prior to time t ($T_{i,j}^t \leq t$ by definition), or more specifically one has the following equality for all $t \in (T, T+s_1-2\varepsilon]$ and $x \geq 0$:

$$\{\mathscr{O}_N^{\downarrow 2\varepsilon}(t) \leq x\} \subseteq \left\{\sum_{i=1}^{K}[A_{i,N}(t) - A_{i,N}(t+x-s_i-V_N^{\downarrow 2\varepsilon}(t-s_i))] \leq N\right\}. \tag{69}$$

Equality (69) is due to the fact that (see (68))

$$\sum_{i=1}^{K} \sum_{j=1}^{A_{i,N}(t)-A_{i,N}(t-s_i-2\varepsilon)} 1_{\{T_{i,j}^t+s_i+V_N^{\downarrow 2\varepsilon}(t-s_i)-t>x\}} < N$$

only if

$$\sum_{i=1}^{K}[A_{i,N}(t) - A_{i,N}(t+x-s_i-V_N^{\downarrow 2\varepsilon}(t-s_i))] \leq N.$$

Intuitively, customers with service requirement s_i arriving to the system during the time interval $(t-s_i-2\varepsilon, t-s_i+2\varepsilon)$ wait for service at least $V_N^{\downarrow 2\varepsilon}(t-s_i)$ time units. Thus, all customers with service requirement s_i arriving after time $t+x-s_i-V_N^{\downarrow 2\varepsilon}(t-s_i)$ depart from the system after time $t+x$. Now, consider a scaled and centered version of the sum in the preceding equation. Namely, it is straightforward to conclude that

$$\frac{\sum_{i=1}^{K}[A_{i,N}(t) - A_{i,N}(t+x-s_i-V_N^{\downarrow 2\varepsilon}(t-s_i))] - N}{\sqrt{N}} = \widetilde{X}_N^{\downarrow 2\varepsilon}(x) - \rho_N\beta_N + \frac{\lambda_N}{\sqrt{N}}(\mathbb{E}_{\langle S\rangle}[V_N^{\downarrow 2\varepsilon}(t-S)] - x), \tag{70}$$

where

$$\widetilde{X}_N^{\downarrow 2\varepsilon}(x) := \sum_{i=1}^{K}[\hat{A}_{i,N}(t) - \hat{A}_{i,N}(t+x-s_i-V_N^{\downarrow 2\varepsilon}(t-s_i))]. \tag{71}$$

Then, combining (69), (70), and (19) results in

$$\begin{aligned}&\{V_N(t) \geq (\mathscr{O}_N^{\downarrow 2\varepsilon}(t))^+,\ \forall\, t \in (T, T+s_1-2\varepsilon]\}\\ &\quad\subseteq \{\widehat{V}_N(t) \geq (\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\downarrow 2\varepsilon}(t-S)] + \rho_N^{-1}\widetilde{X}_N^{\downarrow 2\varepsilon}(V_N(t)) - \beta_N)^+,\ \forall\, t \in (T, T+s_1-2\varepsilon]\}.\end{aligned} \tag{72}$$

In addition, on the event $\mathscr{E}_V$ (see (65)) we have, for $t \in (T, T+s_1-2\varepsilon]$,

$$t-s_i-\varepsilon \leq t+V_N(t)-s_i-V_N^{\downarrow 2\varepsilon}(t-s_i) \leq t-s_i+2\varepsilon.$$

Thus, from (71) it follows that on $\mathscr{E}_V$ for $t \in (T, T+s_1-2\varepsilon]$

$$\widetilde{X}_N^{\downarrow 2\varepsilon}(V_N(t)) \geq \widehat{X}_N^{\downarrow 2\varepsilon}(t), \tag{73}$$

where $\widehat{X}_N^{\downarrow 2\varepsilon}(t)$ is defined in §2.4. Finally, (67), (68), (72), and (73) yield

$$\begin{aligned}1 &= \lim_{N\to\infty} \mathbb{P}[\mathscr{E}_* \cap \mathscr{E}_V]\\ &\leq \lim_{N\to\infty} \mathbb{P}[\mathscr{E}_V \cap \{V_N(t) \geq (\mathscr{O}_N^{\downarrow 2\varepsilon}(t))^+,\ \forall\, t \in (T, T+s_1-\varepsilon]\}]\\ &\leq \lim_{N\to\infty} \mathbb{P}[\widehat{V}_N(t) \geq (\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\downarrow 2\varepsilon}(t-S)] + \rho_N^{-1}\widehat{X}_N^{\downarrow 2\varepsilon}(t) - \beta_N)^+,\ \forall\, t \in (T, T+s_1-2\varepsilon]] \leq 1.\end{aligned}$$

This concludes the proof of the lower bound on the virtual waiting time. □

Proof of Lemma 3.3. The proofs of the two limits are almost identical, and, therefore, we provide a detailed proof only for the first limit (upper bound). Note that

$$\left\{\sup_{t\in[T-s_K-3\varepsilon,\, T]} V_N(t) \leq \varepsilon\right\} = \left\{\sup_{t\in[T-s_K-3\varepsilon,\, T-\varepsilon]} V_N(t) \leq \varepsilon\right\} \cap \left\{\sup_{t\in[T-s_K-2\varepsilon,\, T]} V_N(t) \leq \varepsilon\right\},$$

and, hence, applying Lemma 3.2 (twice) results in

$$\mathbb{P}[\widehat{V}_N(t) \leq (\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 2\varepsilon}(t-S)] + \rho_N^{-1}\widehat{X}_N^{\uparrow 2\varepsilon}(t) - \beta_N + \delta_N)^+,\ \forall\, t \in (T-\varepsilon, T+s_1-2\varepsilon]] \to 1, \tag{74}$$

as $N \to \infty$.

On the other hand, the fact that the supremum of a sum is not less than the sum of suprema yields

$$\begin{aligned}&\{\widehat{V}_N(t) \leq (\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 2\varepsilon}(t-S)] + \rho_N^{-1}\widehat{X}_N^{\uparrow 2\varepsilon}(t) - \beta_N + \delta_N)^+,\ \forall\, t \in (T-\varepsilon, T+s_1-2\varepsilon]\}\\ &\quad\subseteq \{\widehat{V}_N^{\uparrow \varepsilon}(t) \leq (\mathbb{E}_{\langle S\rangle}[\widehat{V}_N^{\uparrow 3\varepsilon}(t-S)] + \rho_N^{-1}\widehat{X}_N^{\uparrow 3\varepsilon}(t) - \beta_N + \delta_N)^+,\ \forall\, t \in (T, T+s_1-3\varepsilon]\}.\end{aligned} \tag{75}$$

The statement follows from (74) and (75). □

Acknowledgments. The authors thank David Gamarnik for helpful discussions and Gennady Samorodnitsky for a useful comment on the main result. They also thank the reviewers for their detailed and constructive comments, which challenged the authors to produce the present version, which is a substantial improvement over the originally submitted manuscript. The research of A. Mandelbaum was partially supported by Binational Science Foundation Grant 2001685/2005175, by Israeli Science Foundation Grants 126/02 and 1046/04, and by Technion funds for the promotion of research and sponsored research.

References

[1] Aldous, D., A. Bandyopadhyay. 2005. A survey of max-type recursive distributional equations. *Ann. Appl. Probab.* **15**(2) 1047–1110.

[2] Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

[3] Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, sequencing rules, and system design. *Oper. Res.* **52**(2) 271–292.

[4] Asmussen, S. 2003. *Applied Probability and Queues*, 2nd ed. Springer-Verlag, New York.

[5] Atar, R. 2005. A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Probab.* **15**(1B) 820–852.

[6] Atar, R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15**(4) 2606–2650.

[7] Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **14**(3) 1084–1134.

[8] Baccelli, F., P. Bremaud. 2003. *Elements of Queueing Theory*, 2nd ed. Springer-Verlag, Berlin.

[9] Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning of large call centers. *Oper. Res.* **52**(1) 17–34.

[10] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, S. Zeltyn, L. Zhao, S. Heipeng. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

[11] Erlang, A. K. 1948. On the rational determination of the number of circuits. E. Brockmeyer, H. L. Halstrom, A. Jensen, eds. *The Life and Works of A. K. Erlang*. The Copenhagen Telephone Company, Copenhagen.

[12] Fleming, P., A. Stolyar, B. Simon. 1994. Heavy traffic limit for a mobile phone system loss model. *Proc. 2nd Int'l Conf. Telecomm. Syst. Mod. Anal.*, Nashville, TN.

[13] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

[14] Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.

[15] Glynn, P., W. Whitt. 1991. A new view of the heavy-traffic limit theorem for the infinite-server queue. *Adv. Appl. Probab.* **23** 188–209.

[16] Gromoll, H. C., A. L. Puha, R. J. Williams. 2002. The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.* **12**(3) 797–859.

[17] Gurvich, I., M. Armory, A. Mandelbaum. 2008. Service level differentiation in call centers with fully flexible servers. *Management Sci.* **54**(2) 279–294.

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

[19] Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* **52**(2) 243–257.

[20] Jacod, J., A. Shiryaev. 2003. Limit theorems for stochastic processes, 2nd. ed. Springer-Verlag, Berlin.

[21] Jagerman, D. 1974. Some properties of the Erlang loss function. *Bell System Techn. J.* **53**(3) 525–551.

[22] Jelenković, P., A. Mandelbaum, P. Momčilović. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Syst. Theory Appl.* **47**(1–2) 53–69.

[23] Karatzas, I., S. Shreve. 1991. *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer-Verlag, New York.

[24] Kiefer, J., J. Wolfowitz. 1955. On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78** 1–18.

[25] Krichagina, E., A. Puhalskii. 1997. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing System Theory Appl.* **25** (1–4) 235–280.

[26] Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

[27] Mandelbaum, A., W. Massey. 1995. Strong approximations for time-dependent queues. *Math. Oper. Res.* **20**(1) 33–64.

[28] Mandelbaum, A., R. Schwartz. 2002. Simulation experiments with $M/G/100$ queues in the Halfin-Whitt (Q.E.D.) regime. Technical report, Technion, http://iew3.technion.ac.il/serveng/References/references.html.

[29] Mandelbaum, A., S. Zeltyn. 2006. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. preprint.

[30] Puhalskii, A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* **19**(4) 946–954.

[31] Puhalskii, A., M. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32**(3) 564–595.

[32] Tezcan, T. 2008. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math. Oper. Res.* **33**(1) 51–90.

[33] Whitt, W. 1980. Some useful functions for functional limit theorems. *Math Oper. Res.* **5**(1) 67–85.

[34] Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

[35] Whitt, W. 2004. A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* **52**(6) 922–941.

[36] Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.

[37] Whitt, W. 2005. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* **30**(1) 1–27.

[38] Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Syst. Theory Appl.* **51**(3–4) 361–402.