A diffusion regime with non-degenerate slowdown

Rami Atar*

June 18, 2009; Revised June 27, 2010 and June 15, 2011

Abstract

We study a diffusion regime, earlier considered by Gurvich, Mandelbaum, Shaikhet and Whitt in the case of the M/M/N queue, that is, in a sense that we make precise, a midpoint between two well-known heavy traffic diffusion regimes, the *conventional* and the *quality and efficiency driven* regimes. Unlike the other two, this regime, that we call the *non-degenerate slowdown* regime, enjoys the property that delay and service time are of the same order of magnitude, a property that is often desirable from a modeling viewpoint. Our main result is that in the case of heterogeneous exponential multi-server systems, this regime gives rise to new limit processes for the sojourn time. In particular, the joint limit law of the delay and service time processes is identified as a reflected Brownian motion and an independent process, whose marginal is a size-biased mixture of exponentials. Our results also motivate the formulation and study of new diffusion control problems, based on sojourn time cost.

AMS subject classifications: 60K25, 60J60, 60F17, 90B22, 68M20

Keywords: diffusion limits, many-server queue, heavy traffic, conventional diffusion regime, ED and QED regimes, non-degenerate slowdown regime

1 Introduction

Diffusion asymptotic analysis of queueing systems in heavy traffic has been the subject of extensive research within stochastic network theory, motivated by the convenient approximations which diffusion models offer. This paper focuses on a many-server diffusion regime where the number of servers grows without bound, that was earlier considered by Mandelbaum and Shaikhet [15], [17], Whitt [22], and Gurvich [10], and is unique in that delay and service time remain comparable under scaling. While the model analyzed by the above mentioned authors was the M/M/N queue, with $N \to \infty$, our main focus is on heterogenous many-server systems, where this asymptotic regime gives rise to new limit processes for the sojourn time, and provides motivation to study new diffusion control problems.

*Research supported in part by the Israel Science Foundation (Grant 1349/08), the US–Israel Binational Science Foundation (Grant 2008466), and the Technion's fund for promotion of research

1.1 On heavy traffic regimes

To place the framework of this paper among other heavy traffic regimes, it is instructive to recall the classification of Garnett, Mandelbaum and Reiman [9]. Consider the M/M/N+M model with parameters λ, μ and θ, of a Markovian N-server queue, where customers may abandon while waiting to be served. The three parameters represent arrival rate, per-server service rate, and per-customer abandonment rate, respectively. The ratio $R = \lambda/\mu$ is often referred to as the *offered load*. As measures of performance consider the steady state probability that an incoming customer does not find an available server immediately upon arrival, and the steady state probability that a customer abandons. Denote these quantities by P_W and P_A, respectively (where W and A are mnemonic for 'wait' and 'abandonment'). This model is analyzed in [9] in three many-server asymptotic regimes, which differ in how N and R are related, and the following results are shown.
(i) When, for some fixed $\beta \in (-\infty, \infty)$, $N = R + \beta\sqrt{R}$, one has $P_W \to \alpha(\beta) \in (0,1)$ and $P_A \to 0$, as $R \to \infty$ (and consequently $N \to \infty$);
(ii) When $N = R + \varepsilon R$, some fixed $\varepsilon > 0$, one has $P_W \to 0$ and $P_A \to 0$, as $R \to \infty$;
(iii) When $N = R - \varepsilon R$, some fixed $\varepsilon \in (0,1)$, one has $P_W \to 1$ while $P_A \to \varepsilon$, as $R \to \infty$.

Result (i) is an extension, to a model that accommodates abandonment, of the well-known *square root rule for safety staffing* (see Whitt [20] as well as references therein to earlier treatments), which proposes how the staffing level, N, should be determined from R so as to achieve a desired level of quality of service. When stated for the G/M/N queue (with no abandonment), this rule asserts that if R is large and N is selected according to the formula

$$N = R + \beta\sqrt{R},$$

then the quality of service, as measured by P_W, is dictated by the parameter β (required in this case to be positive). Whitt [20] rigorously justifies this rule based on Halfin and Whitt [11], where it is shown via diffusion limit techniques, that the large R limit of P_W exists as a number in the interval $(0,1)$, and is solely determined by β and the inter-arrival squared coefficient of variation (assumed finite). Result (i) of [9], alluded to above, thus asserts that an analogous rule continues to hold for the model with abandonment. Note that it also addresses the other measure of performance, P_A.

Results (ii) and (iii) are concerned with a system operating at a high level of quality of service, and, respectively, high efficiency. Accordingly, [9] propose to classify asymptotic regimes for multi-server systems with abandonment according to whether $(N-R)/\sqrt{R}$ converges to

$$+\infty, \qquad -\infty, \qquad \text{or some } \beta \in (-\infty, +\infty),$$

(corresponding to results (ii), (iii), and, respectively, (i)), and to refer to these three regimes, respectively, as

Quality-Driven (QD), *Efficiency-Driven (ED),* and *Quality- and Efficiency-Driven (QED).*

Similarly, for models without abandonment (say, M/M/N), the three limiting values dictating the regimes are, respectively,

$$+\infty, \qquad 0, \qquad \text{or some } \beta \in (0, +\infty),$$

where one assumes $N > R$ so that steady state is well-defined. Because result (i) was first established by [11] (in absence of abandonment), the QED regime is also often called the *Halfin-Whitt* regime.

There is a vast literature on heavy traffic diffusion limits, and the models which have been studied are many, including very general stochastic network systems. The majority of these works analyze models containing a *fixed* number of servers [12, 13, 14, 21]. It is standard to refer to such a setting as the *conventional* diffusion regime (as e.g., in [4]). In recent years there has also been much interest in many-server limits of critically loaded systems, and the model suggested by Halfin and Whitt has been extended in various ways. This body of work, inspired by the approach of [11], addresses systems in the QED regime. On the other hand, the conventional heavy traffic is an ED regime. To see this, consider the M/M/N queue. Fix μ and N, and let $\lambda = \lambda_n = N\mu - cn^{-1/2}$, some $c > 0$. Since $(N-R)/\sqrt{R} = (N - \lambda_n/\mu)(\lambda_n/\mu)^{-1/2} \to 0$, these assumption correspond to the ED regime. As in result (iii) alluded to above, it is well-known that $P_W \to 1$. A well-known limit result states that, with Q_n denoting the corresponding queue-length process, the rescaled process $\hat{Q}_n(t) = n^{-1/2}Q_n(nt)$ converges to a reflected Brownian motion (RBM) on $\mathbb{R}_+$ with specified parameters (assuming convergence of initial conditions). What makes the conventional regime very useful is that these limit results are much more general, and continue to hold, e.g., for the G/G/N queue (under moment assumptions).

However, the conventional diffusion regime is not the only ED regime. Mandelbaum and Shaikhet [15], [17], and Whitt [22] (Theorem 2.2) have identified the following scaling that is ED, with $N \to \infty$. Consider an $M/M/N$ queue with $\mu = 1$ and

$$\lambda = \lambda_N = N - c + o(1), \tag{1}$$

where $c > 0$ is constant. Note that $(N-R)/\sqrt{R} \to 0$ as $R \to \infty$. Denote by Q_N the queueing process, and by X_N the number-in-system process. These authors show that

$$N^{-1}Q_N(Nt) \tag{2}$$

converges to a RBM, where [22] also characterizes the limiting delay process and shows that it is a RBM. Gurvich [10] (Proposition 5.1.1) extended these results to show, under a natural condition on the load, that

$$N^{-\delta}(X_N(N^{2\delta-1}t) - N) \tag{3}$$

converges to a RBM for any $\delta \in (\frac{1}{2}, 1]$.

1.2 ED regimes and slowdown

In both the conventional and the QED regimes, the expected *delay* and *service time* experienced by a customer scale differently as a function of the scaling parameter, n. In fact, delay turns infinitely larger (smaller) than service time in the conventional (respectively, QED) regime. From a modeling viewpoint, it is desirable to allow for these two important performance measures to be comparable under the scaling. As observed by [15], [17], [22], their scaling alluded to above is unique, in that the delay and the service time remain comparable as $n \to \infty$.

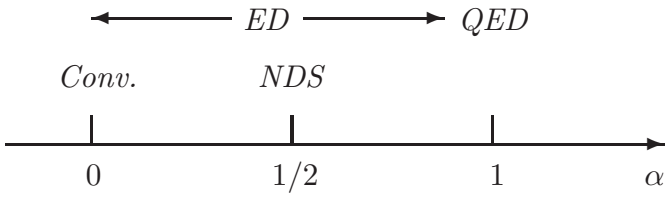


Figure 1: *Diffusion regimes on the α-axis: The interval $[0,1)$ corresponds to the ED regimes, with the conventional ($\alpha = 0$) and the NDS ($\alpha = 1/2$) as special cases. The point 1 corresponds to the QED regime.*

To understand this, consider the M/M/N model where the parameters λ, μ and N all depend on some scaling parameter n. In particular, given $\alpha \in [0,1]$, assume that

$$\lambda_n = n - \hat{\lambda} n^{1/2}, \qquad N_n = [\nu n^{\alpha}], \qquad \mu_n = \nu^{-1} n^{1-\alpha},$$

where ν and $\hat{\lambda}$ are constants (a more general formulation appears below in Section 2). In the case $\alpha = 0$, ν is assumed to be an integer. Note that the system is critically loaded in the sense $\lambda_n \approx N_n \mu_n$. Then the cases $\alpha = 0$ and $\alpha = 1$ are seen to correspond to the conventional, and QED regimes, respectively, and this parametrization can be thought of as an interpolation between the two (see Figure 1). Moreover, any $\alpha \in [0,1)$ represents an ED regime, because

$$\frac{N_n - R_n}{\sqrt{R_n}} = O(n^{-\frac{1}{2}+\frac{\alpha}{2}}) \to 0.$$

By a simple calculation it is seen that the case $\alpha = 1/2$ corresponds to the scaling (1), (2) above, while the interval $\alpha \in [\frac{1}{2}, 1)$ corresponds to (3) with $\delta \in (1/2, 1]$. A typical diffusion limit result in either the conventional or the QED regime shows that queue-length scales as $O(n^{1/2})$. Since the overall processing rate at each station scales as n in both regimes, this means that the expected delay scales as $n^{-1/2}$. On the other hand, the mean service time is the reciprocal service rate, behaving like $O(n^{-1})$ and $O(1)$ in the conventional and the QED regimes, respectively. See columns I and III of Table 1. For a general $\alpha \in [0,1]$, the delay scales as $n^{-1/2}$ just as in both regimes, while service time scales as $n^{-1+\alpha}$. Hence the case $\alpha = 1/2$ is special, in that delay and service time have the same rate of decay in n.

One often defines the *slowdown* as the ratio between the sojourn time and the service time experienced by a typical customer. The foregoing discussion shows that the slowdown degenerates in the limit as $n \to \infty$, to ∞ or to 1, for any α in the interval $[0,1]$ except $\alpha = 1/2$. We therefore refer to the case where $\alpha = 1/2$, considered by [15], [17], [22] and [10], as the *Non-Degenerate Slowdown* (NDS) diffusion regime (column II of Table 1 and point 1/2 in Figure 1).

While the discussion above is on *critically loaded, diffusion* regimes, there are, of course, other useful asymptotic approaches. Whitt [23] proposes an *overloaded, fluid* many-server ED regime, where individual abandonment time and service time are held fixed, while arrival rates and number of servers increase in such a way that the traffic intensity exceeds one. A unique feature of this regime is that the probability of eventually being served approaches a limit strictly between zero and one. Just like the NDS diffusion regime, the ED regime of [23] has the property that delay and service time are comparable.

Diffusion regime	I. Conv.	II. NDS	III. QED
Arrival rate	$O(n)$	$O(n)$	$O(n)$
Number of servers	$O(1)$	$O(n^{1/2})$	$O(n)$
Individual service rates	$O(n)$	$O(n^{1/2})$	$O(1)$
$\Delta :=$ Delay	$O(n^{-1/2})$	$O(n^{-1/2})$	$O(n^{-1/2})$
$\Sigma :=$ Service time	$O(n^{-1})$	$O(n^{-1/2})$	$O(1)$
Relation Δ and Σ	$\Delta \gg \Sigma$	$\Delta \sim \Sigma$	$\Delta \ll \Sigma$

Table 1: *Comparison of three heavy traffic diffusion regimes. Under the conventional and the QED regimes, delay and service time experienced by a typical customer scale differently in n, while they are comparable under the non-degenerate slowdown regime.*

1.3 Non-degenerate slowdown in applications

The NDS diffusion regime is meaningful and useful in applications where

(a) The traffic intensity is close to one,

(b) Delays and service times are of the same order,

(c) The fraction of abandoning customers is of the order of $1/N$ or lower, where N is the number of servers.

Items (a) and (b) merely express the critical load condition and non-degeneracy of the slowdown. As will be argued in Remark 2.2(a) below, the abandonment fraction is of order $n^{-1/2}$ in this regime. Our formulation of item (c) is based on this and the physical meaning of n, as the squared number of servers (another important meaning is the squared ratio between arrival rate and individual service rate).

The main motivation of [9], as well as many other works on many-server asymptotics, arises from the analysis and design of call centers. Whitt [23] argues that his approach, alluded to above, gives rise to efficient approximations for call center applications, provided that abandonment is significant and the quality of service is somewhat low, or when queue length and waiting times are relatively large. Although the reason for comparability in this case is different than in the present paper, and stems from the strong effect of abandonment, the applicability of such ED approximations to call center analysis does indicate that non-degenerate slowdown indeed occurs in practice.

Following is an example of a system that meets conditions (a)–(c) above, and which is therefore meaningful to approach by NDS diffusion approximations. A detailed empirical study of a particular banking call center operations is reported in Brown et al. [7], based on data of more than 1,200,000 calls over the period of a year. This call center is seen to operate with traffic intensity between 0.86–1.22 (as can be seen from [8, Table 6, p. 48] where λ is within 103–116 calls/hour; and $N\mu$ is within 1.7–2 min). Moreover, it operates with comparable delay and service time. Indeed, the

overall average service time reported in this study is 201 sec (with standard deviation of 248 sec) [7, Table 1., p. 39], [8, Table 2, p. 15], while the overall average delay is 98 sec (with standard deviation of 105 sec) [8, Table 3, p. 24], giving a ratio of 2.05, approximately. A more detailed comparison arises when one examines the four types of service that are offered. The types, and the ratio alluded to above for each type, are as follows. Regular service 179 sec/96 sec ≈ 1.86; service to new customers 115 sec/136 sec ≈ 0.85; service related to stock trading 270 sec/114 sec ≈ 2.37; and internet assistance 401 sec/159 sec ≈ 2.52 (this data corresponds to callers who waited for service, i.e., excluding abandoning customers). In summary, the ratio between average service time and delay, under different circumstances, varies between 0.85 and 2.52, and it is very reasonable to consider them comparable.

Further, the fraction of abandoning customers in this empirical study meets condition (c) above. Indeed, it varies from 10% to 24% [8, Table 6, p. 48]. The number of active agents is $N = 6$, as documented in the same table, and $1/N \approx 17\%$ fits well with the range stated above.

These figures clearly indicate a good fit with the working assumptions (a)–(c) of the NDS diffusion regime, and it is therefore expected that the NDS diffusion regime provides, in this case, more efficient approximations than, for example, QED with abandonment (a framework that has often been studied in relation to call centers), or any other critically loaded diffusion regime. The alternative approach of [23] is also useful here when the traffic intensity exceeds 1 (recall that it varies between 0.86–1.22). It is interesting to ask, given a real-world system, which asymptotic framework fits better than the other, and in particular whether the NDS diffusion regime or the ED approach of [23] fits better than the other in this particular case. We will not address these questions here, but note that a serious treatment of this issue requires tools from statistics (such as those implemented in [7] for related questions of model fitting).

In data obtained from one of the authors of [7]–[8] by private communication [16], the behavior of delay over one-hour periods was examined. It was found that its mean and standard deviation over each period were consistently of the same order of magnitude, including periods where the number of callers was relatively high (30–150 calls). This shows that local fluctuations over time are considerable, just as one expects in a diffusion regime. This provides an additional indication that critically loaded *diffusion approximations* are effective (arguably, more than fluid approximations).

We would like to point out that although the NDS regime is obtained by letting the service rates (and arrival rates) grow without bound, it should not be regarded unsuitable for applications where service is, in some sense, "slow" (such as systems operated by humans). In fact, speeding up the rates is merely a convenient way of applying an acceleration of the processes involved, and this is how it should be thought of (this is similar to the situation in the conventional regime, which can be defined with $O(1)$ service and arrival rates and time acceleration, or alternatively, with $O(n)$ rates).

1.4 Theoretical significance of the NDS diffusion regime

The NDS diffusion regime is an ED regime (as is any point on the α-axis, $\alpha \in [0, 1)$), and it resembles the conventional regime in that the limiting delay probability is one and the limiting rescaled queue-

length is a RBM. The main point of this paper is to argue that under server heterogeneity it is unique among all other ED or QED diffusion regimes, in two ways. First, as shown by our main result, *a new limit process arises for the sojourn time.* Although in the case of an $M/M/N$ queue the limiting service time distribution is a plain exponential (as it is in the prelimit), our results establish a more interesting behavior in the heterogenous case. One expects that more general network settings and service time distributions will involve further issues regarding the joint law. A second aspect making this regime unique is related to *control formulations.* There are many papers on diffusion control problems associated with heavy traffic scaling limits of queueing networks. As a consequence of the fact that sojourn time is distinct from delay and service time in this regime, this is the only diffusion regime where it is natural to formulate dynamic control problems with sojourn time costs as opposed to ones based on delay or service time costs. This point is elaborated in Section 3.

Our main result (Theorem 2.2) is concerned with a many server queue with server heterogeneity and customer abandonment, and considers a general class of work-conserving first-come-first-served policies. The result identifies the limiting *joint law* of delay and service time under the NDS regime, in the sense of finite dimensional distribution convergence, in the form of a reflected Ornstein-Uhlenbeck (ROU) process (or, in the case without abandonment, a RBM), and an independent 'white noise' process whose marginal is a size-biased mixture of exponentials. These limit processes do not depend upon the routing policy. The most important consequence is a description of the limiting sojourn time process as the sum of the two processes mentioned above.

On the way to proving the main result, we obtain, in Theorem 2.1, convergence of the diffusion-scaled queue-length process to a reflected diffusion, in a setting which includes as special case the α-parameterized model for any $\alpha \in [0, 1)$.

A standard way of modeling server heterogeneity with $N \to \infty$ is to consider a fixed, finite number of server pools each containing servers with identical characteristics, and let the number at each pool grow without bound. An example of a paper that uses such a setting is Armony [2]. We take a more general approach where servers need not be divided into pools. Instead, assumptions are imposed on the total rate ((6), (7) and (8)), the minimal rate (9), and the empirical measure of the (suitably scaled) rates ((26) and (27)). This is a reminiscent of the setting of Atar and Shwartz [6], where, however, the asymptotic regime is QED.

Let us finally mention additional works on diffusion limits in ED regimes for models with abandonment. Ward and Glynn [18] study a model in the conventional regime and obtain ROU as the limiting process. Whitt [23, Section 4] obtains a diffusion limit, in the form of an Ornstein-Uhlenbeck process, in a scaling very similar to the NDS diffusion regime, but where traffic intensity is kept fixed above one rather than equal to one.

Organization of the paper. The rest of this paper is organized as follows. The model and main results are presented in Section 2, where Theorem 2.1 identifies the limiting queue-length process in a relatively general many-server regime, and Theorem 2.2 treats the delay and service time under the NDS regime. Concluding remarks appear in Section 3. The Appendix [3], published online, contains the proofs, where Theorem 2.1 is proved in Subsection A.1 and Theorem 2.2 in Subsection A.2.

2 Model and results

Let a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ be given, supporting all random variables and stochastic processes defined below. Expectation w.r.t. $\mathbb{P}$ is denoted by $\mathbb{E}$. We will use the following notation. The symbol $\Rightarrow$ denotes convergence in distribution. We write $\mathbb{R}_+^*$ for $[0,\infty) \cup \{\infty\}$. By saying that G^n converge in distribution to G as $n \to \infty$ (and writing $G^n \Rightarrow G$), where G^n, $n \geq 1$, and G are $(\mathbb{R}_+^*)^d$-valued random variables, we mean that $\mathbb{E}[f(G^n); \max_{1\leq i\leq d} G_i^n < \infty] \to \mathbb{E}[f(G); \max_{1\leq i\leq d} G_i < \infty]$ for every $f \in C_b(\mathbb{R}_+ : \mathbb{R})$. We write $\Delta M(t) = M(t) - M(t-)$, $t > 0$, for any càdlàg function M, and $x^{\pm} = \max(\pm x, 0)$ for $x \in \mathbb{R}$. Finally, $|f|_t^* = \sup_{[0,t]} |f|$.

2.1 Model and convergence of queue-length process

The model consists of a parallel server system with a single queue and multiple servers. It is parameterized by $n \in \mathbb{N}$. N_n denotes the number of servers in the nth system, and $K_n := \{1, 2, \ldots, N_n\}$ is an index set for servers. It is assumed that $N_n \to \infty$ as $n \to \infty$, and that $\inf_n N_n \geq 1$. The arrivals are modeled as renewal processes. To this end, we are given parameters $\lambda_n > 0$, $n \in \mathbb{N}$ and a sequence of positive i.i.d. random variables $\{IA(l), l \in \mathbb{N}\}$ (the letters IA are mnemonics for 'inter-arrival'), with mean $\mathbb{E}[IA(1)] = 1$ and variance $C_{IA}^2 = \mathrm{Var}(IA(1)) \in [0,\infty)$. With $\sum_1^0 = 0$, the number of arrivals up to time t for the nth system is given by

$$A_n(t) = \sup\Big\{l \geq 0 : \sum_{i=1}^{l} \lambda_n^{-1} IA(i) \leq t\Big\}, \quad t \geq 0.$$

Denote the time of the first arrival after time t by

$$AT_n(t) = \inf\{s > t : \Delta A_n(s) > 0\}, \qquad t \geq 0. \tag{4}$$

The parameters are assumed to satisfy $\lim_n \lambda_n/n = \lambda > 0$, and

$$\widehat{\lambda}_n := n^{-1/2}(\lambda_n - n\lambda) \to \widehat{\lambda} \in (-\infty, \infty), \tag{5}$$

as $n \to \infty$. Next, let $\mu_{kn} > 0$, $k \in K_n$ be given constants, representing the service rate of server k at the nth system. The sum

$$\mu_n = \sum_{k\in K_n} \mu_{kn} \tag{6}$$

is assumed to satisfy

$$\bar{\mu}_n := n^{-1}\mu_n \to \mu > 0, \quad \text{as } n \to \infty, \tag{7}$$

and

$$\widehat{\mu}_n := n^{-1/2}(\mu_n - n\mu) \to \widehat{\mu} \in (-\infty, \infty), \quad \text{as } n \to \infty. \tag{8}$$

It is also assumed that

$$\mu_n^{\min} := \min_{k\in K_n} \mu_{kn} \to \infty, \quad \text{as } n \to \infty. \tag{9}$$

The system is assumed to be critically loaded in the sense

$$\lambda = \mu. \tag{10}$$

We let X_n, Q_n and I_n be processes representing the number of customers in the system, the number of customers in the buffer, and, respectively, the number of servers that are idle. It is assumed that the routing policy is work conserving, in the sense that

$$Q_n(t) = (X_n(t) - N_n)^+, \quad I_n(t) = (X_n(t) - N_n)^-, \quad t \geq 0. \tag{11}$$

For $k \in K_n$, let B_{kn} be a stochastic process taking values in $\{0, 1\}$, representing the status of server k, as follows: $B_{kn}(t) = 1$ if and only if server k is busy at time t. We set $I_{kn} = 1 - B_{kn}$. Note that $I_n = \sum_{k \in K_n} I_{kn}$. (The letters B and I are mnemonics for busy and idle, resp.). To model service time according to the exponential distribution, assume we are given i.i.d. standard (rate 1) Poisson processes S_k, $k \in \mathbb{N}$, independent of the arrival process. The number of service completions by server k, during the time interval $[0, t]$ is assumed to be given by

$$D_{kn}(t) = S_k(T_{kn}(t)), \qquad k \in K_n, t \geq 0, \tag{12}$$

where

$$T_{kn}(t) = \mu_{kn} \int_0^t B_{kn}(s)ds, \quad k \in K_n. \tag{13}$$

The total number of service completions till time t is given by

$$D_n(t) = \sum_{k \in K_n} D_{kn}(t). \tag{14}$$

We also include customer abandonment in the model. The abandonment rate per unit time, per customer waiting in the queue, is given by the constant $\gamma_n \geq 0$, assumed to satisfy

$$\gamma_n \to \gamma \in [0, \infty). \tag{15}$$

Letting Z be a standard Poisson process, the number of customers abandoning while waiting to be served, by time t, will be given by

$$Z_n(t) := Z(\widetilde{T}_n(t)), \tag{16}$$

where

$$\widetilde{T}_n(t) = \gamma_n \int_0^t Q_n(s)ds. \tag{17}$$

Notice that it is a legitimate special case to let $\gamma_n = 0$ for all n, removing abandonment from the model. The following equation follows from the foregoing verbal description

$$X_n(t) = X_n(0) + A_n(t) - D_n(t) - Z_n(t). \tag{18}$$

The processes A_n, S_k, Z, X_n, Q_n, I_n, B_{kn} are all assumed to have càdlàg sample paths. The primitive processes A, $\{S_k\}$, Z, and the initial condition $(\{B_{kn}(0), k = 1, \ldots, N_n\}, Q_n(0))$ are assumed to be mutually independent (for each n).

It will be assumed throughout that every server can only serve one customer at a time, hence that processor sharing disciplines are not allowed. Apart from the assumptions on the policy, regarding work conservation and ruling out processor sharing, we must require that the routing mechanism does not use information from the future. To this end we impose the following assumption throughout.

Assumption 2.1. *For each n there exists a filtration $\mathbb{F}_n = \{\mathcal{F}_n(t), t \geq 0\}$ that is right-continuous and $\mathbb{P}$-complete, such that the following holds:*

i. The processes A_n, X_n, Q_n, I_n, B_{kn}, T_{kn}, D_{kn}, Z_n are adapted to the filtration;

ii. For each $k \in K_n$,

$$D_{kn} - T_{kn} \text{ is a martingale with respect to } \mathbb{F}_n; \tag{19}$$

iii. Given any a.s.-finite $\mathbb{F}_n$-stopping time τ, the conditional joint law of the N_n processes

$$\{S_k(T_{kn}(\tau) + s) - S_k(T_{kn}(\tau)), s \geq 0, k \in K_n\},$$

conditioned on $\mathcal{F}_n(\tau)$, is that of N_n i.i.d. standard Poisson processes.

iv. For any $t \geq 0$ and any event $E_n \in \mathcal{F}_n(t)$, the N_n-dimensional process

$$\{S_k(T_{kn}(t) + s) - S_k(T_{kn}(t)), s \geq 0, k \in K_n\},$$

the process

$$\{A_n(AT_n(t) + s) - A_n(AT_n(t)), s \geq 0\},$$

and the event E_n are mutually independent.

The combination of items (i)–(iv) above asserts that the distribution of future service times and inter-arrival times is independent of events from the past. This assumption will be violated by (unrealistic) routing policies that can access information from the future, and make decisions based on this. It does hold for any reasonable routing policy that does not have access to information from the future at the time of routing. Consider for example the policy to *always route to the fastest server among those that are available*, and the policy to *always route to the slowest one*. These are special cases of *feedback policies*: each routing decision, say at time t, is performed by selecting a server k according to a given mapping ϕ, say by $k = \phi(B_{kn}(t), D_{kn}(t), k \in K_n)$. These all meet Assumption 2.1. More general policies that meet the assumption may include randomness in the decisions, and yet more generally, decisions may be based on the whole past of the processes listed in item (i) above (for example, one may always select the server that, at the time of routing, has been idle most). The proof of this claim is standard, and thus omitted (for a proof of a closely related statement, albeit for a different model, the reader is referred to the appendix of [5]).

The first result will be concerned with the diffusion scale processes

$$\widehat{X}_n(t) = n^{-1/2}(X_n(t) - N_n), \tag{20}$$

$$\widehat{Q}_n(t) = n^{-1/2}Q_n(t) = \widehat{X}_n(t)^+, \qquad \widehat{I}_n(t) = n^{-1/2}I_n(t) = \widehat{X}_n(t)^-, \tag{21}$$

(where (11) was used), and

$$L_n(t) = n^{-1/2} \sum_{k \in K_n} \mu_{kn} \int_0^t I_{kn}(s)ds. \tag{22}$$

The initial number of customers in the system is assumed to satisfy

$$\widehat{X}_n(0) \Rightarrow \xi_0, \quad \text{as } n \to \infty, \tag{23}$$

where ξ_0 is a random variable satisfying $\xi_0 \geq 0$ with probability one. Let w be a standard Brownian motion, independent of ξ_0, and let $\mathcal{F}_t$ be the $\mathbb{P}$-completion of the smallest σ-field with respect to which w_s, $0 \leq s \leq t$ and ξ_0 are measurable. Denote $\beta = \widehat{\lambda} - \widehat{\mu}$ and $\sigma = (\lambda C_{IA}^2 + \mu)^{1/2} = \lambda^{1/2}(C_{IA}^2 + 1)^{1/2}$. A pair (ξ, l) will be said to be a *solution to the Skorohod equation*

$$\xi(t) = \xi_0 + \beta t - \gamma \int_0^t \xi(s)ds + \sigma w(t) + l(t), \qquad t \geq 0, \tag{24}$$

with data (ξ_0, w), if ξ and l are continuous, $\{\mathcal{F}_t\}$-adapted processes satisfying the following conditions $\mathbb{P}$-a.s.:

- equation (24) holds;
- $\xi(t) \geq 0$, $t \geq 0$;
- l is non-decreasing;
- $\int_{[0,\infty)} \mathbf{1}_{\{\xi(t)>0\}} dl(t) = 0.$

It is well-known [1] that there exists a unique solution (ξ, l) to equation (24) with given data. Because the drift term is linear, the process ξ is often referred to as a ROU process when $\gamma > 0$. When $\gamma = 0$, ξ is a RBM.

Theorem 2.1. *Let* $\{A_n, X_n, Q_n, I_n, B_{kn}, T_{kn}, D_{kn}, Z_n\}$ *be any sequence of processes satisfying all assumptions stated above. Then* $(\widehat{X}_n, L_n, \widehat{Q}_n, \widehat{I}_n)$ *converge in distribution, uniformly on compacts, to* $(\xi, l, \xi, 0)$*, where* (ξ, l) *denotes the unique solution to the Skorohod equation* (24) *with data* (ξ_0, w)*.*

Remark 2.1. Note that the theorem addresses any sequence of processes satisfying the assumptions, not one dictated by a specific routing policy. Yet the limit process ξ does not depend on the policy. As is well known, the ROU process, as a reflected diffusion on $\mathbb{R}_+$, has the property that it is strictly positive at any given time $t > 0$, with probability one. Roughly speaking, this means that, with high probability, for most times, all servers are busy. For a heuristic argument, consider a simplified model in which all server are busy all time. Clearly, the routing policy, affecting which server is busy at each time, plays no role in this scenario. This may explain the asymptotic insensitivity to the policy in the true model. Moreover, in view of this property, it is reasonable to expect that the model is asymptotically equivalent to the M/M/1+N (or M/M/1) with service rate given by $\sum_k \mu_{kn}$. This provides an explanation why the limit should be a ROU (or a RBM) process.

Example 2.1. *α-parametrization, homogeneous servers.* Assume that the number of servers in given by $N_n = [\nu n^\alpha]$, for some $\nu > 0$ and $\alpha \in [0, 1)$. Assume all servers work at the same rate

$$\mu_{kn} = \mu_{1n} = \frac{n}{N_n}(\mu + \widehat{\mu} n^{-1/2}).$$

Then $\bar{\mu}_n = n^{-1} \sum_k \mu_{kn} \to \mu$, while $\widehat{\mu}_n = n^{-1/2}(\sum_k \mu_{kn} - n\mu) \to \widehat{\mu}$. By Theorem 2.1, the limit process is independent of α and of the routing policy (so long as it satisfies the hypotheses). Note that the case $\alpha = 1$ is not covered because (9) requires that $\mu_{1n} \to \infty$.

Example 2.2. *α-parametrization, fixed number of large server pools.* Fix $\alpha \in [0,1)$ and let N_n be as in the above example. Assume that for some constant l, constants $\nu_1, \ldots, \nu_l$ that sum up to 1, and constants $M_1, \widehat{M}_1, \ldots, M_l, \widehat{M}_l$, there is, for each i, a pool of $N_n^{(i)} := \nu_i N_n + O(1)$ servers, each working at rate

$$\mu_{kn} = \frac{\nu_i n}{N_n^{(i)}}(M_i + \widehat{M}_i n^{-1/2}),$$

except for the case $\alpha = 0$, where this should hold without the $O(1)$ term. Note that for pool i, $\mu_{kn} \sim n^{1-\alpha}(M_i + n^{-1/2}\widehat{M}_i)$. The theorem holds with $\bar{\mu}_n \to \mu = \sum_i \nu_i M_i$, and $\widehat{\mu}_n \to \widehat{\mu} = \sum_i \nu_i \widehat{M}_i$.

Example 2.3. *Two server pools with rates at different scales.* This example considers a combination of two regimes on the α-scale. Let μ_1 and μ_2 be positive constants, and let $\widehat{\mu}_1$ and $\widehat{\mu}_2$ be constants. Denote $\mu = \mu_1 + \mu_2$ and $\widehat{\mu} = \widehat{\mu}_1 + \widehat{\mu}_2$. Let α_1 and α_2 be some constants in $[0,1)$, and assume that, for $i \in \{1,2\}$, $N_n^{(i)}$ servers work at rate $\mu_n^{(i)}$, where $N_n^{(i)} = [n^{\alpha_i}]$, and

$$\mu_n^{(i)} = \frac{n}{N_n^{(i)}}(\mu_i + \widehat{\mu}_i n^{-1/2}).$$

The total number of servers is given by $N_n = N_n^{(1)} + N_n^{(2)}$. It is easy to check that $\bar{\mu}_n \to \mu$, and that $\widehat{\mu}_n \to \widehat{\mu}$. Again, Theorem 2.1 applies, and its conclusion is independent of α_1 and α_2.

2.2 Convergence in the NDS regime

We will retain all assumptions imposed thus far and introduce some new ones. First, in accordance with the α-parametrization with $\alpha = 1/2$, we assume that

$$N_n = \nu n^{1/2} + o(n^{1/2}), \tag{25}$$

for some constant $\nu > 0$. However, in contract to Example 2.1, we will allow for server heterogeneity. Letting

$$\widehat{\mu}_{kn} = n^{-1/2}\mu_{kn}, \quad k \in K_n, \tag{26}$$

we assume that the empirical measure of $\{\widehat{\mu}_{kn}\}$ converges weakly, namely that

$$m_n := \frac{1}{N_n}\sum_{k=1}^{N_n} \delta_{\widehat{\mu}_{kn}} \to m, \tag{27}$$

for some probability measure m on $\mathbb{R}_+$. Here, δ_x denotes the unit point mass at x. Recall that conditions (7)–(10), that also concern μ_{kn}, are still in force. To see how they are related to (27), let $\mu^{(1)} = \int y m(dy)$ and $\mu_n^{(1)} = \int y m_n(dy)$ denote respective first moments of m and m_n. Under the conditions just introduced, (7) is equivalent to $\mu_n^{(1)} \to \nu^{-1}\mu$. Hence by Fatou's Lemma

$$\mu^{(1)} \le \nu^{-1}\mu. \tag{28}$$

In general equality need not hold, and we do not require that it does (a case with strict inequality is presented in Example 2.6 below).

The further structure and assumptions we will now introduce are related to the fact that we analyze delay and service time experienced by individual customers. Thus customers must be labeled and information about them, such as their relative positions in the queue, have to be contained in the filtration. Let us then number all customers not initially in the system in order of arrival: for $t \geq 0$, $C_n(t)$ will denote the serial number of the customer to arrive first after time t. In other words, $C_n(t) = A_n(t) + 1$. Given t, the arrival time of customer $C_n(t)$ is $AT_n(t)$ (see (4)). It is assumed that the customers are served according to a *first-come-first-served* discipline, and that service is *non-interruptible*, in the sense that whenever a server is assigned a new job, it works continuously on it until completion.

Next, the stochastic primitive Z and the process Z_n determine how many abandonments occur up to a given time, but it will now be important to identify which customer abandons at each epoch where $\Delta Z_n > 0$. We need an additional stochastic primitive for that. We let U_i, $i \geq 1$, be i.i.d. random variables, uniform over $[0, 1]$, independent of all other primitives. U_i will be used to select a customer uniformly among those present in the queue, according to their positions. Namely, if $\Delta Z_n(t) > 0$ then the customer to abandon at t is the one that at time $t-$ is at position i, where i is the unique $i \in \{1, 2, \ldots, Q_n(t-)\}$ for which $U_{Z_n(t)} Q_n(t-) \in [i-1, i)$. This sets up an independent exponential clock for each customer (see Lemma A.4(i) of [3]).

At a given time t, a customer may be in position $i \in \{1, 2, 3, \ldots\}$ in the queue, it may be in service with server $k \in K_n$, it may have completed service, or it may have abandoned the queue. This information is encoded in the following random variables. Fix s, t. $POS_n(s, t)$ will be an $\mathbb{N} \cup \{\infty\}$-valued random variable representing the position of customer $C_n(s)$ at time t, where position 1 corresponds to the head of the line. It takes the value ∞ if the customer is not in the queue at that time. $AB_n(t)$ is the time when $C_n(t)$ abandons the queue; it takes the value ∞ on the event that this customer never abandons. $RT_n(t)$ and $RD_n(t)$ take values in $\mathbb{R}_+^*$ and $K_n \cup \{\infty\}$, and represent the time when $C_n(t)$ is routed to a server and the identity of the assigned server, respectively (RT and RD are mnemonics for 'routing time' and 'routing decision'). They both take the value ∞ on the event that the customer abandons the queue. $DEP_n(t)$, taking values in $\mathbb{R}_+^*$, will denote the time when the same customer completes service; it is equal to ∞ on the event $\{AB_n(t) < \infty\}$. $EX_n(t) = DEP_n(t) \wedge AB_n(t)$ is the time when the customer exits the system either by completing service or by abandoning. Note that it is always finite.

The result will be concerned with the quantities $\Delta_n(t)$ and $\Sigma_n(t)$, defined as

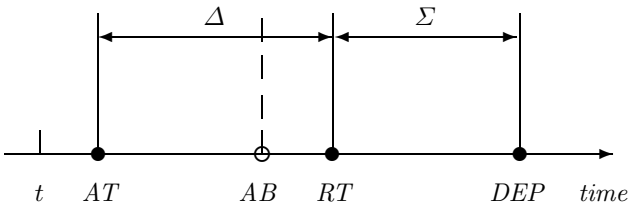
$$\Delta_n(t) = RT_n(t) - AT_n(t), \qquad \Sigma_n(t) = DEP_n(t) - RT_n(t), \qquad \text{on } \{AB_n(t) = \infty\},$$

$$\Delta_n(t) = \Sigma_n(t) = \infty, \qquad \text{on } \{AB_n(t) < \infty\},$$

representing, respectively, the time $C_n(t)$ spends in the queue, since arrival until being accepted to service, and the time it spends in service (the letters Δ and Σ are mnemonics for delay and service time). In case of abandonment we have set these random variables to ∞. See Figure 2.

The filtration $\mathbb{F}_n = \{\mathcal{F}_n(t), t \geq 0\}$ referred to in what follows, is the same as the one from Assumption 2.1.

Assumption 2.2. *Fix n. Given any $t \in [0, \infty)$, the random times $RT_n(t)$, $DEP_n(t)$ and $AB_n(t)$ are stopping times of the filtration $\mathbb{F}_n$. Given any s and t, and any $k \in K_n$, $i \in \mathbb{Z}$ and $b \in \mathbb{R}$, the*



AT = Arrival Time
RT = Routing Time
DEP = Departure Time
AB = Abandonment Time
Δ = Delay
Σ = Service Time

Figure 2: *Quantities associated with the customer to arrive first after time t. On the event of abandonment, Δ and Σ are set to ∞.*

events $\{RT_n(s) \le t, RD_n(s) = k\}$ and $\{Z_n(t) = i, U_i \le b\}$, and the random variable $POS_n(s,t)$ are measurable on $\mathcal{F}_n(t)$.

The comment following Assumption 2.1, regarding its wide scope, applies for Assumption 2.2 as well.

Denote the diffusion scale delay and service time processes by

$$\widehat{\Delta}_n(t) = n^{1/2}\Delta_n(t), \qquad \widehat{\Sigma}_n(t) = n^{1/2}\Sigma_n(t), \qquad t \ge 0,$$

and let

$$\Pi_n(t) = (\widehat{\Delta}_n(t), \widehat{\Sigma}_n(t)), \qquad t \ge 0.$$

Note that the two components of $\Pi_n(t)$ correspond to the delay and service time of the same customer. Our main result asserts that the scaled delay and service time processes jointly converge in the sense of finite dimensional distributions, and identifies the limit.

Theorem 2.2. *Let all assumptions stated in this section hold. Given $j \in \mathbb{N}$ and $0 < t_1 < t_2 < \cdots < t_j < \infty$, we have*

$$(\Pi_n(t_1), \Pi_n(t_2), \ldots, \Pi_n(t_j)) \Rightarrow ((\bar{\xi}(t_1), \eta_1), (\bar{\xi}(t_2), \eta_2), \ldots, (\bar{\xi}(t_j), \eta_j)), \qquad \text{as } n \to \infty. \tag{29}$$

Here, $\bar{\xi}$ is the normalized version $\mu^{-1}\xi$ of the process ξ defined by the Skorohod equation (24)*, and η_i are i.i.d., independent of ξ, and the distribution of η_1 over $[0,\infty)$ is given by*

$$a_0\delta_0(dx) + f(x)\,\mathrm{Leb}(dx), \tag{30}$$

where δ_0 an atom at the origin, Leb *denotes the Lebesgue measure on $[0,\infty)$, and*

$$a_0 = 1 - \frac{\nu\mu^{(1)}}{\mu}, \qquad f(x) = \frac{\nu}{\mu}\int y^2 e^{-yx} m(dy), \quad x \in [0,\infty). \tag{31}$$

Remark 2.2. (a) *The probability to abandon tends to zero but abandonments are not negligible.* Recall that the delay and service time are set to infinity on the event of abandonment. Since the random variables on the r.h.s. of (29) take finite values, the convergence stated in the theorem implies (recalling the convention on convergence of $\mathbb{R}_+^*$-valued r.v.s) that the probability that any one of the customers $C_n(t_i)$ $1 \le i \le j$ abandons, converges to zero. This, however, does not mean

that abandonment can be neglected. Indeed, it affects the limit process via the integral term $\gamma \int \xi$ of equation (24). By a rough calculation, each of the $O(n)$ customers entering the system in a unit time spends $O(n^{-1/2})$ units of time at the queue, and thus $O(n^{1/2})$ customers do abandon (the abandonment fraction is therefore $O(n^{-1/2})$). As a result, the number of abandoning customers is at the same scale as the queue-length, which explains why this ingredient of the model must affect the limiting dynamics (24).
(b) *Limiting service time process is a white noise.* Because the limiting finite dimensional distribution of $(\widehat{\Sigma}_n(t_1), \widehat{\Sigma}_n(t_2), \ldots, \widehat{\Sigma}_n(t_j))$ is that of $(\eta_1, \eta_2, \ldots, \eta_j)$ and η_i are independent, the result expresses convergence of the scaled service time to a 'white noise' process.
(c) *Limiting service time is a size-biased mixture of exponentials.* Consider the case where $a_0 = 0$, so f integrates to one. In this case f is the p.d.f. of a mixed exponential. A corresponding random variable (such as η_1) can be obtained by first drawing a random variable Y from the distribution $ym(dy)/\int zm(dz)$, and then letting η_1 be exponentially distributed with parameter Y. The result thus expresses the fact that, asymptotically, a typical customer is served exponentially with a rate Y, where $\mathbb{P}(Y \in dy)$ is proportional to $ym(dy)$.
(d) *Explanation of the mixture of exponentials.* Remark 2.1 can heuristically explain the distribution of the service time as a size-biased mixture of exponentials. Indeed, in a simplified model where all servers are busy all the time, the departure process is clearly Poisson with rate given by the sum $\sum_k \mu_{kn}$. Moreover, a customer that is at the head of the line will be assigned the "first server whose exponential clock ticks", namely, it will be assigned server k with probability proportional to μ_{kn}, and the resulting service time will be a mixture of exponentials. Taking into account that the limiting empirical distribution of the (normalized) service rates is m, this model gives rise to the p.d.f. f (31).
(e) *What may cause a_0 to be positive.*

Recall the foregoing discussion on $\mu^{(1)}$ versus μ. In case the inequality (28) holds with equality, the weight a_0 of the Dirac measure vanishes, and we obtain a pure mixture of exponentials for the distribution of η_1, just as discussed in the above remark. When $a_0 > 0$, the result asserts that, with probability a_0, a typical customer experiences an extremely short service time, that is $o(n^{-1/2})$. This may happen when relatively few servers work much faster than most other servers, so they affect the overall service rate $\mu_n^{(1)}$ considerably, while they are too few to affect the limiting distribution m. Example 2.6 below identifies such a situation.

The following result is an immediate outcome of Theorem 2.2. It demonstrates that the sojourn time behavior is dramatically different than in the conventional and QED asymptotic regimes. Denote by $SO_n(t)$ the sojourn time experienced by $C_n(t)$, namely $\Delta_n(t) + \Sigma_n(t)$, and set $\widehat{SO}_n(t) = n^{1/2} SO_n(t)$.

Corollary 2.1. *Under the assumptions and notation of Theorem 2.2, one has the finite dimensional convergence of sojourn time, as follows*

$$(\widehat{SO}_n(t_1), \widehat{SO}_n(t_2), \ldots, \widehat{SO}_n(t_j)) \Rightarrow (\bar{\xi}(t_1) + \eta_1,\, \bar{\xi}(t_2) + \eta_2,\, \ldots,\, \bar{\xi}(t_j) + \eta_j), \qquad \text{as } n \to \infty.$$

Following are some examples of settings that are covered by the framework of Theorem 2.2.

Example 2.4. *Homogeneous servers.* Assume that $\mu_{kn} = \mu n^{1/2} + \widehat{\mu}$ for all n and all $k \in K_n$, where

N_n satisfies (25). In this case f is simply exponential with mean μ. Also, $a_0 = 0$ and there is no atom at the origin.

Example 2.5. *Fixed number of large server pools.* Assume now that for some constant l and constants $\nu_1, \ldots, \nu_l$ summing up to 1, one has $N_n^{(i)} := \nu_i N_n + O(1)$ of the servers working at rate

$$\mu_{kn} = \frac{\nu_i n}{N_n^{(i)}} M_i, \qquad i = 1, 2, \ldots, l \tag{32}$$

(where N_n satisfies (25)). Note that this is a special case of Example 2.2, and thus $\mu = \sum \nu_i M_i$, $\widehat{\mu} = 0$ (one can work with a more general $\widehat{\mu}$). Now, m is given by $\sum_{i=1}^{l} \nu_i \delta_{M_i}$, and $a_0 = 0$. Hence η_1 is a mixed exponential, having the distribution $f(x)\mathrm{Leb}(dx)$, where

$$f(x) = \frac{1}{\mu} \sum_{i=1}^{l} \nu_i M_i^2 e^{-M_i x}, \quad x \in [0, \infty).$$

Example 2.6. *Fixed number of large server pools, and few fast servers.* This example is based on the previous one, but identifies a situation where Fatou's inequality does not hold as equality. As before, assume there are $N_n^{(i)} = \nu_i N_n + O(1)$ servers working at rate according to (32), but now assume in addition that N_0 servers serve at rate $\mu_{kn} = M_0 n$, where both $N_0 \geq 1$ and $M_0 > 0$ are constants independent of n. Then $\lim_n n^{-1} \sum_k \mu_{kn} = \mu + M_0 N_0$. η_1 has distribution $a_0 \delta_0(dx) + f(x)\mathrm{Leb}(dx)$, where now

$$a_0 = 1 - \frac{\mu}{\mu + M_0 N_0}, \qquad f(x) = \frac{1}{\mu + M_0 N_0} \sum_{i=1}^{l} \nu_i M_i^2 e^{-M_i x}, \quad x \in [0, \infty).$$

Example 2.7. *Uniform distribution of rates.* For some fixed $0 < a < b$, assume $\mu_{kn} = (b + kn^{-1/2})n^{1/2}$ for $k \in \mathbb{Z}$ such that $|k| < an^{1/2}$ (in this example the index set K_n is not of the form $\{1, 2, \ldots, N_n\}$ but this will not cause confusion). The distribution of the rates is symmetric about b, $N_n = 2an^{1/2} + O(1)$, and so $\nu = 2a$, $\nu^{-1}\mu = b$, and $\widehat{\mu} = 0$. The limiting distribution m is clearly uniform over $[b - a, b + a]$. Of course, one may consider much more general distributions than the uniform.

Finally, let us mention that the steady state distribution of both the limiting processes $\bar{\xi}$ and η are not hard to compute. Clearly, the distribution of η does not depend on time, and is given by Theorem 2.2. A simple calculation gives its mean, namely

$$\mathbb{E}[\eta_1] = \int_0^\infty x f(x) dx = \frac{\nu}{\mu}. \tag{33}$$

The process ξ is positive recurrent when $\gamma > 0$. Its steady state distribution is given by the conditional distribution of N given $N \geq 0$, where N is normally distributed with parameters $(b, c^2) := (\frac{\beta}{\gamma}, \frac{\sigma^2}{2\gamma})$ (see [19]). When $\gamma = 0$, a stability condition is needed, namely that $\beta < 0$, and the steady state distribution is then exponential with mean $\frac{\sigma^2}{2|\beta|}$ [19]. Denoting by ξ_∞ (resp., $\bar{\xi}_\infty$) a random variable distributed according to the steady state of ξ (resp., $\bar{\xi}$), we have

$$\mathbb{E}[\bar{\xi}_\infty] = \mu^{-1}\mathbb{E}[\xi_\infty] = \frac{1}{\mu} \frac{\int_0^\infty x e^{-(x-b)^2/2c^2} dx}{\int_0^\infty e^{-(x-b)^2/2c^2} dx} \qquad \text{when } \gamma > 0, \tag{34}$$

and

$$\mathbb{E}[\bar{\xi}_\infty] = \frac{\sigma^2}{2\mu|\beta|} \qquad \text{when } \gamma = 0 \text{ and } \beta < 0. \tag{35}$$

One may regard (33)–(35) as the expected service and delay asymptotic values under the NDS regime in steady state, and accordingly define the slowdown as the ratio between the corresponding expected sojourn time and expected service time. This gives

$$\text{slowdown} = 1 + \frac{1}{\nu} \frac{\int_0^\infty x e^{-(x-b)^2/2c^2} dx}{\int_0^\infty e^{-(x-b)^2/2c^2} dx} \qquad \text{when } \gamma > 0, \tag{36}$$

and

$$\text{slowdown} = 1 + \frac{\sigma^2}{2\nu|\beta|} \qquad \text{when } \gamma = 0 \text{ and } \beta < 0. \tag{37}$$

Note that this discussion is only formal as we have not proved that the steady state distributions of the prelimit processes converge, in any sense, to those of $\bar{\xi}$ and η. This will be addressed in future work.

3 Conclusion

We have identified the joint limit law of delay and service time, and consequently, that of sojourn time, under the NDS regime for a system with heterogenous servers. The limit law of sojourn time is distinct from that under any other asymptotic regime, and captures the comparability of delay and service time. We expect that delay and service time are comparable under similar parametric assumptions in far greater generality. It is thus desirable to consider models with general service time distributions as well as network settings. One would also like to complete the steady state analysis discussed at the end of Section 2. It is plausible that explicit expressions for slowdown such as (36)–(37) can be developed for a general set up.

Control of queueing networks under both the conventional and QED diffusion regimes (as well as fluid regimes) has been a very active research topic in recent years. From the viewpoint of the customer, sojourn time is an important measure of performance, and it is desired to seek control schemes that optimize it. In the conventional regime, a control policy that is good for minimizing a delay-related cost is also good for minimizing a sojourn time cost, because service time is negligible. A control problem with sojourn time cost (or, more generally, cost defined in terms of delay and service time simultaneously) is natural to be considered in the NDS regime, where it is distinct from formulations based on either delay or service time. Because it involves the limiting distribution of service time in addition to reflected diffusion processes representing delay, such a formulation would lead to diffusion control problems and asymptotically optimal control schemes distinct from those obtained under the conventional regime.

Acknowledgment. I thank Avi Mandelbaum for referring me to information from [8] relevant to the discussion in Section 1.3, and him and the Technion's SEELab staff for providing me with the additional statistical data alluded to in that discussion. I also thank three referees for valuable comments that helped improve the paper considerably.

References

[1] R. F. Anderson and S. Orey. Small random perturbation of dynamical systems with reflecting boundary. *Nagoya Math. J.* Vol. 60 (1976), 189–216

[2] M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3-4) (2005), 287–329.

[3] R. Atar. A diffusion regime with non-degenerate slowdown: Appendix (published online).

[4] R. Atar, A. Mandelbaum and M. I. Reiman. Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy-traffic. *Ann. Appl. Probab.* 14 (2004), no. 3, 1084–1134

[5] R. Atar and M. I. Reiman. Asymptotically optimal dynamic pricing for network revenue management. Work in progress.

[6] R. Atar and A. Shwartz. Efficient routing in heavy traffic under partial sampling of service times. *Math. Oper. Res.*, 33: 899 - 909 (2008)

[7] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association*, March 2005, Vol. 100, No. 469

[8] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. Statistical analysis of a telephone call center: a queueing-science perspective. An extended version of [7]. Available online at http://ie.technion.ac.il/serveng/References/references.html

[9] O. Garnett, A. Mandelbaum and M. Reiman. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* 4, 208-227 (2002)

[10] I. Gurvich. *Design and control of the M/M/N queue with multi-class customers and many servers.* M.Sc. Thesis, Technion, July 2004

[11] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29 (1981), no. 3, 567–588.

[12] J. M. Harrison. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Ann. Appl. Probab.* 8 (1998), no. 3, 822–848.

[13] J. M. Harrison. A broader view of Brownian networks. *Ann. Appl. Probab.* 13 (2003), no. 3, 1119–1150.

[14] H. J. Kushner. *Heavy traffic analysis of controlled queueing and communication networks.* Applications of Mathematics (New York), 47. Stochastic Modelling and Applied Probability. Springer-Verlag, New York, 2001.

[15] A. Mandelbaum. QED Q's. Notes from a lecture delivered at the Workshop on Heavy Traffic Analysis and Process Limits of Stochastic Networks, EURANDOM, September 2003 http://ie.technion.ac.il/serveng/References/references.html

[16] A. Mandelbaum. Private communication.

[17] A. Mandelbaum and G. Shaikhet. Private communication.

[18] A. R. Ward and P. W. Glynn. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* 43, 103-128.

[19] A. R. Ward and P. W. Glynn. Properties of the reflected Ornstein-Uhlenbeck process. *Queueing Systems.* Vol. 44, No. 2 June, 2003

[20] W. Whitt. Understanding the efficiency of multi-server service systems. *Management Sci.* 38 (5) 708723.

[21] W. Whitt. *Stochastic-process limits.* Springer Series in Operations Research. Springer-Verlag, New York, 2002

[22] W. Whitt. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* Vol. 51 No. 4 531–542 (2003)

[23] W. Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* Vol. 50, No. 10, October 2004, pp. 1449–1461

Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel