A New Transitory Queueing Model and its Process Limits

Harsha Honnappa
EE Department
Univ. of Southern California
honnappa@usc.edu

Rahul Jain
EE & ISE Department
Univ. of Southern California
rahul.jain@usc.edu

Amy Ward
Marshall School of Business
Univ. of Southern California
amyward@usc.edu

***Abstract*—We introduce the $\Delta_{(i)}/GI/1$ queue, a new queueing model. In this model, customers from a given population independently arrive according to some given distribution F. Thus, the arrival times are an ordered statistics, and the inter-arrival times are differences of consecutive ordered statistics. They are served by a single server which provides service according to a general distribution G, with independent service times. We develop fluid and diffusion limits for the various stochastic processes, and performance metrics. The fluid limit of the queue length is observed to be a reflected process while the diffusion limit is observed to be a function of a Brownian motion and a Brownian bridge, reflected through a directional derivative of the usual Skorokhod reflection map. We also observe what may be interpreted as a 'transient' Little's law. Sample path analysis reveals various operating regimes where the diffusion limit switches between a free diffusion, a reflected diffusion process and the zero process, with possible discontinuities during regime switches.**

***Index Terms*—Queueing models; transitory queueing systems; fluid and diffusion limits, distributional approximations; directional derivatives, M_1 topology**

I. Introduction

Most of modern queueing theory is concerned with scenarios where arrival and service processes are stationary and ergodic. That the arrival process is a renewal process with i.i.d. inter-arrival times is a common modeling assumption. This is mathematically convenient as it allows full use of the tools that renewal theory and ergodic theory provide. But it need not be true in some queueing scenarios. For example, in some queueing scenarios, each arriving customer takes an independent decision of when to arrive. Even if we assume that every arriving customer draws an arrival time from the same distribution, this need not lead to a renewal arrival process. Moreover, such a distribution may only have finite support meaning that the system is transient. Thus, these scenarios do not seem to fit the standard, single-server models in queueing theory such as $M/M/1$, $M/G/1$, etc.

There has been an interest in developing a theory for transient queues [15]. The first such models for time-dependent queues were the early attempts of Newell [22] (see also [15], [21], [8], [20]), and the more recent developments for $M_t/M_t/1$ in [18] and state dependent Markovian queues in [19]. However, in all of these the assumption of a renewal arrival process (albeit time-inhomogeneous) remains ubiquitous. Furthermore, all such models still assume a queueing system operating forever, with an infinite population of customers and a steady state. In contrast, many queueing systems serve only a finite number of customers, and in fact, the queueing system itself may be *transitory*, i.e., it may operate only in a finite window of time, meaning that the concept of a steady state does not exist. A goal of the present work is to propose queueing models, and develop their theory, that are relevant for such *transitory queueing systems*.

Such models can arise when service starts at a certain time, and customers may choose to arrive early. For example, when customers go to a rock concert in a Greek theater, they may choose to arrive before the gates open, or arrive any time after until the gates close. Such a scenario was studied as the *concert arrival game* in [13], [10]. Other scenarios where such a model may be relevant include queueing outside stores for black Friday sales, outside Apple store before new product launches, DMV or postal offices, lunch cafeterias, etc.

We introduce a new queueing model that has a finite population of customers whose arrival process is not a renewal process, and arrivals happen in a finite time window; in particular, it is a transitory queueing model. Consider n customers who arrive into a single-server queue. Each customer's time of arrival is sampled i.i.d. from a distribution F. Then, the times of arrival are an ordered statistics. Service times have a general distribution G and are i.i.d. We call this the $\Delta_{(i)}/GI/1$ *queueing model*. The $\Delta_{(i)}/GI/1$ queue in the notation of Kendall [16] is a $\Delta_{(i)}/GI/1$ queue, where $X_{(i)}$ is the ith order statistic from a sample of size n from the distribution F and $\Delta_{(i)} = (X_{(i)} - X_{(i-1)})$. We note, without proof, that the exact analysis of the $\Delta_{(i)}/GI/1$ parallels that of a $M_t/GI/1$ queue, but with the added complication that the first and the last "inter-arrival" periods are not renewal periods complicating the boundary conditions used for the transient queue length probability distributions. Therefore, we develop fluid and diffusion limits for this queueing model.

To develop our fluid and diffusion limits for the $\Delta_{(i)}/GI/1$ model, we scale up the population size and accelerate the service process, and use the functional strong law of large numbers (FSLLN) for random walks and the Glivenko-Cantelli theorem to establish fluid limits for the service and arrival processes respectively. Then, using Skorokhod's reflection mapping theorem [26], [9], we can obtain fluid limits for the queue length process and the workload (the time it

would take to serve current jobs in the queue) process. The diffusion process limits of the service and arrival processes can be obtained by use of the functional central limit theorem (FCLT) for renewal processes and for empirical processes respectively. The queue length diffusion process limit involves a map which can be interpreted as the directional derivative of the Skorokhod reflected fluid netput in the direction of a diffusion refinement of the netput process. We also note that our diffusion process convergence results are in Skorokhod's M_1 topology on the space $\mathcal{D}_{\lim}[0,\infty)$, the space of functions that are right or left continuous at every point, and right continuous at 0.

The most standard heavy traffic approximation in the literature for a single server queue is a reflected Brownian motion diffusion approximation; see, for example, Chapter 6 in [5] and the pioneering work in [12]. The reflected Brownian motion approximation is relevant for a $G/G/1$ queue in which the inter-arrival and service times are either independent or only exhibit weak dependence, and in which both the inter-arrival and service time distributions have finite second moments. In the case that there is strong dependence in the inter-arrival or service times and/ or in the case that the distributions are heavy-tailed (with an infinite second moment), non-Brownian limits arise; see, for example Chapter 4 in [31] and the overview paper [29]. However, the $\Delta_{(i)}/GI/1$ queue does not fit that framework. In fact, the $\Delta_{(i)}/GI/1$ queue has a closer connection with single server queues that have a time-varying arrival rate.

Thus, it is pertinent to compare the $\Delta_{(i)}/GI/1$ model to the already studied $M_t/M_t/1$ queueing model. One of the earliest papers on this model is [20]. Strong approximations for the model were later developed in [18]. However, all such models still assume a renewal arrival process (albeit a time-inhomogeneous one). Thus, the arrival process in the $\Delta_{(i)}/GI/1$ model cannot be obtained simply by windowing the process for the $M_t/GI/1$ queue since the first and last inter-arrival time in such a time-window need not be renewal periods.

Perhaps the work closest to the current paper is [17], where the author considers the same setup as we have, but does not allow early arrivals. The paper develops diffusion approximations to the queue length in separate, distinct intervals and the maximum queue length process. However, without establishing a "process-level" convergence over all time, such a result is rather incomplete. In fact, it is not difficult to derive point-wise limits to the queue length process. Establishing "process-level" convergence for such limiting processes in an appropriate topology is the main mathematical difficulty, as was also observed in [18] for the $M_t/M_t/1$ model.

We note that there is a 'transient' Little's Law that holds for the $\Delta_{(i)}/GI/1$ model, i.e., the diffusion limit of the work load process converges to diffusion limit of the queue length process divided by the service rate plus a diffusion term which is non-zero only in the 'overloaded' regime. This is a handy result analogous to a similar Little's Law for diffusion approximations to the $GI/GI/1$ queues (see Chapter 6 ,[5]). However, it is to be noted that, unlike the 'standard' Little's Law, this result holds only in the case of a First-Come-First-Serve (FCFS) service discipline.

II. Preliminaries

A. Queue Model

Consider a single server, infinite buffer queue that is non-preemptive and non-idling, and say, starts empty. Customers in the queue will be assumed served on a first-come-first-served (FCFS) basis, though this is not essential. Arriving customers do not balk or renege once they commit to queueing. We also allow early-bird arrivals so customers can queue up before service starts . Most importantly, it is assumed there is a large but finite number of customers who arrive for service over some finite time interval.

Let n be the customer population size. Arriving customers independently sample an *arrival time* T_i, $i = 1, \ldots, n$, from a fixed cumulative distribution function F that is assumed to have support $[-T_0, T] \subset \mathbb{R}$, where $-T_0 \leq 0$ and $T > 0$. Thus, $F(-T_0) = 0$ and $F(T) = 1$. Customers enter the queue in increasing order of the sampled arrival times. Let $A(t)$ be the number of customers who have arrived by time t, which is usually called the *arrival process*. This is defined to be

$$A(t) := \sum_{i=1}^{n} \mathbf{1}_{\{T_i \leq t\}}. \tag{1}$$

Let $\nu_i, i \geq 1$ be a sequence of independent and identically distributed (IID) random variables, independent of the arrival times T_i, with mean $\mathbb{E}\nu_i = 1/\mu$ and finite variance σ^2, whose cumulative distribution function G has support $[0,\infty)$. Define S to be a renewal process in terms of ν_i as

$$S(t) := \sup\{m \geq 1 | V(m) \leq t\},\ \forall t \geq 0, \tag{2}$$

where $V(m) := \sum_{i=1}^{m} \nu_i$. $S(t)$ is interpreted as the *service process* of the queue. Here, ν_i is the service time for the ith potential customer, and μ as the service rate offered by the system. Note that S is defined for all $t \geq 0$, so service starts at time 0 in the $\Delta_{(i)}/GI/1$ model, and we define $S(t) = 0$ for all $t < 0$. Thus, the service process $S(t)$ can be interpreted as the number of customers that could be served if the server were busy all the time in the interval $[0, t]$.

Now, $V(m)$ can be interpreted as the amount of work (in units of time) presented by m customers. Let Z represent the *virtual waiting time* process, defined using V as

$$Z(t) := V(A(t)) - B(t) - t\mathbf{1}_{\{t \leq 0\}}, \tag{3}$$

where $B(t)$ is the busy time process defined as

$$B(t) := \left(\int_0^t \mathbf{1}_{\{Q(s)>0\}} ds \right) \mathbf{1}_{\{t \geq 0\}},\ \forall t \in [-T_0, \infty). \tag{4}$$

$B(t)$ is the amount of time the server has been busy in the interval $[0, t]$, and clearly $B(t) \leq t$. Thus, $Z(t)$ is the difference between the total amount of work presented by the arrivals up to time t and the amount of work completed by the server by t. $Z(t)$ can also be interpreted as the amount

of time a *virtual* arrival at time t would have to wait till it enters service. Note that this definition varies slightly from the standard definition due to the fact that an arrival at time $t < 0$ before service starts has to wait an extra t units of time for service to start, which accounts for the $-t\mathbf{1}_{\{t\leq 0\}}$ term.

Finally, let Q represent the *queue length process*, or the number of customers in service and waiting in the buffer at time t. This is defined in terms of the arrival and service processes as

$$Q(t) := A(t) - S(B(t)), \quad \forall t \in [-T_0, \infty). \tag{5}$$

As noted before, $S(t)$ is the number of service completions if the server is always busy in $[0, t]$. Thus, $S(B(t))$ counts the number of service completions by time t given that it is busy only for time $B(t)$ until that time. This then is the number of departures from the queue in $[0, t]$, and the queue length is the difference between the number of arrivals and the number of departures by time t.

We observe that this model is intractable to exact analysis. Therefore, we develop asymptotic approximations as the population size increases.

B. Basic results

Notation Unless noted otherwise, all intervals of time are subsets of $[-T_0, \infty)$, for a given $-T_0 \leq 0$. Let $\mathcal{D}_{\lim} := \mathcal{D}_{\lim}[-T_0, \infty)$ be the space of functions $x : [-T_0, \infty) \to \mathbb{R}$ that are right-continuous at $-T_0$, and are either right or left continuous at every point $t > -T_0$. Note that this differs from the usual definition of the space $\mathcal{D}$ as the space of functions that are right continuous with left limits (cadlág functions). We denote almost sure convergence by $\xrightarrow{a.s}$ and weak convergence by $\Rightarrow$. The topology of convergence is indicated by the tuple (S, m), where S is the metric space of interest and m is the metric inducing a metric topology on S. Thus, $X_n \xrightarrow{a.s} X$ in $(\mathcal{D}_{\lim}, U)$ as $n \to \infty$ indicates that $X_n \in \mathcal{D}_{\lim}$ converges to $X \in \mathcal{D}_{\lim}$ uniformly on compact sets (u.o.c.) of $[-T_0, \infty)$ almost surely. Similarly, $X_n \Rightarrow X$ in $(\mathcal{D}_{\lim}, U)$ as $n \to \infty$ indicates that $X_n \in \mathcal{D}_{\lim}$ converges weakly to $X \in \mathcal{D}_{\lim}$ uniformly on compact sets of $[-T_0, \infty)$. $(\mathcal{D}_{\lim}, M_1)$ indicates that the topology of convergence is the M_1 topology. $\bar{X}$ indicates a fluid-scaled or fluid limit process. $\hat{X}$ and $\tilde{X}$ are used to indicate diffusion-scaled and diffusion limit processes. We use $\circ$ is used to indicate composition of functions or processes. The indicator function is denoted by $\mathbf{1}_{\{\cdot\}}$ and the positive part operator by $(\cdot)_+$.

We now present known functional strong law of large numbers (FSLLN) or *fluid* limits, and functional central limit theorem (FCLT) or *diffusion* limits, for the arrival and service processes, as the population size n increases to ∞. Our convention is to superscript any process associated with the model having population size n by n. These limits are presented for the reader's convenience, and will be useful in later sections where fluid and diffusion limits for the queue length and the virtual waiting time processes are derived.

We start with the arrival process. For the rest of this section, let $A^n := A$ be the arrival process indexed by the population size n, and define the *fluid-scaled arrival process* as

$$\bar{A}^n := \frac{A^n}{n}.$$

Now, for the service process, we use an *acceleration* technique wherein the service rate is accelerated by multiplication with the population size so that $\mu^n := n\mu$. Correspondingly, the *scaled* service times are $\nu_i^n := \nu_i/n$ for $i = 1, \ldots, n$. Using these, we define the *accelerated service process* as

$$S^n(t) := \sup\left\{m \geq 1 | \sum_{i=1}^{m} \frac{\nu_i}{n} \leq t\right\} \quad t \geq 0,$$

and the *fluid-scaled service process* as

$$\bar{S}^n := \frac{1}{n} S^n.$$

Finally, Recall that we have

$$V(m) := \sum_{i=1}^{m} \nu_i.$$

This random variable can be interpreted as the amount of work (in units of time) presented to the server by m customers. We can define the *offered work process* (the amount of work offered by time t by a population of size n), as

$$\bar{V}^n(t) := \sum_{i=1}^{\lfloor nt \rfloor} \nu_i^n \quad \forall t \in [0, \infty). \tag{6}$$

Here, we have used the acceleration argument presented for the service process.

The following lemma establishes the fluid limits for these processes.

Proposition 1: As $n \to \infty$, in $(\mathcal{D}_{\lim}, U)$

$$(\bar{A}^n(t), \bar{S}^n(t)\mathbf{1}_{t\geq 0}, \bar{V}^n(t)\mathbf{1}_{t\geq 0}) \xrightarrow{a.s} (F(t), \mu t\mathbf{1}_{t\geq 0}, \frac{t}{\mu}\mathbf{1}_{t\geq 0}). \tag{7}$$

Remarks 1. The proof of Proposition 1 follows easily from standard results: The fluid arrival process limit is given by the Glivenko-Cantelli Theorem (see [6]). The fluid limits of the service process and the offered work process follow from the functional strong law of large numbers theorem for renewal processes (see [5]).

Next, using the fluid limits from Proposition 1, we present functional central limit theorem or diffusion limits, to the appropriately standardized or *diffusion-scaled* processes. The *diffusion-scaled arrival process* is defined as

$$\hat{A}^n(t) := \sqrt{n}\left(\bar{A}^n(t) - F(t)\right) \quad \forall t \in [-T_0, \infty).$$

Similarly, the *diffusion-scaled service and offered work processes* are given by

$$\begin{aligned} \hat{S}^n(t) &:= \sqrt{n}\left(\bar{S}^n(t) - \mu t\right) \quad t \geq 0 \\ \hat{V}^n(t) &:= \sqrt{n}\left(\bar{V}^n(t) - \frac{1}{\mu} t\right) \quad t \geq 0. \end{aligned}$$

The following proposition presents the diffusion limits for these processes.

Proposition 2: As $n \to \infty$, in $(\mathcal{D}_{\lim}, U)$,

$$(\hat{A}^n, \hat{S}^n, \hat{V}^n) \Rightarrow \left(W^0 \circ F, \sigma\mu^{3/2} W \circ e, -\sigma\mu^{1/2} W \circ \frac{e}{\mu} \right) \quad (8)$$

where W^0 is the standard Brownian Bridge process and W is the standard Brownian motion process, both are independent processes. $e : [0, \infty) \to [0, \infty)$ is the identity map.

Remarks 1. The proof of this proposition follows easily from standard results: The FCLT limit for the diffusion-scaled arrival process, also called the empirical process, is a Brownian Bridge process by Donsker's Theorem (see Sections 13 and 16 in [3]). Note that this limit also arises in the study of the invariance principle associated with the Kolmogorov-Smirnov statistic used to compare empirical distributions with candidate ones (see [31] for more detail). The limits for the diffusion-scaled service and offered work processes follow from the FCLT for renewal processes (see Section 16 in [3] and Chapter 5 in [5]).

2. Note that we have not placed any restriction on the arrival distribution F other than that of finite support. The proofs of the fluid limits to the performance metrics will hold for arbitrary distribution functions F. However, the diffusion limits require F to be absolutely continuous, since the form of the limit process depends on this fact. Extending the result to arbitrary F appears non-trivial, and left for future work.

III. Fluid Approximations

We now derive the fluid limit to the queue length process for an arbitrary continuous arrival distribution and constant service rate. Recall the queue length process from (5). The corresponding *fluid-scaled queue length process* is

$$\frac{Q^n(t)}{n} = \frac{1}{n} A^n(t) - \frac{1}{n} S^n(B^n(t)), \quad (9)$$

where $B^n(t)$ is the fluid-scaled version of the busy time process (4) defined as

$$B^n(t) := \left(\int_0^t \mathbf{1}_{\{Q^n(s)>0\}} ds \right) \mathbf{1}_{\{t \geq 0\}}. \quad (10)$$

Now, we can write (9) by adding and subtracting the functions F and μB^n to obtain $\frac{Q^n(t)}{n} =$

$$\left(\frac{A^n(t)}{n} - F(t) \right) - \left(\frac{S^n(B^n(t))}{n} - \mu B^n(t) \right) + \left(F(t) - \mu B^n(t) \right).$$

Similarly, adding and subtracting the function $\mu t \mathbf{1}_{\{t \geq 0\}}$, we get $\frac{Q^n(t)}{n} := \mu I^n(t) +$

$$\left(\frac{A^n(t)}{n} - F(t) \right) - \left(\frac{S^n(B^n(t)}{n} - \mu B^n(t) \right) + \left(F(t) - \mu t \mathbf{1}_{\{t \geq 0\}} \right),$$

where $I^n(t) = t\mathbf{1}_{\{t \geq 0\}} - B^n(t)$ is the *fluid-scaled idle time process*. Now, we rewrite the fluid-scaled queue length process (9) as follows:

$$\frac{Q^n(t)}{n} = \bar{X}^n(t) + \mu I^n(t), \quad \forall t \in [-T_0, \infty), \quad (11)$$

where $\bar{X}^n(t)$ is defined to be $\bar{X}^n(t) :=$

$$\left(\frac{A^n(t)}{n} - F(t) \right) - \left(\frac{S^n(B^n(t))}{n} - \mu B^n(t) \right) + (F(t) - \mu t \mathbf{1}_{\{t \geq 0\}}). \quad (12)$$

Let $\bar{Q}^n(t) := Q^n(t)/n$ denote the fluid-scaled queue length process. Theorem 1 below proves two things about $\bar{Q}^n$. First, that it satisfies the Skorokhod reflection mapping theorem (see Chapter 6 of [5]), and can be expressed uniquely in terms of $\bar{X}^n$. Second, using this unique representation and the continuous mapping theorem, we obtain the the fluid limit for the queue length process.

Recall that the Skorokhod reflection map is a continuous functional $(\Phi, \Psi) : \mathcal{D}_{\lim} \to \mathcal{D}_{\lim} \times \mathcal{D}_{\lim}$ defined as

$$x \mapsto \Psi(x) := \sup_{-T_0 \leq s \leq t} (-x(s))_+,$$

and

$$x \mapsto \Phi(x) := x + \Psi(x), \quad \forall x \in \mathcal{D}_{\lim}.$$

Theorem 1 (Fluid Limit): The pair $(\bar{Q}^n, \mu I^n)$ has a unique representation $(\Phi(\bar{X}^n), \Psi(\bar{X}^n))$ in terms of $\bar{X}^n$. Furthermore, as $n \to \infty$,

$$(\bar{Q}^n, \mu I^n) \xrightarrow{a.s} (\Phi(\bar{X}), \Psi(\bar{X})) \text{ in } (\mathcal{D}_{\lim}, U),$$

where $\bar{X}(t) = (F(t) - \mu t \mathbf{1}_{\{t \geq 0\}})$.

Proof: First note that $\bar{Q}^n(t) \geq 0$, $\forall t \in [-T_0, \infty)$. It is also true that $I^n(-T_0) = 0$ and $dI^n(t) \geq 0$, $\forall t \in [-T_0, \infty)$. By definition of $I^n(t)$, it follows that $\int_{-T_0}^{\infty} \bar{Q}^n(t) dI^n(t) = 0$. Thus, by the Skorokhod reflection mapping theorem [25], [5], the joint process $(\bar{Q}^n(t), \mu I^n(t))$ has a unique reflection mapping representation in terms of $\bar{X}^n(t)$ as $(\Phi(\bar{X}^n), \Psi(\bar{X}^n))$.

Note that by definition of $B^n(t) \leq t$ and from Proposition 1, it follows that as $n \to \infty$,

$$\left| \frac{S^n \circ B^n}{n} - \mu B^n \right| \xrightarrow{a.s} 0 \text{ in } (\mathcal{D}_{\lim}, U). \quad (13)$$

Using (13) and Proposition 1 it follows that as $n \to \infty$

$$\bar{X}^n \xrightarrow{a.s} \bar{X} \text{ in } (\mathcal{D}_{\lim}, U),$$

where $\bar{X} := (F(t) - \mu t \mathbf{1}_{\{t \geq 0\}})$. Finally, it is easy to verify that $\Psi : \mathcal{D} \longrightarrow \mathcal{D}$ is a continuous operator. So, using the limit derived above and the Continuous Mapping Theorem (Theorem 5.2 of [5]) it follows that, as $n \to \infty$

$$(\bar{Q}^n, \mu I^n) = (\Phi(\bar{X}^n), \Psi(\bar{X}^n)) \xrightarrow{a.s} (\Phi(\bar{X}), \Psi(\bar{X})) \text{ in } (\mathcal{D}_{\lim}, U).$$

■

Remarks 1. Note that $\bar{X}$ is the difference between the fluid limits of the arrival and service processes, and is often referred to as the fluid limit of the *netput* process. In effect, it is the amount of net *potential* fluid inflow to the system.

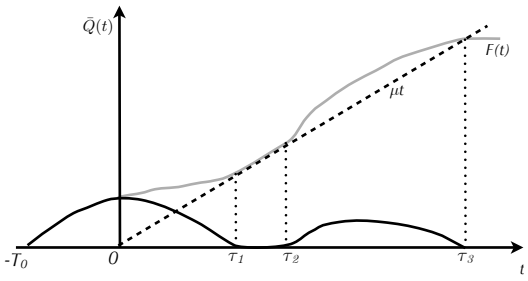


Fig. 1. An example of a $\Delta_{(i)}/GI/1$ queue that will undergo multiple "regime changes". The fluid queue length process is positive on $[-T_0, \tau_1)$ and $[\tau_2, \tau_3)$, and 0 on $[\tau_1, \tau_2)$ and $[\tau, \infty)$

2. Theorem 1 shows that the fluid limit of the queue length process for $t \in [-T_0, \infty)$ is

$$\bar{Q}(t) \;=\; (F(t) - \mu t \mathbf{1}_{\{t \geq 0\}}) + \sup_{-T_0 \leq s \leq t} (-(F(s) - \mu s \mathbf{1}_{\{s \geq 0\}}))_+.$$

$\bar{Q}$ can be interpreted as the sum of the fluid netput process and the potential amount of fluid lost from the system. Suppose that $(F(t) - \mu t \mathbf{1}_{\{t \geq 0\}}) < 0$ so that the fluid service process has "caught up" and exceeded the cumulative amount of fluid arrived in the system by time t (for simplicity assume $t > 0$). Further, suppose $f(t) - \mu < 0$, implying that the netput process is decreasing at t. In this case, $\sup_{-T_0 \leq s \leq t}(-(F(s) - \mu s \mathbf{1}_{\{s \geq 0\}}))_+ = -(F(t) - \mu t)$. This is the amount of extra fluid that could have been served, but is now lost.

From Theorem 1, we can see that the fluid limit of the fluid-scaled busy time process is

$$\bar{B}(t) \;:=\; t \mathbf{1}_{\{t \geq 0\}} - \frac{1}{\mu} \Psi(\bar{X}(t)),\; \forall t \in [-T_0, \infty) \qquad (14)$$

Note that $\bar{B}(t) = 0$ for all $t \leq 0$, as $\Psi(\bar{X})(t) = 0$ on that interval.

The limit for the busy time process will prove useful in establishing a fluid limit for the virtual waiting time process (3). The *fluid-scaled virtual waiting time process* is

$$Z^n(t) \quad = \quad V^n\left(n\left(\frac{A^n(t)}{n}\right)\right) - \; B^n(t) - t \mathbf{1}_{\{t \leq 0\}}. \qquad (15)$$

The following result establishes a fluid limit for the virtual waiting time process.

Proposition 3 (Fluid Transient Little's Law): As $n \to \infty$,

$$Z^n \xrightarrow{a.s} \bar{Z} \;\text{ in } (\mathcal{D}_{\lim}, U), \qquad (16)$$

where $\bar{Z}(t) := \bar{Q}(t)/\mu - t \mathbf{1}_{\{t \leq 0\}}$.

Remarks 1. The fluid limit in Corollary 3 can be interpreted as a *fluid 'transient' Little's Law* in the fluid limit for this model since it relates the virtual waiting time fluid queue length but there is a transient term, $t \mathbf{1}_{\{t \leq 0\}}$ which accounts for the fact that an arrival at time $t < 0$ would have to have $-t$ time units for service to start.

IV. Diffusion Approximations

A. Queue Length Process

Define the *diffusion-scaled queue length process* as

$$\frac{Q^n(t)}{\sqrt{n}} \;:=\; \frac{A^n(t)}{\sqrt{n}} - \frac{S^n(B^n(t))}{\sqrt{n}} \quad \forall t \in [-T_0, \infty) \qquad (17)$$

Rewriting it after introducing the term $\sqrt{n}\mu t \mathbf{1}_{\{t \geq 0\}}$, we have

$$\frac{Q^n(t)}{\sqrt{n}} = \left(\frac{A^n(t)}{\sqrt{n}} - \sqrt{n}F(t)\right) - \left(\frac{S^n(B^n(t))}{\sqrt{n}} - \sqrt{n}\mu B^n(t)\right) + \sqrt{n}(F(t) - \mu t \mathbf{1}_{\{t \geq 0\}}) + \sqrt{n}\mu(t \mathbf{1}_{\{t \geq 0\}} - B^n(t)).$$

Using the definition of the idle time process

$$\sqrt{n} I^n(t) \;=\; \sqrt{n}(t \mathbf{1}_{\{t \geq 0\}} - B^n(t)),$$

and we can express $Q^n/\sqrt{n}$ as

$$\frac{Q^n}{\sqrt{n}} = \hat{X}^n + \sqrt{n}\bar{X} + \sqrt{n}\mu I^n, \qquad (18)$$

where $\hat{X}^n(t) :=$

$$\left(\frac{A^n(t)}{\sqrt{n}} - \sqrt{n}F(t)\right) - \left(\frac{S^n(B^n(t))}{\sqrt{n}} - \sqrt{n}\mu B^n(t)\right)$$
$$= \quad \hat{A}^n(t) - \hat{S}^n(B^n(t)), \quad \forall t \in [-T_0, \infty). \qquad (19)$$

Recall from Theorem 1 that $\bar{X}(t) = (F(t) - \mu t \mathbf{1}_{t \geq 0})$ is the fluid netput process. We can think of $\hat{X}^n$ as a diffusion refinement of the netput process. Now, Lemma 1 gives a diffusion limit of $\hat{X}^n(t)$ as a direct consequence of Proposition 2.

Lemma 1: As $n \to \infty$,

$$\hat{X}^n \Rightarrow \hat{X} \;:=\; W^0 \circ F - \sigma \mu^{3/2} W \circ \bar{B} \;\text{ in } (\mathcal{D}_{\lim}, U) \qquad (20)$$

where $\bar{B}$ is defined in (14), and W^0 and W are independent standard Brownian Bridge and standard Brownian motion processes.

Proof: First note that $B^n(t) \leq t, \forall t \in [0, \infty)$, implying that $S^n \circ B^n \in \mathcal{D}$. Using Proposition 2, (14) and the random time change theorem (see Section 17 of [3]), it follows that as $n \to \infty$

$$\sqrt{n}\left(\frac{S^n \circ B^n}{n} - \mu B^n\right) \Rightarrow \sigma \mu^{3/2} W \circ \bar{B}. \qquad (21)$$

Now, it follows from Proposition 2 and the weak limit (21) that, as $n \to \infty$,

$$\hat{X}^n \Rightarrow \hat{X}(t) \;:=\; W^0 \circ F - \sigma \mu^{3/2} W \circ \bar{B}.$$

■

Remarks 1. Note that using a classical time change (see [14]) it is possible to see that the Brownian Bridge process is equal in distribution to a time changed Brownian motion process, and $\hat{X}$ is equal in distribution to a stochastic integral

$$\hat{X}(t) \stackrel{d}{=} \begin{cases} \int_{-T_0}^t \sqrt{g'(s)} d\tilde{W}_s, & \forall t \in [-T_0, T] \\ -\sigma \mu^{3/2} W(\bar{B}(T)), & \forall t > T. \end{cases} \qquad (22)$$

where $g(t) = F(t)(1 - F(t)) + \sigma^2 \mu^3 \bar{B}(t)$ and $\tilde{W}$ is a Brownian motion independent of W^0 and W. Thus, the process $\hat{X}$ can

also be interpreted as a time-changed Brownian motion, on the interval $[-T_0, T]$, and then its sample path is a constant on (T, ∞).

In the rest of this section, we will use Skorokhod's almost sure representation theorem [25], [30], and replace the random processes above that converge in distribution by those defined on a new probability space, that have the same distribution as the original processes and converge almost surely. Remarkably, the requirements for the almost sure representation are quite mild - the underlying topological space needs to be Polish (a separable and complete metric space). We note without proof that the space $\mathcal{D}_{\lim}$, as defined in this paper, is Polish when endowed with the M_1 topology. This conclusion follows from Theorem 2.6 of [28] and the fact that the proof there extends easily to the case of the M_1 topology. The authors in [18] also point out that [23] has a more general proof of this fact.

Thus, we replace the weak convergence in (8) by

$$(\hat{A}^n, \hat{S}^n, \hat{V}^n) \xrightarrow{a.s} \left(W^0 \circ F, \sigma\mu^{3/2} W, -\sigma\mu^{1/2} W \circ \frac{h}{\mu} \right)$$

in $(\mathcal{D}_{\lim}, U)$, where abusing notation we denote the new limit random processes by the same letters as the old ones. This implies that in Lemma 1, as $n \to \infty$, we actually have

$$\hat{X}^n \xrightarrow{a.s} \hat{X} \text{ in } (\mathcal{D}_{\lim}, U).$$

Now, our goal is to establish a functional central limit theorem for the centered queue length process

$$\hat{Q}^n(t) := \sqrt{n}\left(\frac{Q^n(t)}{n} - \bar{Q}(t) \right). \tag{23}$$

We achieve this by using the Skorokhod reflection mapping theorem [25], [5], [31] and express $(Q^n(t)/\sqrt{n}, \sqrt{n}\mu I^n(t))$ uniquely in terms of $\hat{X}^n$ and $\bar{X}$. Using this representation, we redefine $\hat{Q}^n$ in terms of $\hat{X}^n$ and $\bar{X}$, and then establish the necessary limit as $n \to \infty$. Note that we will establish the limit in the weaker topology M_1, as opposed to the more common U (uniform) or J_1 topologies. This is because a directional derivative reflection mapping lemma (Lemma 2) we will use is only available for $(\mathcal{D}_{\lim}, M_1)$. For convenience, we define the function

$$\tilde{Y}^n(t) := \sqrt{n}\mu I^n(t) - \sqrt{n}\Psi(\bar{X}(t)). \tag{24}$$

Recall that (Φ, Ψ) is the Skorokhod reflection map. We denote the directional derivative of the Skorokhod reflection map by

$$\nabla_t^{\bar{X}} = \{-T_0 \le s \le t | \bar{X}(s) = \Psi(\bar{X})(t)\}, \tag{25}$$

which is a set correspondence of points upto time t where the fluid netput process achieves an infimum.

The following theorem proves the diffusion limit for the queue length process.

Theorem 2 (Diffusion Limit): The pair $(\hat{Q}^n, \tilde{Y}^n)$ has a unique representation in terms of $\hat{X}^n$ and $\sqrt{n}\bar{X}$ given by

$$\left(\Phi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\bar{Q}, \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X}) \right),$$

where $\bar{Q} = \bar{X} + \Psi(\bar{X})$ is the fluid limit of the queue length process. Furthermore, as $n \to \infty$

$$(\hat{Q}^n, \tilde{Y}^n) \Rightarrow (\hat{X} + \tilde{Y}, \tilde{Y}) \text{ in } (\mathcal{D}_{\lim}, M_1),$$

where $\hat{X}(t) = W^0(F(t)) - \sigma\mu^{3/2} W(\bar{B}(t))$, and $\tilde{Y}(t) = \max_{s \in \nabla_t^{\bar{X}}}(-\hat{X}(s)) \ \forall t \in [-T_0, \infty)$.

The limit result follows by use of the following directional derivative reflection mapping lemma which is adapted from Lemma 5.2 in [18].

Lemma 2: Let x and y be real-valued continuous functions on $[0, \infty)$, and $\Psi(z)(t) = \sup_{0 \le s \le t}(-z(s))$, for any process $z \in \mathcal{D}_{\lim}$. Let $\{y_n\} \subset \mathcal{D}_{lim}$ be a sequence of functions such that $y_n \xrightarrow{a.s} y$ as $n \to \infty$. Then, with respect to Skorokhod's M_1 topology, $\tilde{y}_n := \Psi(\sqrt{n}x + y_n) - \sqrt{n}\Psi(x) \longrightarrow \tilde{y} := \sup_{s \in \nabla_t^x}(-y(s))$ as $n \to \infty$, where $\nabla_t^x = \{0 \le s \le t | x(s) = \bar{x}(t)\}$.

Proof: Rewrite $\tilde{y}_n$ as

$$\tilde{y}_n = (\Psi(\sqrt{n}x + y_n) - \Psi(\sqrt{n}x + y)) - (\Psi(\sqrt{n}x + y) - \sqrt{n}\Psi(x)).$$

Now, using the fact that the Skorokhod reflection map is Lipschitz continuous under the uniform metric (see Lemma 13.4.1 and Theorem 13.4.1 of [31]) we have

$$(\Psi(\sqrt{n}x + y_n) - \Psi(\sqrt{n}x + y)) \le \|y_n - y\|,$$

where $\|\cdot\|$ is the uniform metric. It follows that

$$\tilde{y}_n \le \|y_n - y\| + (\Psi(\sqrt{n}x + y) - \sqrt{n}\Psi(x)),$$

Now, by Lemma 5.2 of [18] we know that as $n \to \infty$

$$(\Psi(\sqrt{n}x + y) - \sqrt{n}\Psi(x)) \xrightarrow{a.s} \tilde{y}, \text{ in } (\mathcal{D}_{lim}, M_1).$$

Using this result, and the fact that by hypothesis y_n converges to y in $(\mathcal{D}_{lim}, U)$ we have, as $n \to \infty$,

$$\tilde{y}_n \xrightarrow{a.s} \tilde{y}, \text{ in } (\mathcal{D}_{lim}, M_1).$$

■

Proof: [Proof of Theorem 2] First using (18), it follows by the Skorokhod reflection mapping theorem that

$$\left(\frac{Q^n}{\sqrt{n}}, \sqrt{n}\mu I^n \right) = \left(\Phi(\hat{X}^n + \sqrt{n}\bar{X}), \Psi(\hat{X}^n + \sqrt{n}\bar{X}) \right). \tag{26}$$

This implies that

$$\hat{Q}^n = \frac{Q^n}{\sqrt{n}} - \sqrt{n}\bar{Q} = \Phi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\bar{Q}. \tag{27}$$

Recall from Theorem 1 that $\bar{Q} = \bar{X} + \Psi(\bar{X})$. Substituting this expression into (27), and using the fact that $\Phi(x) = x + \Psi(x)$, for any $x \in \mathcal{D}_{\lim}$, we have

$$\begin{aligned} \hat{Q}^n &= \hat{X}^n + \sqrt{n}\bar{X} + \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}(\bar{X} + \Psi(\bar{X})), \\ &= \hat{X}^n + \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X}). \end{aligned} \tag{28}$$

Similarly, utilizing the expression for $\sqrt{n}\mu I^n$ in (26) we have

$$\begin{aligned} \tilde{Y}^n &= \sqrt{n}\mu I^n - \sqrt{n}\Psi(\bar{X}), \\ &= \Psi(\hat{X}^n + \sqrt{n}\bar{X}) - \sqrt{n}\Psi(\bar{X}), \end{aligned} \tag{29}$$

implying that $\hat{Q}^n = \hat{X}^n + \tilde{Y}^n$.

Observe that $\tilde{Y}_n$ is exactly in the form of $\tilde{y}_n$ defined in Lemma 2 above. Since it has been shown that $\hat{X}^n$ converges uniformly on compact sets of $[-T_0, \infty)$ to the continuous process $\hat{X}$ in Lemma 1, applying Lemma 2 above it follows that, as $n \to \infty$,

$$\tilde{Y}_n \xrightarrow{a.s} \tilde{Y} := \max_{s \in \nabla_{\cdot}^{\bar{X}}} (-\hat{X}(s)) \text{ in } (\mathcal{D}_{\lim}, M_1)$$

Now, using Lemma 1 and the immediate result above we have

$$\hat{Q}^n = \hat{X}^n + \tilde{Y}^n \xrightarrow{a.s} \hat{X} + \max_{s \in \nabla_{\cdot}^{\bar{X}}} (-\hat{X}(s)) \text{ in } (D, M_1) \text{ as } n \to \infty.$$

This completes the proof. ∎

Remarks 1. Observe that the diffusion limit to the queue length process is a function of a Brownian Bridge and a Brownian motion process. This is significantly different from the usual limits obtained in a heavy-traffic or large population approximation to a single server queue. For instance, in the $G/G/1$ queue one would expect a reflected Brownian motion in the heavy-traffic setting. In [18] it was shown that the diffusion limit process to the $M_t/M_t/1$ queue is a time changed Brownian motion reflected through the directional derivative reflection map we used in Lemma 2. There are very few examples of heavy-traffic limits involving a diffusion that is a function of a Brownian Bridge and a Brownian motion process. However, there have been some results in non-conventional queueing models where a Brownian bridge arises in the limit. In [24], for instance, a Brownian Bridge limit arises in the study of a many-server queue in the Halfin-Whitt regime.

2. We noted in the remarks after Theorem 1 that the fluid limit can change between being positive and zero in the arrival interval for a completely general F. One can then expect the diffusion limit to change as well, and switch between being a 'free' diffusion, a reflected diffusion and a zero process. This is indeed the case. Figure 2 illustrates this for the example in Figure 1. Note that $\forall t \in [-T_0, \tau_1)$ $\Psi(\bar{X})(t) = -\bar{X}(-T_0)$, implying that the set $\nabla_t^{\bar{X}}$ is a singleton. On the other hand, at τ_1 $\nabla_t^{\bar{X}} = \{-T_0, \tau_1\}$. For $t \in (\tau_1, \tau_2]$, $\Psi(\bar{X})(t) = 0 = \bar{X}(t)$, implying that $\nabla_t^{\bar{X}} = (\tau_1, t]$. On (τ_2, τ_3), $\Psi(\bar{X})(t) = 0$, but $\bar{X}(t) > 0$, so that $\nabla_t^{\bar{X}} = (\tau_1, \tau_2]$. Finally, the fluid queue length becomes zero when the fluid service process exceeds the fluid arrival process in $[\tau_3, \infty)$, implying that $\Psi(\bar{X})(t) = -(F(t) - \mu t) > 0$. It can be seen that $\nabla_t^{\bar{X}} = \{t\}$ in this case.

1) Uniform Arrival Distribution: Note that $\nabla_t^{\bar{X}}$ is a set correspondence that maps each time t to the set of points (upto t) at which the fluid netput process is equal to its infimum at t. Theorem 2 shows that the diffusion limit to the queue length process is in fact piecewise continuous since $\tilde{Y}$ is. We now specialize the diffusion limit results to the case of a uniform F with early-bird arrivals. This illustrates with greater clarity the discontinuous nature of the limit processes.

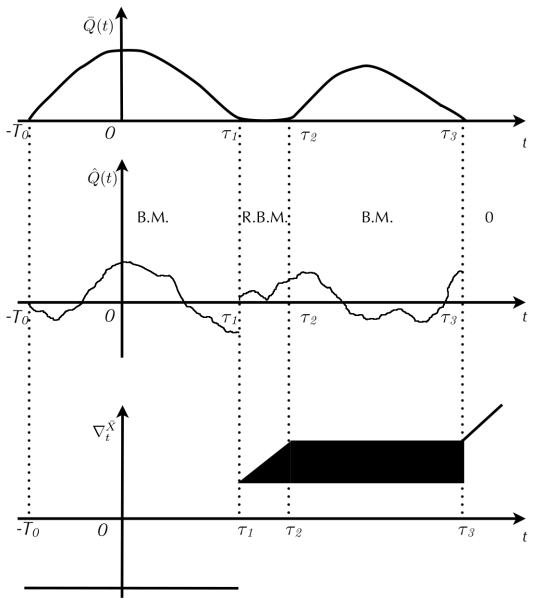


Fig. 2. An example of a $\Delta_{(i)}/GI/1$ queue that will undergo multiple "regime changes". The diffusion limit switches between a free Brownian motion (BM), a reflected Brownian motion (RBM), and the zero process

Corollary 1: Let F be the uniform distribution on $[-T_0, T]$, where $-T_0 < 0$. Then,

$$\hat{Q}(t) = \begin{cases} W^0(F(t)) - \sigma\mu^{\frac{3}{2}} W(t), & \forall t \in [-T_0, \tau) \\ (W^0(F(\tau)) - \sigma\mu^{\frac{3}{2}} W(\tau)) & \\ +(-(W^0(F(\tau)) - \sigma\mu^{\frac{3}{2}} W(\tau)))_+, & t = \tau \\ 0, & \forall t \in (\tau, \infty), \end{cases}$$

where $\tau = \{-T_0 \le t < \infty \,|\, F(t) = \mu t\}$.

The time τ can be interpreted as the first time that the fluid service process catches up with the fluid arrival process. For a uniform F there is at most one such point, but in general there can be many such points.

Interestingly, the nature of the discontinuity at $\hat{Q}(\tau)$ depends on the the sign of $\hat{X}(\tau)$. The following corollary clarifies this statement. Recall that t is a point of *right-discontinuity* for a function $x \in \mathcal{D}_{\lim}$ if x is left-continuous at t, and $x(t-) > x(t+)$. On the other hand, t is a point of *left-discontinuity* if x is right-continuous at t, and $x(t+) > x(t-)$.

Corollary 2: Let F be the uniform distribution function over $[-T_0, T]$, where $T_0 > 0$, and $\tau = \{-T_0 \le t < \infty | F(t) = \mu t \mathbf{1}_{\{t \ge 0\}}\}$. Then, for the process $\hat{Q}$ in Corollary 1, we have

(i) $[-T_0, \tau) \cup (\tau, \infty)$ are points of continuity.
(ii) τ is a point of right-discontinuity, when $\hat{X}(\tau) \ge 0$.
(iii) τ is point of left-discontinuity, when $\hat{X}(\tau) < 0$.

The proof is available in [11].

Remarks: 1. In Corollary 1, $\hat{Q}$ is piecewise continuous on $[-T_0, \infty)$, with a single point of discontinuity at τ. Interestingly, $\hat{Q}(\tau)$ is determined by the value of the process at $\tau-$. If $\hat{Q}(\tau-)$ is non-positive, then the value of $\hat{Q}(\tau)$ is 0. On the other hand, if $\hat{Q}(\tau-) = \hat{X}(\tau) > 0$, then $\hat{Q}(\tau) = \hat{X}(\tau) = \hat{Q}(\tau-)$. However, at $\tau+$ the queue length

immediately falls to 0 and remains there forever after, as the reflection regulator map becomes positive for all time that follows τ.

2. A useful way to interpret the discontinuity at τ in Corollary 1 is to consider the process on the two sub-intervals separately and try to "patch" them together. If $\hat{Q}(\tau-) = \hat{X}(\tau) = \hat{Q}(\tau) > 0$ we should expect a free diffusion path on the interval $[-T_0, \tau]$, and a reflected process such that the path is 0 on (τ, ∞). Further, $\hat{Q}(\tau)$ becomes the "starting state" for the process on the interval (τ, ∞), and the reflection operator is applied an instant after τ. On the other hand, if $\hat{Q}(\tau-) = \hat{X}(\tau-) \leq 0$ we have a free diffusion on $[-T_0, \tau)$ and the zero process on $[\tau, \infty)$; i.e., the process drops to zero at τ. Thus, $\hat{Q}(\tau-)$ provides the starting conditions for the new "regime" of the diffusion, as the process transitions from $[-T_0, \tau)$ to (τ, ∞).

3. We note that in [17], a diffusion approximation to the queue length process is derived independently for different operating regimes. However, these limit results have not been "patched" together to obtain a "process-level" convergence result, which is precisely where the mathematical challenges lie. We also note that diffusion limits in [17] do not involve directional derivative maps since the processes are continuous over the intervals on which they are studied.

4. It is also pertinent to mention that the limit results in [17] are obtained in the uniform topology at what are the continuity points of the limit process between regime changes. However, as we noted above, there are discontinuities in the limit process at points such as τ where regimes change, precluding the possibility of establishing a "process-level" limit in the uniform topology and necessitating the need to establish the limit in a weaker topology.

B. Virtual Waiting Time Process

Now, consider the centered virtual waiting time process given by

$$\hat{Z}^n(t) \, = \, \sqrt{n}(Z^n(t) - \bar{Z}(t)) \quad \forall t \in [-T_0, \infty), \tag{30}$$

where $\bar{Z}(t)$ is defined in (16) and $Z^n(t)$ is defined in (15). Corollary 4 proves the diffusion limit to this process.

Proposition 4 (Diff. transient Little's Law): As $n \to \infty$,

$$\hat{Z}^n \Rightarrow \hat{Z} := \frac{1}{\mu}\hat{Q} + \sigma\mu^{1/2} W \circ \bar{B} - \sigma\mu^{1/2} W \circ F \;\text{ in } (\mathcal{D}_{\lim}, M_1). \tag{31}$$

V. Sample-Path Analysis

As noted in Section IV, the limit process is piecewise continuous, with discontinuity points determined by the fluid limit. Indeed, the discontinuity points are precisely where the fluid limit switches regimes between 'overloaded', 'underloaded' and 'critically-loaded' states. We now provide formal definitions of these notions, in terms of the fluid limit arrival and service processes.

We then characterize the sample path of the queue length limit process, and the points at which it has discontinuities. Developments in this section follow the study of the directional derivative process in [18]. However, the limit processes and the setting of our models are completely different. Thus, where necessary, we prove some of the facts about the sample paths. The operating regimes for the $M_t/M_t/1$ model in [18] and our $\Delta_{(i)}/GI/1$ model are quite similar, and we adapt the definitions to our model.

A. Regimes of $\bar{Q}$

It can be useful to characterize the state of a queue in terms of a "traffic intensity" measure. For instance, in the case of a $G/G/1$ queue, the traffic intensity is the ratio of the arrival rate to the service rate. In [21] a traffic intensity function for the $M_t/M_t/1$ queue with arrival rate $\lambda(s)$ and service rate $\mu(s)$ was introduced as the ratio

$$\rho^*(t) := \sup_{0 \leq s \leq t} \frac{\int_s^t \lambda(u)du}{\int_s^t \mu(u)du}, \; t > 0.$$

Here, we adapt the form of this function and define the traffic intensity for the $\Delta_{(i)}/GI/1$ queue in terms of the fluid limit as

$$\rho(t) = \begin{cases} \infty, & \forall t \in [-T_0, 0] \\ \sup_{0 \leq r \leq t} \frac{F(t)-F(r)}{\mu(t-r)}, & \forall t \in [0, \tilde{T}] \\ 0, & \forall t > \tilde{T}, \end{cases} \tag{32}$$

where $\tilde{T} := \inf\{t > 0 | F(t) = 1 \text{ and } \bar{Q}(t) = 0\}$. Note that we define the traffic intensity to be ∞ in the interval $[-T_0, 0]$ as there is no service, but there can be fluid arrivals. It is also important to note that the definition of ρ^* follows from the pre-limit system describing the arrival and service processes in the $M_t/M_t/1$ queue, whereas the definition of ρ is contingent on the establishment of the fluid limit processes as there is no explicit arrival 'rate' associated with the arrival process. In the case of a uniform F over an interval $[-T_0, T]$ ρ is given by

$$\rho(t) = \frac{t \wedge T}{t} \frac{1}{\mu(T + T_0)}, \; \forall \; t \in [0, \tilde{T}].$$

Now, consider the following obvious definitions of the some of the operating states of the fluid $\Delta_{(i)}/GI/1$ queue.

Definition 1 (Operating regimes.): The $\Delta_{(i)}/GI/1$ queue is

*(i)*overloaded if $\rho(t) > 1$.

(ii) critically loaded if $\rho(t) = 1$.

(iii) underloaded if $\rho(t) < 1$.

Notice that these regimes correspond to the operating regimes of a time homogeneous $G/G/1$ queue. However, since the queue length fluid limit in the $\Delta_{(i)}/GI/1$ queue can also vary with time, and analogous to the $M_t/M_t/1$ queue in [18] we also identify the following "finer" operating states. In particular, these states are useful in studying the approximation to the distribution of queue length process on local time scales.

Definition 2 (Operating states.): The $\Delta_{(i)}/GI/1$ queue is at

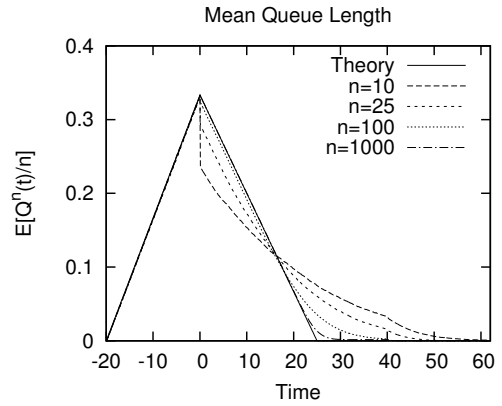
(i) end of overloading at time t if $\rho(t) = 1$ and there exists an open interval (a, t) or (t, a) such that $\rho(r) > 1$ for all r in that interval.

(ii) onset of critical loading at time t if $\rho(t) = 1$ and there exists a sequence $\lambda_n \uparrow t$ such that $\rho(\lambda_n) < 1$ for all n.
(iii) end of critical loading at time t if $\rho(t) = 1$, and there exists a sequence $\lambda_n \uparrow t$ such that $\rho(\lambda_n) = 1$ for all n and a sequence $\gamma_n \downarrow t$ such that $\rho(\gamma_n) < 1$ for all n.
(iv) middle of critical loading at time t if $\rho(t) = 1$, and t is in an open interval (a, b), such that $\sup_{t\in(a,b)} \rho(t) \geq 1$ and there exists a sequence $\lambda_n \uparrow t$ such that $\rho(\lambda_n) = 1$ for all n.

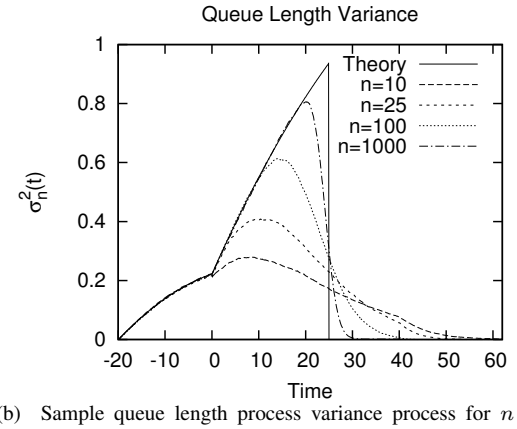
The following Lemma shows the equivalence of the definitions of the operating regimes to the process $\bar{Q}$.

Lemma 3: The $\Delta_{(i)}/GI/1$ queue is
(i) overloaded at time t if $\bar{Q}(t) > 0$.
(ii) critically loaded at time t if $\bar{Q}(t) = 0$, $\bar{X}(t) = \Psi(\bar{X})(t)$ and there exists an $r < t$ such that $\Psi(\bar{X})(t) = \Psi(\bar{X})(s)$ for all $s \in [r, t]$.
(iii) underloaded at time t if $\bar{Q} = 0$, $\bar{X}(t) = \Psi(\bar{X})(t)$ and there exists an $r < t$ such that $\Psi(\bar{X})(t) > \Psi(\bar{X})(s)$ for all $s \in (r, t)$.
The proof of the lemma is available in [11].

VI. Simulations



(a) Sample queue length process mean for $n = 10, 25, 100, 1000$, averaged over 10000 simulation runs.



(b) Sample queue length process variance process for $n = 10, 25, 100, 1000$, averaged over 10000 simulation runs.

Fig. 3. Mean and variance envelopes for F uniform over $[-20, 40]$, and exponentially distributed service times with rate $\mu = 0.03$

We now present some simulation results to illustrate the validity of the approximations in Theorems 1-2 as the population size increases. Consider a uniform arrival distribution over the interval $[-20, 40]$, with service times i.i.d. and exponentially distributed with parameter $\mu = 0.03$. We simulated the $\Delta_{(i)}/GI/1$ queue with population sizes $n = 10, 25, 100, 1000$.

It is straightforward to show that the variance of the diffusion limit process at time t is given by

$$\sigma^2(t) = \begin{cases} F(t)(1-F(t)), & \forall t \in [-T_0, 0] \\ F(t)(1-F(t)) + \sigma^2\mu^3 t, & \forall t \in (0, \tau) \\ 0, & \forall t > \tau. \end{cases}$$

Here, $F(t) = \frac{t+20}{60}$, $1/\sigma = \mu = 0.03$ and $\tau = 25$.

Observe from Figure 3(a) that even for small n, the sample mean is quite close to the fluid limit for $t < 0$. However, once queueing dynamics come into play, the fluid limit is a good approximation only for $n = 100$ or larger. Similarly, Figure 3(b) shows that the diffusion limit is a good approximation to population sizes of around $n = 1000$.

VII. Conclusions and Future Work

In this paper, we introduce a novel single server queueing model which we call the $\Delta_{(i)}/GI/1$ queue. In this model, customers from a finite population independently choose (sample) their time of arrival at the queue from a common distribution function. The arrival times are, thus, order statistics, and the inter-arrival times are differences of consecutive order statistics, and thus not renewal intervals. Service times are i.i.d. with some general distribution G, and the service rate is fixed.

Our original motivation for introducing the $\Delta_{(i)}/GI/1$ model came from the 'concert arrival game', a game of arrival timing introduced in [13]. Customers choose to arrive at a queue to minimize an expected cost functional that depends on the waiting time and the number of people who have already arrived. The Nash equilibrium analysis was done in the fluid limit, and it was established that for linear cost functionals, the uniform arrival distribution is a Nash equilibrium. With a given arrival distribution, this is a $\Delta_{(i)}/GI/1$ queueing model, and the details of the queueing dynamics in the game are also of interest. The second question is whether the equilibrium derived from the fluid model approximates in any way the equilibrium of the finite population 'concert arrival game'. Note that this is in the spirit of mean field equilibrium approximations of finite stochastic game models [4], [1], [27].

The new queue model we introduce should also be of interest in other scenarios. Thus, another thing we plan to do in the future is to acquire data for some common queueing situations, for example, queues at lunch-time cafeterias, postal and DMV offices, some enterprise call centers, and check if the $\Delta_{(i)}/GI/1$ queue could be a reasonable model.

References

[1] S. Adlakha and J. R. Mean field equilibrium in dynamic games with strategic complementarities. submitted, 2010.
[2] N. Bambos and J. Walrand. On queues with periodic inputs. *J. of Applied Probability*, pages 381–389, 1989.

[3] P. Billingsley. *Convergence of Probability Measures*. Wiley & Sons, 1968.

[4] A. L. Bodoh-Creed. Approximation of Large Dynamic Games. *Working paper*, 2012.

[5] H. Chen and D. Yao. *Fundamentals of Queueing Networks: Performance, asymptotics, and optimization*. Springer, 2001.

[6] R. Durrett. *Probability: Theory and Examples, 4th Ed.* Cambridge University Press, 2010.

[7] B. Hajek. A Queue with Periodic Arrivals and Constant Service Rate. *Probability, Statistics and Optimization - A Tribute To Peter Whittle*, pages 147–157, 1994.

[8] R. W. Hall. *Queueing Methods: For Services and Manufacturing*. Prentice Hall, 1990.

[9] J. Harrison. *Brownian motion and stochastic flow systems*. J. Wiley, 1985.

[10] H. Honnappa and R. Jain. Strategic Arrivals into Queueing Networks: The Network Concert Queueing Game. Submitted, 2011.

[11] H. Honnappa, R. Jain, and A. Ward. $\Delta_{(i)}/GI/1$:A New Queueing Model For Transitory Queueing Systems. Submitted, 2012.

[12] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, I. *Adv. in Applied Probability*, 2:150–177, 1970.

[13] R. Jain, S. Juneja, and N. Shimkin. The Concert Queueing Game: To Wait or To be Late. *Discrete Event Dynamic Systems*, 21(1):103–134, 2011.

[14] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 1991.

[15] J. Keller. Time-dependent queues. *SIAM Review*, pages 401–412, 1982.

[16] D. G. Kendall. Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of Imbedded Markov Chain. *Annals of Mathematical Statistics*, 24(3):338–354, 1953.

[17] G. Louchard. Large finite population queueing systems. The single-server model. *Stochastic Processes and their Applications*, 53(1):117 – 145, 1994.

[18] A. Mandelbaum and W. Massey. Strong Approximations For Time-dependent Queues. *Math. of Operations Research*, 20(1), 1995.

[19] A. Mandelbaum and G. Pats. State-dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits. *Annals of Applied Probability*, 8(2):569–646, 1998.

[20] W. Massey. Asymptotic analysis of the time dependent M/M/1 queue. *Math. of operations research*, pages 305–327, 1985.

[21] W. A. Massey. Non-Stationary Queues. *Ph.D. Dissertation, Stanford University*, 1981.

[22] G. Newell. Queues with time-dependent arrival rates I, II and III. *J. of Applied Probability*, 5:436–451 (I); 436–451 (II); 591–606 (III), 1968.

[23] J. L. Pomarede. A Unified Approach via Graphs to Skorohod's Topologies on the Function Space D. *Ph.D. Thesis, Yale University*, 1976.

[24] A. Puhalskii and J. Reed. On many-server queues in heavy traffic. *Annals of Applied Probability*, 20(1):129–195, 2010.

[25] A. Skorokhod. Limit Theorems For Stochastic Processes. *Theory of Probability And Its Applications*, 1(3), 1956.

[26] A. Skorokhod. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability and its Applications*, 6:264, 1961.

[27] G. Y. Weintraub and B. Van Roy. Industry dynamics: Foundations for models with an infinite number of firms. submitted, 2010.

[28] W. Whitt. Some Useful Functions for Functional Limit Theorems. *Math. of Operations Research*, 5(1):67–85, 1980.

[29] W. Whitt. An overview of Brownian and non-Brownian FCLTs for the single-server queue. *Queueing Systems*, 36:39–70, 2000.

[30] W. Whitt. *Internet Supplement To Stochastic Process Limits*. 2001.

[31] W. Whitt. *Stochastic Process Limits*. Springer, 2001.