

Many-server diffusion limits for $G/Ph/n + GI$ queues

J. G. Dai,¹ Shuangchi He² and Tolga Tezcan³

First draft: December 10, 2008

Current version: November 11, 2009

Abstract

This paper studies many-server limits for multi-server queues that have a phase-type service time distribution and allow for customer abandonment. The first set of limit theorems is for critically loaded $G/Ph/n + GI$ queues, where the patience times are independent, identically distributed following a general distribution. The next limit theorem is for overloaded $G/Ph/n + M$ queues, where the patience time distribution is restricted to be exponential. We prove that a pair of diffusion-scaled total-customer-count and server-allocation processes, properly centered, converges in distribution to a continuous Markov process as the number of servers n goes to infinity. In the overloaded case, the limit is a multi-dimensional diffusion process, and in the critically loaded case, the limit is a simple transformation of a diffusion process. When the queues are critically loaded, our diffusion limit generalizes the result by Puhalskii and Reiman (2000) for $GI/Ph/n$ queues without customer abandonment. When the queues are overloaded, the diffusion limit provides a refinement to a fluid limit and it generalizes a result by Whitt (2004) for $M/M/n + M$ queues with an exponential service time distribution. The proof techniques employed in this paper are innovative. First, a perturbed system is shown to be equivalent to the original system. Next, two maps are employed in both fluid and diffusion scalings. These maps allow one to prove the limit theorems by applying the standard continuous-mapping theorem and the standard random-time-change theorem.

Keywords: multi-server queues, customer abandonment, many-server heavy traffic, Halfin-Whitt regime, quality- and efficiency-driven regime, efficiency-driven regime, phase-type distribution.

AMS Subject Classification (2000): 90B20, 68M20, 60J70

1 Introduction

This paper studies many-server limits for multi-server queues that allow for customer abandonment. These queues are assumed to have a phase-type service time distribution. We

¹H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, Email: dai@gatech.edu

²H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, Email: heshuangchi@gatech.edu

³Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, Email: ttezcan@uiuc.edu

consider two separate parameter regimes: one for critically loaded many-server queues and the other for overloaded many-server queues.

As argued in the seminal paper of Halfin and Whitt (1981), for a critically loaded many-server queue, the system provides high-quality service and at the same time achieves high server utilization. Thus, the critically loaded parameter regime is also known as the Quality- and Efficiency-Driven (QED) limiting regime or the Halfin-Whitt limiting regime. For the overloaded $M/M/n + M$ model, Whitt (2004) demonstrates that a certain fluid approximation can be useful in predicting the *steady-state* performance of the multi-server system. He further demonstrates that a diffusion limit provides a refined approximation.

Our first set of results is for critically loaded $G/Ph/n + GI$ queues, whose patience times are independent, identically distributed (iid) following a general distribution. In Theorem 1, we prove that a pair of diffusion-scaled *total-customer-count* and *server-allocation processes* converges in distribution to a continuous Markov process (\tilde{X}, \tilde{Z}) . In Theorem 2, we prove that the diffusion-scaled *customer-count-vector process* converges to a diffusion process \tilde{Y} . In Theorem 3, the diffusion-scaled *virtual waiting time process* converges in distribution to a constant multiple of $(\tilde{X})^+$, which serves as the limit of the diffusion-scaled queue-length process. Our second result is for overloaded $G/Ph/n + M$ queues, whose patience time distribution is restricted to be exponential. In Theorem 4, we prove that the pair of diffusion-scaled *total-customer-count* and *server-allocation processes* converges in distribution to a diffusion process. Although the limit (\tilde{X}, \tilde{Z}) in Theorem 1 is not a diffusion process in a strict sense (see discussions below the statement of Theorem 2), we still call it a diffusion limit because it is a simple transformation of the diffusion process \tilde{Y} in Theorem 2. This terminology is consistent with the usage in conventional heavy traffic, where the limit process is often a constrained diffusion process; see, for example, Reiman (1984).

In the critically-loaded regime, Halfin and Whitt (1981) is the first paper to establish a diffusion limit for the $GI/M/n$ model. Puhalskii and Reiman (2000) establishes a diffusion limit for the $GI/Ph/n$ model, where the service time distribution is phase-type. Garnett et al. (2002) proves a diffusion limit for the $M/M/n + M$ model, which allows for customer abandonment. Whitt (2005) generalizes the result to the $G/M/n + M$ model. In the same paper, Whitt proves a stochastic-process limit for the $G/H_2^*/n$ model; this limit is not a diffusion process although a simple transformation of it is a diffusion process. Our first set of results, Theorems 1 and 2, extends the result of Puhalskii and Reiman (2000) to the $G/Ph/n + GI$ model, which allows for customer abandonment with a general patience time distribution. It also extends the work of Garnett et al. (2002) and Whitt (2005) to allow for phase-type service time distributions. For the overloaded $G/Ph/n + M$ model, Theorem 4 generalizes Whitt (2005) to the $G/Ph/n + M$ model. The diffusion limit in the theorem provides a refinement to a fluid limit.

In addition to these limit theorems, the techniques used in the proofs are innovative. When the patience time distribution is exponential, we first establish a sample-path representation for our $G/Ph/n + M$ model. This representation allows us to obtain the total-

customer-count and the server-allocation processes as a map of primitive processes with a random time change. These primitive processes are either assumed or known to satisfy functional central limit theorems (FCLTs). Therefore, our limits follow from the standard continuous-mapping theorem and the standard random-time-change theorem; see, for examples, Ethier and Kurtz (1986) and Billingsley (1999). When the queues are critically loaded, a result of Dai and He (2009) proves that the performance of these queues is not sensitive to the distribution of patience times, thus allowing us to prove Theorems 1–3 for general patience time distributions with a mild regularity condition.

Halfin and Whitt (1981), Garnett et al. (2002) and Whitt (2004) all use Stone’s theorem to prove diffusion limit theorems in the critically-loaded regime. Stone (1963) is set up for convergence of Markov chains to a diffusion process. This setting makes the generalization to non-renewal arrival processes difficult. Puhalskii and Reiman (2000) also uses the continuous-mapping approach for the $GI/Ph/n$ model. They employ a different sample-path representation for the total-customer-count process. Their representation requires them to use extensively *martingale FCLTs* in their proofs, whereas our approach uses standard FCLTs for random walks and Poisson processes. Pang et al. (2007) reviews a number of sample-path representations and martingale proofs for many-server heavy traffic limits, and Whitt (2007) surveys the proof techniques for establishing martingale FCLTs. Our proofs show that for multi-server queues with a phase-type service time distribution and an exponential patience time distribution, there is a general approach to proving limit theorems, without employing martingale FCLTs. Note that when the patience time distribution is general, our proofs for the diffusion limits in the critically-loaded regime rely on a result of Dai and He (2009), which is proved by using a martingale FCLT.

Our sample-path representation is based on the equivalence of our multi-server system to a perturbed system as illustrated in Tezcan (2006). This representation has been used in Dai and Tezcan (2005), Tezcan and Dai (2008) and Dai and Tezcan (2008) for multi-server-pool systems when service and patience times have exponential distributions. The sample-path argument has been explored previously in the setting of Markovian networks in Mandelbaum et al. (1998) for strong approximations and Mandelbaum and Pats (1998) for general state-dependent networks.

In our continuous-mapping approach, we have heavily exploited some maps from \mathbb{D}^{K+1} to \mathbb{D}^{K+1} , where K is the number of phases in the service time distribution. Variants of these maps have been employed in the literature; see, for examples, Mandelbaum et al. (1998), Reed (2007), Tezcan and Dai (2008), Dai and Tezcan (2008) and Pang et al. (2007). We use these maps not just in diffusion scaling but also in fluid scaling. Using a map twice, one for each scaling, allows us to obtain diffusion limits as a simple consequence of the standard continuous-mapping theorem and the random-time-change theorem. In the seminal paper, Reiman (1984) proves a *conventional* heavy traffic limit theorem for generalized Jackson networks. Our approach resembles the work of Johnson (1983), which also uses a multi-dimensional Skorohod map twice and provides a significant simplification of Reiman’s original proof.

For the $G/GI/n$ model, Reed (2007) proves a many-server limit for the total-customer-count process in the critically-loaded regime; his assumption on the service time distribution is completely general, and his limit is not a Markov process. This work is generalized in Mandelbaum and Momčilović (2009) to allow for customer abandonment. For the overloaded $G/GI/n$ model, Puhalskii and Reed (2008) proves a finite-dimensional-distribution limit for the total-customer-count process. Jelenković et al. (2004) proves a limit theorem for the $GI/D/n$ model. Gamarnik and Momčilović (2008) studies a many-server limit of the steady-state distribution of the $GI/GI/n$ model, where the service times are lattice-valued with a finite support. When the service time distribution is general, measure-valued processes have been used to give a Markovian description of the system. Kaspi and Ramanan (2007) obtains a measure-valued fluid limit for the $G/GI/n$ model. Kang and Ramanan (2008) obtains a measure-valued fluid limit for the $G/GI/n + G$ model with customer abandonment, and Zhang (2009) obtains a similar measure-valued fluid limit independently. Their work partially justifies the fluid model in Whitt (2006).

The remainder of the paper is organized as follows. In Section 2, we introduce the $G/GI/n + GI$ model, in an asymptotic framework, and phase-type distributions. The main results, Theorems 1–4, are stated in Section 3; a roadmap for the proofs is introduced in Section 3.3. In Section 4 we introduce a perturbed system that is equivalent to a $G/Ph/n + M$ queue and derive the dynamical equations that the perturbed system must obey. The proofs for the diffusion limits of $G/Ph/n + M$ queues, in both the critically-loaded and the overloaded regimes, are given in Section 5. Section 6 is dedicated to the proof for the diffusion limit of $G/Ph/n + GI$ queues in the critically-loaded regime. In Appendix A, we introduce a continuous map and establish various properties for the map. The state-space-collapse lemma is proved in Appendix B, in which Theorem 3 and Lemma 3 are also proved.

Notation

All random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ unless otherwise specified. The symbols \mathbb{Z} , \mathbb{Z}_+ , \mathbb{N} , \mathbb{R} and \mathbb{R}_+ are used to denote the sets of integers, nonnegative integers, positive integers, real numbers and nonnegative real numbers, respectively. For $d \in \mathbb{N}$, \mathbb{R}^d denotes the d -dimensional Euclidean space; thus, $\mathbb{R} = \mathbb{R}^1$. The space of functions $f : \mathbb{R}_+ \rightarrow \mathbb{R}^d$ that are right-continuous on $[0, \infty)$ and have left limits in $(0, \infty)$ is denoted by $\mathbb{D}(\mathbb{R}_+, \mathbb{R}^d)$ or simply \mathbb{D}^d ; similarly, with $T > 0$, the space of functions $f : [0, T] \rightarrow \mathbb{R}^d$ that are right-continuous on $[0, T)$ and have left limits in $(0, T]$ is denoted by $\mathbb{D}([0, T], \mathbb{R}^d)$. For $f \in \mathbb{D}^d$, $f(t-)$ denotes its left limit at $t > 0$. For a sequence of random elements $\{X^n, n \in \mathbb{N}\}$ taking values in a metric space, we write $X^n \Rightarrow X$ to denote the convergence of X^n to X in distribution. Each stochastic process whose sample paths are in \mathbb{D}^d is considered to be a \mathbb{D}^d -valued random element. The space \mathbb{D}^d is assumed to be endowed with the Skorohod J_1 -topology (see Ethier and Kurtz (1986) or Billingsley (1999)). Given $x \in \mathbb{R}$, we set $x^+ = \max\{x, 0\}$, $x^- = \max\{-x, 0\}$

and $\lfloor x \rfloor = \max\{j \in \mathbb{Z} : j \leq x\}$. All vectors are envisioned as column vectors. For a K -dimensional vector $u \in \mathbb{R}^K$, we use u_k to denote its k th entry and $\text{diag}(u)$ for the $K \times K$ diagonal matrix with k th diagonal entry u_k ; we put $|u| = \max_{1 \leq k \leq K} |u_k|$. For a matrix M , M' denotes its transpose and $M_{j,k}$ denotes its (j, k) th entry. We reserve I for the $K \times K$ identity matrix, e for the K -dimensional vector of ones, and e^k for the K -dimensional vector with k th entry one and all other entries zero.

2 A $G/Ph/n + GI$ queue

We first introduce a $G/GI/n + GI$ queue in Section 2.1. We then define a $G/Ph/n + GI$ queue by restricting the service time distribution to be phase-type. Phase-type distributions are defined in Section 2.2.

2.1 A $G/GI/n + GI$ queue

A $G/GI/n$ queue is a classical stochastic system that has been extensively studied in the literature; see, for examples, Kiefer and Wolfowitz (1955), Borovkov (1967) and Iglehart and Whitt (1970), among others. In such a system, there are n identical servers. The service times $\{v_i, i \in \mathbb{N}\}$ are a sequence of iid random variables, where v_i is the service time of the i th customer entering service after time 0. The service time distribution is general (the GI in the $G/GI/n$ notation), although for the rest of this paper we restrict it to be a phase-type distribution. The arrival process $E = \{E(t), t \geq 0\}$ is assumed to be general (the first G), where $E(t)$ denotes the number of customer arrivals to the system by time t . Upon his arrival to the system, a customer gets into service immediately if there is an idle server; otherwise, he waits in a *waiting buffer* that holds a first-in-first-out (FIFO) queue. The buffer size is assumed to be infinite. When a server finishes a service, the server removes the leading customer from the waiting buffer and starts to serve the customer; when the queue is empty, the server begins to idle. In our model, each customer has a patience time: when a customer's waiting time in queue exceeds his patience time, the customer abandons the system without any service. Retrial is not modeled in this paper. We assume that the patience times of customers who arrive after time 0, form a sequence of iid random variables that have a general distribution. Thus our model is a $G/GI/n + GI$ queue, where $+GI$ signifies the general patience time distribution. When the patience times are iid following an exponential distribution, the resulting system is a $G/GI/n + M$ queue.

We focus on systems when the arrival rate is high. Specifically, we consider a sequence of $G/GI/n + GI$ systems indexed by n , with E^n being the arrival process of the n th system. We assume that for the n th system, the arrival rate $\lambda^n \rightarrow \infty$ as the number of servers $n \rightarrow \infty$, while the service time and the patience time distributions do not change

with n . We further assume that

$$\lim_{n \rightarrow \infty} \frac{\lambda^n}{n} = \lambda > 0, \quad \lim_{n \rightarrow \infty} \sqrt{n} \left(\lambda - \frac{\lambda^n}{n} \right) = \beta \mu \quad \text{for some } \beta \in \mathbb{R}, \quad (2.1)$$

and

$$\tilde{E}^n \Rightarrow \tilde{E} \quad \text{as } n \rightarrow \infty, \quad (2.2)$$

where \tilde{E} is a Brownian motion and

$$\tilde{E}^n(t) = \frac{1}{\sqrt{n}} \hat{E}^n(t), \quad \hat{E}^n(t) = E^n(t) - \lambda^n t \quad \text{for } t \geq 0. \quad (2.3)$$

We use m to denote the mean service time; thus $\mu = 1/m$ is the mean service rate. For future purposes, let

$$\rho^n = \frac{\lambda^n}{n\mu} \quad \text{and} \quad \rho = \frac{\lambda}{\mu}. \quad (2.4)$$

Because customer abandonment is allowed, it is not necessary to assume $\rho^n < 1$ or $\rho \leq 1$. Condition (2.1) implies that

$$\lim_{n \rightarrow \infty} \sqrt{n}(\rho - \rho^n) = \beta.$$

When $\rho = 1$, the sequence of systems is critically loaded in the limit, and is said to be in the Quality- and Efficiency-Driven (QED) regime or the Halfin-Whitt regime. When $\rho > 1$, the sequence of systems is overloaded, and is said to be in the Efficiency-Driven (ED) regime. Our focus is both the QED and the ED regimes.

2.2 Phase-type distributions

In this section, we introduce phase-type distributions. Such a distribution is assumed to have $K \geq 1$ phases. The set of phases is assumed to be $\mathcal{K} = \{1, \dots, K\}$. Each phase-type distribution has a set of parameters p , ν and P , where p is a K -dimensional vector of nonnegative entries whose sum is equal to one, ν is a K -dimensional vector of positive entries, and P is a $K \times K$ sub-stochastic matrix. We assume that the diagonals of P are zero, namely,

$$P_{ii} = 0 \quad \text{for } i = 1, \dots, K, \quad (2.5)$$

and P is transient, namely,

$$I - P \text{ is invertible.} \quad (2.6)$$

A (continuous) phase-type random variable v is defined as the time until absorption in an absorbing state of a continuous-time Markov chain. With p , ν and P , the continuous-time Markov chain can be described as follows. It has $K + 1$ states, $1, \dots, K, K + 1$, with state $K + 1$ being absorbing. The rate matrix (or generator) of the Markov chain is

$$G = \begin{pmatrix} F & h \\ 0 & 0 \end{pmatrix},$$

where $F = \text{diag}(\nu)(P - I)$ is a $K \times K$ matrix and $h = -Fe$ is a K -dimensional vector.

Definition 1. A continuous phase-type random variable v with parameters p , ν and P , denoted as $v \sim \text{Ph}(p, \nu, P)$, is defined to be the first time until the continuous-time Markov chain with initial distribution p and rate matrix G reaches state $K + 1$.

Given condition (2.5), the rate matrix G and (ν, P) are uniquely determined from each other. It is well known (see, for example, Latouche and Ramaswami (1999)) that

$$\mathbb{P}[v \leq x] = 1 - p' \exp(Fx)e \quad \text{for } x \geq 0.$$

Because parameters p , ν and P uniquely determine a phase-type distribution, we free symbols F and G so that they can be reused in the rest of the paper.

For our purposes, we provide an alternative way to sample a $\text{Ph}(p, \nu, P)$ random variable. We first sample a sequence of phases k_1, \dots, k_L in $\mathcal{K} = \{1, \dots, K\}$ as follows. We sample phase k_1 following distribution p on \mathcal{K} . Assume $k_1, \dots, k_i \in \mathcal{K}$ have been sampled. Setting $j = k_i$, sample a phase from $\{1, \dots, K, K + 1\}$ following a distribution that is determined by the j th row of P ; the probability of getting phase $K + 1$ is $1 - \sum_{\ell=1}^K P_{j\ell} \geq 0$. The resulting phase is denoted by k_{i+1} . When $k_{i+1} = K + 1$, terminate the process and set $L = i$; otherwise, continue the sampling process. Because the matrix P is assumed to be transient, one has $L < \infty$ almost surely. Let ξ_1, \dots, ξ_L be independently sampled from exponential distributions with respective rates $\nu_{k_1}, \dots, \nu_{k_L}$. Then

$$v = \sum_{i=1}^L \xi_i. \tag{2.7}$$

3 Main results

In this section, we present two sets of results. The first set, presented in Section 3.1, is for critically loaded $G/Ph/n + GI$ queues. The second set, presented in Section 3.2, is for overloaded $G/Ph/n + M$ queues. A roadmap for proving these results is given in Section 3.3.

We consider a sequence of $G/Ph/n + GI$ queues, indexed by the number of servers n , which satisfies condition (2.1). We assume that the service times follow a phase-type distribution $\text{Ph}(p, \nu, P)$. Each customer's service time can be decomposed into a number of phases as in (2.7). When a customer is in service, it must be in one of the K phases of service. Let $Z_k^n(t)$ denote the number of customers *in phase k service* in the n th system at time t ; service times in phase k are exponentially distributed with rate ν_k . We use $Z^n(t)$ to denote the corresponding K -dimensional vector. We call $Z^n = \{Z^n(t), t \geq 0\}$ the *server-allocation process*. Let $N^n(t)$ denote the number of customers in the n th system at time t , either in queue or in service. Setting

$$X^n(t) = N^n(t) - n \quad \text{for } t \geq 0, \tag{3.1}$$

we call $X^n = \{X^n(t), t \geq 0\}$ the *total-customer-count process* in the n th system. One can check that $(X^n(t))^+$ is the number of customers waiting in queue at time t , and $(X^n(t))^-$ is the number of idle servers at time t . Clearly,

$$e'Z^n(t) = n - (X^n(t))^- \quad \text{for } t \geq 0. \quad (3.2)$$

The processes X^n and Z^n describe the “state” of the system as time evolves. Hereafter, they are called the *state processes* for the n th system.

The customers in service are distributed among the K phases following a distribution γ , given by

$$\gamma = \mu R^{-1} p, \quad (3.3)$$

$$R = (I - P') \text{diag}(\nu). \quad (3.4)$$

One can check that $\sum_{k=1}^K \gamma_k = 1$, and one interprets γ_k to be the fraction of phase k load on the n servers.

The preceding paragraph suggests the following centering for the server-allocation process:

$$\hat{Z}^n(t) = Z^n(t) - n\gamma \quad \text{for } t \geq 0.$$

Define the corresponding diffusion-scaled process

$$\tilde{Z}^n(t) = \frac{1}{\sqrt{n}} \hat{Z}^n(t) \quad \text{for } t \geq 0.$$

3.1 Diffusion limits for critically loaded $G/Ph/n + GI$ queues

Throughout Section 3.1, we assume that $\rho = 1$ and that the patience times of customers who arrive after time 0 are iid having distribution function F , which satisfies

$$F(0) = 0 \quad \text{and} \quad \alpha = \lim_{x \downarrow 0} x^{-1} F(x) < \infty. \quad (3.5)$$

Note that for exponentially distributed patience times, α turns out to be the rate of the exponential distribution.

Since ρ is assumed to be 1, we define the diffusion-scaled total-customer-count process \tilde{X}^n by

$$\tilde{X}^n(t) = \frac{1}{\sqrt{n}} X^n(t) \quad \text{for } t \geq 0. \quad (3.6)$$

(When $\rho > 1$, the definition of $\tilde{X}^n(t)$ will be modified, which is given in (3.27).) We assume that

$$(\tilde{X}^n(0), \tilde{Z}^n(0)) \Rightarrow (\tilde{X}(0), \tilde{Z}(0)) \quad \text{as } n \rightarrow \infty \quad (3.7)$$

for a pair of random variables $(\tilde{X}(0), \tilde{Z}(0))$.

The random variables $\tilde{X}(0)$ and $\tilde{Z}(0)$ are assumed to be defined on some probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, which is rich enough so that stochastic processes $\tilde{E}, \tilde{\Phi}^0, \dots, \tilde{\Phi}^K$ and \tilde{S} defined on this space are independent of $(\tilde{X}(0), \tilde{Z}(0))$. Here, \tilde{E} is a one-dimensional driftless Brownian motion, and $\tilde{\Phi}^0, \dots, \tilde{\Phi}^K$ and \tilde{S} are K -dimensional driftless Brownian motions. These Brownian motions, possibly degenerate, are mutually independent and start from 0; the variance of \tilde{E} is λc_a^2 for some constant $c_a^2 \geq 0$, and the covariance matrices of $\tilde{\Phi}^0, \dots, \tilde{\Phi}^K$ and \tilde{S} are H^0, \dots, H^K and $\text{diag}(\nu)$, respectively, where for $k = 0, \dots, K$, the $K \times K$ matrix H^k is given by

$$H_{ij}^k = \begin{cases} p_i^k(1 - p_j^k) & \text{if } i = j \\ -p_i^k p_j^k & \text{otherwise} \end{cases} \quad (3.8)$$

with $p^0 = p$ and p^k being the k th column of P' .

To state the main theorems of this paper, let

$$\tilde{U}(t) = \tilde{X}(0) + \tilde{E}(t) - \mu\beta t + e' \tilde{M}(t), \quad (3.9)$$

$$\tilde{V}(t) = (I - pe')\tilde{Z}(0) + \tilde{\Phi}^0(\mu t) + (I - pe')\tilde{M}(t), \quad (3.10)$$

where

$$\tilde{M}(t) = \sum_{k=1}^K \tilde{\Phi}^k(\nu_k \gamma_k t) - (I - P')\tilde{S}(\gamma t) \quad (3.11)$$

for $t \geq 0$. The process (\tilde{U}, \tilde{V}) is a $(K + 1)$ -dimensional Brownian motion; it is degenerate because $e'\tilde{V}(t) = 0$ for $t \geq 0$. (When $\rho > 1$, the definition of \tilde{U} will be modified in (3.28).) Before we state the first theorem of this paper, we present the following lemma, which is a corollary of Lemma 9 in Appendix A.

Lemma 1. *Let p be a K -dimensional vector that is the distribution of initial phases of the phase-type service times, R be the $K \times K$ matrix defined by (3.4), and $\alpha \geq 0$ be defined by (3.5). (a) For each $(u, v) \in \mathbb{D}^{K+1}$ with $u(t) \in \mathbb{R}$ and $v(t) \in \mathbb{R}^K$ for $t \geq 0$, there exists a unique $(x, z) \in \mathbb{D}^{K+1}$ with $x(t) \in \mathbb{R}$ and $z(t) \in \mathbb{R}^K$ for $t \geq 0$, such that*

$$x(t) = u(t) - \alpha \int_0^t (x(s))^+ ds - e'R \int_0^t z(s) ds, \quad (3.12)$$

$$z(t) = v(t) - p(x(t))^- - (I - pe')R \int_0^t z(s) ds \quad (3.13)$$

for $t \geq 0$. (b) For each $(u, v) \in \mathbb{D}^{K+1}$, define $\Phi(u, v) = (x, z) \in \mathbb{D}^{K+1}$, where (x, z) satisfies (3.12) and (3.13). The map Φ is well defined and is continuous when both the domain and the range \mathbb{D}^{K+1} are endowed with the Skorohod J_1 -topology. (c) The map Φ is Lipschitz continuous in the sense that for any $T > 0$, there exists a constant $C_T^1 > 0$ such that

$$\sup_{0 \leq t \leq T} |\Phi(y)(t) - \Phi(\tilde{y})(t)| \leq C_T^1 \sup_{0 \leq t \leq T} |y(t) - \tilde{y}(t)| \quad \text{for any } y, \tilde{y} \in \mathbb{D}^{K+1}. \quad (3.14)$$

(d) The map Φ is positively homogeneous in the sense that

$$\Phi(ay) = a\Phi(y) \quad \text{for each } a > 0 \text{ and each } y \in \mathbb{D}^{K+1}. \quad (3.15)$$

Let A_0^n be the number of customers who are waiting in queue at time 0 but will eventually abandon the system, and

$$\tilde{A}_0^n = \frac{1}{\sqrt{n}} A_0^n.$$

To state Theorem 1, we assume that

$$\tilde{A}_0^n \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3.16)$$

Clearly, condition (3.16) is satisfied if no customers are waiting in queue at time 0. The validity of this initial condition will be further discussed at the end of Section 3.1.

Theorem 1. *Consider a sequence of $G/Ph/n + GI$ queues satisfying (2.1) and (2.2). Assume that $\rho = 1$ and that (3.5), (3.7) and (3.16) hold. Then*

$$(\tilde{X}^n, \tilde{Z}^n) \Rightarrow (\tilde{X}, \tilde{Z}) \quad \text{as } n \rightarrow \infty,$$

where

$$(\tilde{X}, \tilde{Z}) = \Phi(\tilde{U}, \tilde{V}). \quad (3.17)$$

Suppose that each customer, including those initial customers who are waiting in queue at time 0, samples his first service phase that he is yet to enter following distribution p at his arrival time to the system. One can stratify the customers in the waiting buffer according to their first service phases. For $k = 1, \dots, K$, we use $Q_k^n(t)$ to denote the number of waiting customers at time t whose initial service phase will be phase k ($Q_k^n(t) = 0$ for $t \geq 0$ if phase k is not a first service phase for any customer), and we use $Y_k^n(t)$ to denote the number of phase k customers in the system at time t , either waiting or in service. Let $Q^n(t)$ and $Y^n(t)$ denote the corresponding K -dimensional vectors. Set

$$\tilde{Q}_k^n(t) = \frac{1}{\sqrt{n}} Q_k^n(t), \quad \tilde{Y}_k^n(t) = \frac{1}{\sqrt{n}} \hat{Y}_k^n(t), \quad \hat{Y}_k^n(t) = Y_k^n(t) - n\gamma_k \quad \text{for } t \geq 0. \quad (3.18)$$

Clearly,

$$\tilde{Y}_k^n(t) = \tilde{Q}_k^n(t) + \tilde{Z}_k^n(t) \quad \text{and} \quad \tilde{X}^n(t) = e' \tilde{Y}^n(t) \quad \text{for } t \geq 0. \quad (3.19)$$

The following lemma says that for critically loaded systems, the waiting customers are distributed among the K phases following distribution p . It is known as the state-space-collapse (SSC) result.

Lemma 2. *Under the conditions of Theorem 1, for any $T > 0$,*

$$\frac{1}{\sqrt{n}} \sup_{0 \leq t \leq T} |Q^n(t) - p(X^n(t))^+| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3.20)$$

The following theorem is a corollary to Theorem 1 and Lemma 2. When there is no customer abandonment and the arrival process is renewal, it is identical to Theorem 2.3 of Puhalskii and Reiman (2000).

Theorem 2. *Under the conditions of Theorem 1,*

$$(\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n) \Rightarrow (\tilde{X}, \tilde{Y}, \tilde{Z}) \quad \text{as } n \rightarrow \infty,$$

where (\tilde{X}, \tilde{Z}) is defined in (3.17) and

$$\tilde{Y}(t) = p(\tilde{X}(t))^+ + \tilde{Z}(t). \quad (3.21)$$

The process \tilde{Y} satisfies

$$\begin{aligned} \tilde{Y}(t) = & \tilde{Y}(0) - \beta\mu pt + \tilde{\Phi}^0(\mu t) + p\tilde{E}(t) + \tilde{M}(t) \\ & - R \int_0^t \tilde{Y}(s) ds + (R - \alpha I)p \int_0^t (e'\tilde{Y}(s))^+ ds \quad \text{for } t \geq 0. \end{aligned}$$

The process \tilde{Y} in Theorem 2 is a diffusion process; see Rogers and Williams (2000, page 110) or Karlin and Taylor (1981, page 159) for a definition of diffusion processes. Therefore, \tilde{Y} is a continuous Markov process. The map $\tilde{\Phi}$ in (3.17) defines (\tilde{X}, \tilde{Z}) as a $(K + 1)$ -dimensional continuous process, which is degenerate because it lives on a K -dimensional manifold. From the K -dimensional process \tilde{Y} , one can recover the $(K + 1)$ -dimensional process (\tilde{X}, \tilde{Z}) via

$$\tilde{X}(t) = e'\tilde{Y}(t) \quad \text{and} \quad \tilde{Z}(t) = \tilde{Y}(t) - p(\tilde{X}(t))^+ \quad \text{for } t \geq 0. \quad (3.22)$$

Therefore, (\tilde{X}, \tilde{Z}) is also a continuous Markov process. However, the process (\tilde{X}, \tilde{Z}) is not a diffusion process by the common definition because the function x^+ in (3.22) is not twice differentiable in x at 0. Whitt (2005, Remark 2.2) makes a similar observation that his limit process is not a diffusion process, but a simple transformation of his limit process is a diffusion process.

Our next theorem is concerned with the *virtual waiting time process* $W^n = \{W^n(t), t \geq 0\}$. Here, $W^n(t)$ is the potential waiting time of a hypothetical, infinitely patient customer who arrives at the queue at time t . When there is no customer abandonment and the arrival process is renewal, the theorem is implied by Corollary 2.3 and Remark 2.6 of Puhalskii and Reiman (2000).

Theorem 3. *Under the conditions of Theorem 1,*

$$\sqrt{n}W^n \Rightarrow \frac{(\tilde{X})^+}{\mu} \quad \text{as } n \rightarrow \infty. \quad (3.23)$$

We end this section by the following lemma, which gives a justification for imposing initial condition (3.16) in Theorems 1–3. Let $A_Q^n(t)$ be the number of customers in the n th system who are waiting in queue at time t , but will eventually abandon the system. Clearly,

$$A_0^n = A_Q^n(0).$$

Its diffusion-scaled version is given by

$$\tilde{A}_Q^n(t) = \frac{1}{\sqrt{n}} A_Q^n(t) \quad \text{for } t \geq 0. \quad (3.24)$$

Regarding the process $\tilde{A}_Q^n = \{\tilde{A}_Q^n(t), t \geq 0\}$, we have the following result.

Lemma 3. *Under the conditions of Theorem 1,*

$$\tilde{A}_Q^n \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3.25)$$

The proof of Lemma 3 is presented in Appendix B. Assume that the queue is initially empty. Then condition (3.16) is satisfied at time $t = 0$. Under an additional assumption (3.7), Theorem 1 and Lemma 3 imply that for any $t > 0$,

$$\tilde{A}_Q^n(t) \Rightarrow 0 \quad \text{and} \quad (\tilde{X}^n(t), \tilde{Z}^n(t)) \Rightarrow (\tilde{X}(t), \tilde{Z}(t)) \quad \text{as } n \rightarrow \infty.$$

Thus, if we start to observe the system at any fixed time $t > 0$, initial conditions (3.7) and (3.16) are indeed satisfied at time t . Condition (3.16) is used to prove asymptotic relationship (6.4) in the critically-loaded regime; this relationship between the abandonment-count process and the queue-length process is the key to extending the diffusion limits for $G/Ph/n+M$ queues to the $G/Ph/n+GI$ model with a general patience time distribution. Condition (3.16) is necessary for the asymptotic relationship to hold. In Mandelbaum and Momčilović (2009), an initial assumption similar to (3.16) is made for the $G/GI/n+GI$ model in the critically-loaded regime.

3.2 A diffusion limit for overloaded $G/Ph/n+M$ queues

Our next result is for overloaded $G/Ph/n+M$ systems, where the patience times of all customers, including those waiting in queue at time 0, are assumed to be iid following an exponential distribution. We use α to denote the rate of the exponential patience time distribution. Note that this definition of α is consistent with the definition in (3.5). Assume that $\rho > 1$. Intuitively, when n is large, all n servers are 100% busy, and there should be nq customers on average waiting in the buffer, where

$$q = \frac{\lambda - \mu}{\alpha}. \quad (3.26)$$

An intuitive explanation is as follows: $\lambda - \mu$ is the number of customers per unit of time that the system must “delete” in order for the system to reach an equilibrium. While in

equilibrium, each waiting customer abandons the system at rate α , and collectively all q waiting customers abandon the system at rate $q\alpha$ customers per unit of time. Thus, one should have $q\alpha = \lambda - \mu$, which leads to (3.26). Readers are referred to Whitt (2004) for further discussion on the derivation of (3.26).

Now we modify the definition of \tilde{X}^n in (3.6) and \tilde{U} in (3.9). Let

$$\tilde{X}^n(t) = \frac{1}{\sqrt{n}} \hat{X}^n(t), \quad \hat{X}^n(t) = X^n(t) - nq, \quad (3.27)$$

$$\tilde{U}(t) = \tilde{X}(0) + \tilde{E}(t) - \mu\beta t + e' \tilde{M}(t) - \tilde{G}(qt), \quad (3.28)$$

for $t \geq 0$. In (3.28), the process $\tilde{G} = \{\tilde{G}(t), t \geq 0\}$ is a one-dimensional driftless Brownian motion starting from 0, which has variance α and is independent of $\tilde{E}, \tilde{\Phi}^0, \dots, \tilde{\Phi}^K$ and \tilde{S} (recall that $\tilde{E}, \tilde{\Phi}^0, \dots, \tilde{\Phi}^K$ and \tilde{S} are Brownian motions defined in Section 3.1, and \tilde{M} is given by (3.11)). When $\rho = 1$, one has $q = 0$. Thus, definitions in (3.27) and (3.28) are consistent with (3.6) and (3.9). Assume that

$$(\tilde{X}^n(0), \tilde{Z}^n(0)) \Rightarrow (\tilde{X}(0), \tilde{Z}(0)) \quad \text{as } n \rightarrow \infty \quad (3.29)$$

for a pair of random variables $(\tilde{X}(0), \tilde{Z}(0))$.

Before presenting Theorem 4, we introduce the next lemma, which is also a corollary of Lemma 9.

Lemma 4. *Let p be a K -dimensional vector that is the distribution of initial phases of the phase-type service times, R be the $K \times K$ matrix defined by (3.4), and $\alpha \geq 0$ be defined by (3.5). (a) For each $(u, v) \in \mathbb{D}^{K+1}$ with $u(t) \in \mathbb{R}$ and $v(t) \in \mathbb{R}^K$ for $t \geq 0$, there exists a unique $(x, z) \in \mathbb{D}^{K+1}$ with $x(t) \in \mathbb{R}$ and $z(t) \in \mathbb{R}^K$ for $t \geq 0$, such that*

$$x(t) = u(t) - \alpha \int_0^t x(s) ds - e' R \int_0^t z(s) ds, \quad (3.30)$$

$$z(t) = v(t) - (I - pe') R \int_0^t z(s) ds \quad (3.31)$$

for $t \geq 0$. (b) For each $(u, v) \in \mathbb{D}^{K+1}$, define $\Psi(u, v) = (x, z) \in \mathbb{D}^{K+1}$, where (x, z) satisfies (3.30) and (3.31). The map Ψ is well defined and is continuous when both the domain and the range \mathbb{D}^{K+1} are endowed with the Skorohod J_1 -topology. (c) The map Ψ is Lipschitz continuous in the sense that for any $T > 0$, there exists a constant $C_T^2 > 0$ such that

$$\sup_{0 \leq t \leq T} |\Psi(y)(t) - \Psi(\tilde{y})(t)| \leq C_T^2 \sup_{0 \leq t \leq T} |y(s) - \tilde{y}(s)| \quad \text{for any } y, \tilde{y} \in \mathbb{D}^{K+1}. \quad (3.32)$$

(d) The map Ψ is positively homogeneous in the sense that

$$\Psi(ay) = a\Psi(y) \quad \text{for each } a > 0 \text{ and each } y \in \mathbb{D}^{K+1}. \quad (3.33)$$

Theorem 4. Consider a sequence of $G/Ph/n + M$ queues satisfying (2.1) and (2.2). Assume that $\rho > 1$ and that (3.29) holds. Then

$$(\tilde{X}^n, \tilde{Z}^n) \Rightarrow (\tilde{X}, \tilde{Z}) \quad \text{as } n \rightarrow \infty,$$

where

$$(\tilde{X}, \tilde{Z}) = \Psi(\tilde{U}, \tilde{V}). \quad (3.34)$$

Equation (3.34) defines (\tilde{X}, \tilde{Z}) as a $(K+1)$ -dimensional diffusion process, which is also degenerate and lives on a K -dimensional manifold.

3.3 A roadmap for proofs

Theorems 1 and 4 are the main results of this paper. Theorem 4 and a restricted version of Theorem 1 are proved in Section 5; the restriction is to assume that the patience times are exponentially distributed. These proofs use standard FCLTs and then apply the continuous-mapping theorem and the random-time-change theorem. To construct appropriate continuous maps, we introduce a perturbed system in Section 4.1, which is equal in distribution to the original system when the patience time distribution is exponential. Using the perturbed system, we are able to construct a set of system equations in Section 4.2, which is critical to define the continuous maps.

Section 6 is devoted to proving the general version of Theorem 1. When the patience time distribution is general and the systems are critically loaded, we first modify the preceding system equations slightly by replacing the cumulative number of customer abandonments by an integral of the queue-length process. We then apply an asymptotic relationship in Dai and He (2009) to establish a result that the error from the replacement is negligible under a stochastic boundedness assumption of the diffusion-scaled queue-length process; the latter assumption holds by a comparison result in Dai and He (2009) and the restricted version of Theorem 1 proved in Section 5.

Theorem 2 is a corollary to Theorem 1 and Lemma 2; the latter is proved in Appendix B. Theorem 3 is also proved in Appendix B.

4 System representation for a $G/Ph/n + M$ queue

In this section, we first describe a perturbed system of the $G/Ph/n + M$ model, and show that this perturbed system is equivalent to the $G/Ph/n + M$ queue. We then develop the dynamical equations that the perturbed system must satisfy.

4.1 A perturbed system

Now we describe a perturbation of the $G/Ph/n + M$ model. In the perturbed system, each phase has a *service queue* for the customers “in service”. Only the leading customer in

the service queue is actually in service; all others are waiting in the service queue, ordered according to the FIFO discipline. We use $Z_k^*(t)$ to denote the number of customers in phase k service queue at time t . (Star-version quantities are associated with the perturbed system; the corresponding quantities in the original system are denoted by the same symbols without the star.) Each customer in the service queue is attached to exactly one server. Thus, there are exactly $Z_k^*(t)$ servers in phase k at time t . All these $Z_k^*(t)$ servers simultaneously work on the leading customer in the service queue. The service effort received by the leading customer is additive, proportional to the number of servers working on the customer. Each customer has a service requirement for each phase that he visits, with phase k service requirement being exponentially distributed with mean $1/\nu_k$.

When the total service effort spent on a customer reaches his service requirement in a phase, the service in the phase is completed. When a customer completes a phase k service, he immediately moves to the next phase following a sampling procedure to be specified below, taking with him the associated server. If the service queue in the new phase is not empty at their arrival, the server joins the service immediately, collaborating with other servers who are already in service to work on the leading customer in the service queue. The newly arrived customer joins the end of the service queue. If the new service queue is empty, the server works on her customer who is the only one in the new phase of service.

When a customer finishes a phase k service, it uses the k th row of P to sample the next phase of service to join among phases $\{1, \dots, K, K+1\}$; the probability of selecting phase $K+1$ is $1 - \sum_{\ell=1}^K P_{k\ell}$. If $\ell \in \{1, \dots, K\}$ is selected, both the server and the customer move next to phase ℓ . If $K+1$ is selected, the customer exits the system and the associated server is released. The released server checks the FIFO real queue to select the next customer to work on if the real queue is not empty. The selected customer is attached to the server until the customer exits the system. If the real queue is empty, the server becomes idle.

At a customer's arrival time to the system, if there is an idle server, the customer grabs a free server and is attached to the server until the customer exits the system. Together, they move into the customer's first phase of service, which is selected according to distribution p . The service and waiting mechanism is identical to the one described in the previous paragraph. If all servers are busy at the customer's arrival time, the customer joins the end of the FIFO real queue. Only the leading customer in the FIFO real queue can abandon the system; other waiting customers are infinitely patient until they become a leading customer. The patience time of the leading customer is exponentially distributed with mean $1/\alpha$. The customer abandons the system without service if his patience clock exceeds the patience time. The patience clock starts from 0 when the customer becomes a leading customer and increases at rate k when the queue length is k .

For each n fixed, now we show that when the arrival process E^n is a renewal process, the perturbed system and the original system are equivalent in a precise mathematical sense. For that, recall that the waiting buffer in the perturbed system maintains a FIFO

queue for waiting customers. Let

$$\mathcal{Q}^*(t) = (i_1, \dots, i_{L^*(t)}),$$

where $L^*(t)$ is the total number of customers waiting in queue at time t , and i_ℓ is the *first service phase* that the ℓ th customer is yet to enter.

Let $\xi(t)$ be the remaining interarrival time at time t . (ξ has no star because the arrival processes in the perturbed system and in the original system are identical.) It follows that, $\{(\xi(t), \mathcal{Q}^*(t), Z^*(t)), t \geq 0\}$ is a continuous-time Markov process living in state space $\mathbb{R}_+ \times \mathcal{K}^\infty \times \mathbb{Z}_+^K$, where \mathcal{K}^∞ is the space of finite sequences taking values in $\mathcal{K} = \{1, \dots, K\}$.

Let $\{(\xi(t), \mathcal{Q}(t), Z(t)), t \geq 0\}$ be the corresponding process of the original system. The process is also a continuous-time Markov process. At any time t , the phase k service rate is $Z_k^*(t)\nu_k$ in the perturbed system and $Z_k(t)\nu_k$ in the original system, while the abandonment rate is $L^*(t)\alpha$ in the perturbed system and $L(t)\alpha$ in the original system. One can check that the two Markov processes

$$\{(\xi(t), \mathcal{Q}(t), Z(t)), t \geq 0\} \quad \text{and} \quad \{(\xi(t), \mathcal{Q}^*(t), Z^*(t)), t \geq 0\}$$

have the same generator. Thus, when they have the same initial distribution, they have the same distribution for the entire processes. In the following, we always choose the initial condition of the perturbed system to be identical to that of the original system.

Even if the arrival process is not a renewal process, the perturbed system can still have the same distribution as the original system. The rest of the paper does not require the arrival process to be renewal. Rather, we assume that each arrival process satisfies the requirement that the perturbed system has the same distribution as the original one. See Tezcan (2006) for a more general treatment of perturbed systems.

4.2 System equations

From now on, we focus on the perturbed system of the $G/Ph/n + M$ queue and drop the stars attached to its quantities. We assume that the patience times of all customers, including those who are waiting in queue at time 0, are exponentially distributed with rate α . In this section, we describe the dynamical equations that the system must obey. For $k = 1, \dots, K$, let $\phi^k = \{\phi^k(j), j \in \mathbb{N}\}$ be a sequence of iid ‘‘Bernoulli random vectors’’. For each j , the K -dimensional random vector $\phi^k(j)$ takes vector e^ℓ with probability $P_{k\ell}$ and takes the K -dimensional zero vector with probability $1 - \sum_{\ell=1}^K P_{k\ell}$. Similarly, let $\phi^0 = \{\phi^0(j), j \in \mathbb{N}\}$ be a sequence of iid K -dimensional random vectors; the probability that $\phi^0(j) = e^\ell$ is p_ℓ . For $k = 0, \dots, K$, define the routing process

$$\Phi^k(N) = \sum_{j=1}^N \phi^k(j) \quad \text{for } N \in \mathbb{N}.$$

For each $k = 1, \dots, K$, let S_k be a Poisson process with rate ν_k , and let G be a Poisson process with rate α . We assume that

$$X^n(0), E^n, S_1, \dots, S_K, \Phi^0, \dots, \Phi^K \text{ and } G \text{ are mutually independent.} \quad (4.1)$$

Let $T_k^n(t)$ be the cumulative amount of service effort received by customers in phase k service in $(0, t]$, $B^n(t)$ be the cumulative number of customers who have entered service in $(0, t]$, and $D^n(t)$ be the cumulative number of customers who have completed service in $(0, t]$. Clearly,

$$T_k^n(t) = \int_0^t Z_k^n(s) ds \quad \text{for } t \geq 0. \quad (4.2)$$

Then $S_k(T_k^n(t))$ is the cumulative number of phase k service completions by time t . Also $G(\int_0^t (X^n(s))^+ ds)$ is the cumulative number of customers who have abandoned the system by time t . One can check that for $t \geq 0$, the processes X^n and Z^n satisfy the following dynamical equations:

$$Z^n(t) = Z^n(0) + \Phi^0(B^n(t)) + \sum_{k=1}^K \Phi^k(S_k(T_k^n(t))) - S(T^n(t)), \quad (4.3)$$

$$X^n(t) = X^n(0) + E^n(t) - D^n(t) - G\left(\int_0^t (X^n(s))^+ ds\right), \quad (4.4)$$

$$\begin{aligned} D^n(t) &= \sum_{k=1}^K \left(S_k(T_k^n(t)) - e' \Phi^k(S_k(T_k^n(t))) \right) \\ &= -e' \left(\sum_{k=1}^K \Phi^k(S_k(T_k^n(t))) - S(T^n(t)) \right), \end{aligned} \quad (4.5)$$

where

$$S(T^n(t)) = (S_1(T_1^n(t)), \dots, S_K(T_K^n(t)))'.$$

4.3 State-process representation

Define the centered processes

$$\hat{S}(t) = S(t) - \nu t, \quad \hat{G}(t) = G(t) - \alpha t, \quad \hat{\Phi}^\ell(N) = \sum_{j=1}^N (\phi^\ell(j) - p^\ell),$$

for $t \geq 0$, $\ell = 0, \dots, K$ and $N \in \mathbb{N}$, where $p^0 = p$ and p^k is the k th column of P' for $k = 1, \dots, K$. Setting

$$M^n(t) = \sum_{k=1}^K \hat{\Phi}^k(S_k(T_k^n(t))) - (I - P') \hat{S}(T^n(t)), \quad (4.6)$$

one then has

$$\sum_{k=1}^K \Phi^k(S_k(T_k^n(t))) - S(T^n(t)) = M^n(t) - R \int_0^t Z^n(s) ds,$$

where R is defined in (3.4). By (4.3) and (4.5),

$$e'Z^n(t) = e'Z^n(0) + B^n(t) - D^n(t), \quad (4.7)$$

$$D^n(t) = -e'M^n(t) + e'R \int_0^t Z^n(s) ds. \quad (4.8)$$

It follows from (3.2) and (4.3)–(4.8) that

$$\begin{aligned} Z^n(t) &= Z^n(0) + p(X^n(0))^- + \hat{\Phi}^0(B^n(t)) - p(X^n(t))^- \\ &\quad + (I - pe')M^n(t) - (I - pe')R \int_0^t Z^n(s) ds, \\ X^n(t) &= X^n(0) + \hat{E}^n(t) + \lambda^n t + e'M^n(t) - e'R \int_0^t Z^n(s) ds \\ &\quad - \hat{G} \left(\int_0^t (X^n(s))^+ ds \right) - \alpha \int_0^t (X^n(s))^+ ds. \end{aligned}$$

Recall that $\hat{Z}^n(t) = Z^n(t) - n\gamma$. We then have

$$\begin{aligned} \hat{Z}^n(t) &= (I - pe')\hat{Z}^n(0) + \hat{\Phi}^0(B^n(t)) - p(X^n(t))^- \\ &\quad + (I - pe')M^n(t) - (I - pe')R \int_0^t \hat{Z}^n(s) ds, \\ X^n(t) &= X^n(0) + \hat{E}^n(t) + (\lambda^n - n\mu)t + e'M^n(t) - e'R \int_0^t \hat{Z}^n(s) ds \\ &\quad - \hat{G} \left(\int_0^t (X^n(s))^+ ds \right) - \alpha \int_0^t (X^n(s))^+ ds, \end{aligned}$$

where we have used (3.3) and (3.4) in the derivations. Setting

$$U^n(t) = X^n(0) + \hat{E}^n(t) + (\lambda^n - n\mu)t + e'M^n(t) - \hat{G} \left(\int_0^t (X^n(s))^+ ds \right), \quad (4.9)$$

$$V^n(t) = (I - pe')\hat{Z}^n(0) + \hat{\Phi}^0(B^n(t)) + (I - pe')M^n(t) \quad (4.10)$$

for $t \geq 0$, we finally have

$$X^n(t) = U^n(t) - \alpha \int_0^t (X^n(s))^+ ds - e'R \int_0^t \hat{Z}^n(s) ds, \quad (4.11)$$

$$\hat{Z}^n(t) = V^n(t) - p(X^n(t))^- - (I - pe')R \int_0^t \hat{Z}^n(s) ds. \quad (4.12)$$

By Lemma 1, we have obtained the following representation for the state processes

$$(X^n, \hat{Z}^n) = \Phi(U^n, V^n). \quad (4.13)$$

5 Proofs for $G/Ph/n + M$ queues

This section provides proofs for Theorem 4 and a special version of Theorem 1 when the patience time distribution is exponential. Section 5.1 first establishes a fluid limit, which is needed in applying the random-time-change theorem to prove the theorems in Section 5.2.

5.1 Fluid limits

For $t \geq 0$, define the fluid-scaled processes $\bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{T}^n, \bar{X}^n$ and \bar{Z}^n via

$$\begin{aligned} \bar{B}^n(t) &= \frac{1}{n}B^n(t), & \bar{D}^n(t) &= \frac{1}{n}D^n(t), & \bar{E}^n(t) &= \frac{1}{n}E^n(t), \\ \bar{T}^n(t) &= \frac{1}{n}T^n(t), & \bar{X}^n(t) &= \frac{1}{n}X^n(t), & \bar{Z}^n(t) &= \frac{1}{n}Z^n(t). \end{aligned}$$

Theorem 5. *Consider a sequence of $G/Ph/n + M$ queues satisfying (2.1) and (2.2). Assume (3.29) holds. Then*

$$(\bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{T}^n, \bar{X}^n, \bar{Z}^n) \Rightarrow (\bar{B}, \bar{D}, \bar{E}, \bar{T}, \bar{X}, \bar{Z}) \quad \text{as } n \rightarrow \infty, \quad (5.1)$$

where $\bar{B}(t) = \mu t$, $\bar{D}(t) = \mu t$, $\bar{E}(t) = \lambda t$, $\bar{T}(t) = \gamma t$, $\bar{X}(t) = q$ and $\bar{Z}(t) = \gamma$ for $t \geq 0$.

Proof. For $t \geq 0$, let

$$\bar{M}^n(t) = \frac{1}{n}M^n(t), \quad \bar{U}^n(t) = \frac{1}{n}U^n(t), \quad \bar{V}^n(t) = \frac{1}{n}V^n(t), \quad \bar{L}^n(t) = \frac{1}{n}\hat{Z}^n(t). \quad (5.2)$$

By (4.13) and the positively homogeneous property of Φ , we have

$$(\bar{X}^n, \bar{L}^n) = \Phi(\bar{U}^n, \bar{V}^n).$$

Setting

$$\bar{U}(t) = q + (\lambda - \mu)t \quad \text{and} \quad \bar{V}(t) = 0 \quad \text{for } t \geq 0, \quad (5.3)$$

one can check that $\Phi(\bar{U}, \bar{V}) = (\bar{X}, 0)$. We are going to show that

$$(\bar{M}^n, \bar{U}^n, \bar{V}^n) \Rightarrow (0, \bar{U}, 0) \quad \text{as } n \rightarrow \infty. \quad (5.4)$$

Assuming (5.4), we now complete the proof of the theorem. The continuity of the map Φ implies that

$$(\bar{X}^n, \bar{L}^n) = \Phi(\bar{U}^n, \bar{V}^n) \Rightarrow \Phi(\bar{U}, \bar{V}) = (\bar{X}, 0) \quad \text{as } n \rightarrow \infty.$$

Since $\bar{Z}^n(t) = \bar{L}^n(t) + \gamma$ for $t \geq 0$, then $\bar{Z}^n \Rightarrow \bar{Z}$ as $n \rightarrow \infty$, from which one has $\bar{T}^n \Rightarrow \bar{T}$ as $n \rightarrow \infty$. By (4.8),

$$\bar{D}^n(t) = -e' \bar{M}^n(t) + e' R \int_0^t \bar{Z}^n(s) ds.$$

Since $e' R \int_0^t \bar{Z}(s) ds = \mu t$ for $t \geq 0$, by the continuous-mapping theorem $\bar{D}^n \Rightarrow \bar{D}$ as $n \rightarrow \infty$. The convergence of \bar{D}^n and (4.7) imply that $\bar{B}^n \Rightarrow \bar{B}$ as $n \rightarrow \infty$, and \bar{B} satisfies

$$e' \bar{Z}(t) = e' \bar{Z}(0) + \bar{B}(t) - \bar{D}(t) \quad \text{for } t \geq 0.$$

Since $\bar{Z}(t) = \bar{Z}(0) = \gamma$, we conclude that $\bar{B}(t) = \mu t$ for $t \geq 0$. By assumptions (2.1) and (2.2), for each $T > 0$,

$$\frac{1}{n} \sup_{0 \leq t \leq T} |\hat{E}^n(t)| \Rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (5.5)$$

which implies that $\bar{E}^n \Rightarrow \bar{E}$. This proves the theorem when (5.4) holds.

It remains to prove (5.4). By the functional strong law of large numbers (FSLLN),

$$\frac{1}{n} \sup_{0 \leq t \leq T} |\hat{G}(nt)| \Rightarrow 0, \quad \frac{1}{n} \sup_{0 \leq t \leq T} |\hat{S}(nt)| \Rightarrow 0, \quad \frac{1}{n} \sup_{0 \leq t \leq T} |\hat{\Phi}^k(\lfloor nt \rfloor)| \Rightarrow 0 \quad (5.6)$$

as $n \rightarrow \infty$, for $k = 0, \dots, K$. Let $\bar{S}^n(t) = S^n(nt)/n$ for $t \geq 0$. The FSLLN also leads to $\bar{S}^n \Rightarrow \bar{S}$ as $n \rightarrow \infty$ where $\bar{S}(t) = \nu t$ for $t \geq 0$. This, together with (4.6), (5.6) and the fact $\bar{T}_k^n(t) \leq t$ for $t \geq 0$, implies that

$$\sup_{0 \leq t \leq T} |\bar{M}^n(t)| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (5.7)$$

Note that $\bar{B}^n(t) \leq (\bar{X}^n(0))^+ + \bar{E}^n(t)$. By (3.7) and the convergence of \bar{E}^n , the sequence of processes $\{\bar{B}^n, n \in \mathbb{N}\}$ is stochastically bounded, that is, for each $T > 0$,

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[\sup_{0 \leq t \leq T} \bar{B}^n(t) > a \right] = 0.$$

Using this and (5.6), we deduce that

$$\sup_{0 \leq t \leq T} \frac{1}{n} \hat{\Phi}^0(B^n(t)) \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (5.8)$$

Condition (3.7) implies that $\hat{Z}^n(0)/n \Rightarrow 0$ as $n \rightarrow \infty$, which, together with (5.7) and (5.8), leads to $\bar{V}^n \Rightarrow 0$ as $n \rightarrow \infty$. Since $\sup_{0 \leq t \leq T} (\bar{X}^n(t))^+ \leq (\bar{X}^n(0))^+ + \bar{E}^n(T)$, one can argue similarly that $\bar{U}^n \Rightarrow \bar{U}$ as $n \rightarrow \infty$. Hence, we have shown (5.4) holds, and thus proved the theorem. \square

5.2 Diffusion limits

In this section, we prove Theorems 1 and 4, under the assumption that the patience times are exponentially distributed.

Define the diffusion-scaled processes \tilde{G}^n , \tilde{S}^n and $\tilde{\Phi}^0, \dots, \tilde{\Phi}^K$ via

$$\tilde{G}^n(t) = \frac{1}{\sqrt{n}}\hat{G}(nt), \quad \tilde{S}^n(t) = \frac{1}{\sqrt{n}}\hat{S}(nt), \quad \tilde{\Phi}^{k,n}(t) = \frac{1}{\sqrt{n}}\hat{\Phi}^k(\lfloor nt \rfloor)$$

for $t \geq 0$ and $k = 0, \dots, K$. By the FCLT, one has

$$(\tilde{G}^n, \tilde{S}^n, \tilde{\Phi}^{0,n}, \dots, \tilde{\Phi}^{K,n}) \Rightarrow (\tilde{G}, \tilde{S}, \tilde{\Phi}^0, \dots, \tilde{\Phi}^K) \quad \text{as } n \rightarrow \infty,$$

where \tilde{G} is a one-dimensional driftless Brownian motion, and \tilde{S} and $\tilde{\Phi}^0, \dots, \tilde{\Phi}^K$ are K -dimensional driftless Brownian motions. As mentioned previously, the variance of \tilde{G} is α , the covariance matrix for \tilde{S} is $\text{diag}(\nu)$, and for $k = 0, \dots, K$, the covariance matrix for $\tilde{\Phi}^k$ is H^k given by (3.8). By the FCLT assumption (2.2) for the arrival process E^n , the initial condition (3.29), and the independence assumption (4.1), one has

$$(\tilde{X}^n(0), \tilde{Z}^n(0), \tilde{E}^n, \tilde{G}^n, \tilde{S}^n, \tilde{\Phi}^{0,n}, \dots, \tilde{\Phi}^{K,n}) \Rightarrow (\tilde{X}(0), \tilde{Z}(0), \tilde{E}, \tilde{G}, \tilde{S}, \tilde{\Phi}^0, \dots, \tilde{\Phi}^K) \quad (5.9)$$

as $n \rightarrow \infty$. The components of $(\tilde{E}, \tilde{G}, \tilde{S}, \tilde{\Phi}^0, \dots, \tilde{\Phi}^K)$ are mutually independent, and they are independent of $(\tilde{X}(0), \tilde{Z}(0))$.

Let $\tilde{U}^n(t) = U^n(t) - n\bar{U}(t)$, and define the diffusion-scaled processes

$$\tilde{U}^n(t) = \frac{1}{\sqrt{n}}\hat{U}^n(t) \quad \text{and} \quad \tilde{V}^n(t) = \frac{1}{\sqrt{n}}V^n(t) \quad \text{for } t \geq 0,$$

where \bar{U} is defined in (5.3). We now have the following lemma.

Lemma 5. *Consider a sequence of $G/Ph/n+M$ queues satisfying (2.1) and (2.2). Assume that (3.29) holds. Then*

$$(\tilde{U}^n, \tilde{V}^n) \Rightarrow (\tilde{U}, \tilde{V}) \quad \text{as } n \rightarrow \infty,$$

where (\tilde{U}, \tilde{V}) is a $(K+1)$ -dimensional Brownian motion defined by (3.28) and (3.10).

Proof. By (4.9) and (4.10),

$$\tilde{U}^n(t) = \tilde{X}^n(0) + \tilde{E}^n(t) + \sqrt{n} \left(\frac{1}{n}\lambda^n - \lambda \right) + e'\tilde{M}^n(t) - \tilde{G}^n \left(\int_0^t (\tilde{X}^n(s))^+ ds \right), \quad (5.10)$$

$$\tilde{V}^n(t) = (I - pe')\tilde{Z}^n(0) + \tilde{\Phi}^{0,n}(\tilde{B}^n(t)) + (I - pe')\tilde{M}^n(t), \quad (5.11)$$

where

$$\tilde{M}^n(t) = \frac{1}{\sqrt{n}}M^n(t) = \sum_{k=1}^K \tilde{\Phi}^{k,n}(\tilde{S}_k^n(\tilde{T}_k^n(t))) - (I - P')\tilde{S}^n(\tilde{T}^n(t))$$

and $\tilde{S}^n(t) = S(nt)/n$ for $t \geq 0$. By the FSLLN, $\tilde{S}^n \Rightarrow \tilde{S}$ as $n \rightarrow \infty$, where $\tilde{S}(t) = \nu t$ for $t \geq 0$. The lemma now follows from (5.9), Theorem 5, the continuous-mapping theorem, and the random-time-change theorem. \square

Proof of Theorem 1 (assuming an exponential patience time distribution). Since $\rho = 1$, it follows that $q = 0$ and $\lambda = \mu$. Then $\bar{U}(t) = 0$ for $t \geq 0$. It follows from the state-process representation (4.13) and the positively homogeneous property of the map Φ that

$$(\tilde{X}^n, \tilde{Z}^n) = \Phi(\tilde{U}^n, \tilde{V}^n).$$

The theorem now follows from Lemma 5 and the continuous-mapping theorem. \square

Although condition (3.16) is not explicitly required in the above proof, it can be deduced by using initial condition (3.7) and the assumption that the patience times of all customers, including those in queue initially, are iid following an exponential distribution.

Before proving Theorem 4, we first establish a lemma. It says that when $\rho > 1$, the number of idle servers goes to zero under diffusion scaling.

Lemma 6. *Let $I^n(t) = (X^n(t))^-$ and $\tilde{I}^n(t) = I^n(t)/\sqrt{n}$ for $t \geq 0$. Then under the conditions of Theorem 4,*

$$\tilde{I}^n \Rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Proof. It follows from (4.11) and (4.12) that

$$\begin{aligned} \frac{1}{\sqrt{n}}X^n(t) &= \sqrt{n}\bar{U}(t) + \tilde{U}^n(t) - \frac{\alpha}{\sqrt{n}} \int_0^t (X^n(s))^+ ds - e'R \int_0^t \tilde{Z}^n(s) ds, \\ \tilde{Z}^n(t) &= \tilde{V}^n(t) - \frac{p}{\sqrt{n}}(X^n(t))^- - (I - pe')R \int_0^t \tilde{Z}^n(s) ds. \end{aligned}$$

Therefore, by Lemma 1

$$\left(\frac{1}{\sqrt{n}}X^n, \tilde{Z}^n \right) = \Phi(\tilde{U}^n + \sqrt{n}\bar{U}, \tilde{V}^n).$$

By the Lipschitz continuity property (3.14) of the map Φ , for any $T > 0$, there exists a constant $C_T^1 > 0$ such that

$$\sup_{0 \leq t \leq T} \left| \Phi(\tilde{U}^n + \sqrt{n}\bar{U}, \tilde{V}^n)(t) - \Phi(\sqrt{n}\bar{U}, 0)(t) \right| \leq C_T^1 \sup_{0 \leq t \leq T} \left\{ |\tilde{U}^n(t)| + |\tilde{V}^n(t)| \right\}$$

for all n and all sample paths. One can check that $\Phi(\sqrt{n}\bar{U}, 0) = (\sqrt{n}q, 0)$. Therefore,

$$\inf_{0 \leq t \leq T} \frac{1}{\sqrt{n}}X^n(t) \geq \sqrt{n}q - C_T^1 \sup_{0 \leq t \leq T} \left\{ |\tilde{U}^n(t)| + |\tilde{V}^n(t)| \right\}. \quad (5.12)$$

By Lemma 5,

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[\sup_{0 \leq t \leq T} \left\{ |\tilde{U}^n(t)| + |\tilde{V}^n(t)| \right\} > a \right] = 0,$$

which, together with (5.12), implies that $\sup_{0 \leq t \leq T} \tilde{I}^n(t) \Rightarrow 0$ as $n \rightarrow \infty$. The lemma is proved. \square

Proof of Theorem 4. It follows from (4.11) and (4.12) that

$$\begin{aligned}\tilde{X}^n(t) &= \tilde{U}^n(t) - \frac{\alpha}{\sqrt{n}} \int_0^t (X^n(s))^- ds - \alpha \int_0^t \tilde{X}^n(s) ds - e'R \int_0^t \tilde{Z}^n(s) ds, \\ \tilde{Z}^n(t) &= \tilde{V}^n(t) - \frac{p}{\sqrt{n}} (X^n(t))^- - (I - pe')R \int_0^t \tilde{Z}^n(s) ds.\end{aligned}$$

Let

$$\delta^n(t) = \frac{\alpha}{\sqrt{n}} \int_0^t (X^n(s))^- ds \quad \text{and} \quad \epsilon^n(t) = \frac{p}{\sqrt{n}} (X^n(t))^- \quad \text{for } t \geq 0.$$

By Lemma 4,

$$(\tilde{X}^n, \tilde{Z}^n) = \Psi(\tilde{U}^n - \delta^n, \tilde{V}^n - \epsilon^n).$$

By Lemma 6,

$$(\delta^n, \epsilon^n) \Rightarrow (0, 0) \quad \text{as } n \rightarrow \infty. \quad (5.13)$$

The theorem follows from Lemma 5, (5.13), and the continuity of the map Ψ . \square

6 Proofs for critically loaded $G/Ph/n + GI$ queues

In this section, we prove Theorem 1 for a general patience time distribution. Consider a sequence of $G/Ph/n + GI$ queues indexed by n . Our starting point is the perturbed system described in Section 4.1 with the following modification: each customer in queue can abandon the system, not just the leading customer; when a customer's waiting time in the real FIFO queue exceeds his patience time, the customer abandons the system. By the same argument as in Section 4.1, for each n , the modified perturbed system is equivalent in distribution to the original $G/Ph/n + GI$ queue. In particular, the system equations (4.3)–(4.5) derived in Section 4.2 hold, except that (4.4) is modified as follows:

$$X^n(t) = X^n(0) + E^n(t) - D^n(t) - A^n(t), \quad (6.1)$$

where $A^n(t)$ denotes the cumulative number of customers that have abandoned the system by time t . We call $A^n = \{A^n(t), t \geq 0\}$ the *abandonment-count process* in the n th system.

With systems equations (4.3), (6.1) and (4.5), one can derive representation (4.13)

$$(X^n, \hat{Z}^n) = \Phi(U^n, V^n)$$

with U^n modified as

$$U^n(t) = X^n(0) + \hat{E}^n(t) + (\lambda^n - n\mu)t + e'M^n(t) - A^n(t) + \alpha \int_0^t (X^n(s))^+ ds. \quad (6.2)$$

The derivation is identical to the one in Section 4.3 and is not repeated here.

Before we prove Theorem 1, we state two lemmas, which will be proved at the end of this section. The first lemma follows from a main result in Dai and He (2009). In the lemma, the diffusion-scaled abandonment-count process $\tilde{A}^n = \{\tilde{A}^n(t), t \geq 0\}$ is defined by

$$\tilde{A}^n(t) = \frac{1}{\sqrt{n}} A^n(t) \quad \text{for } t \geq 0. \quad (6.3)$$

Lemma 7. *Under the conditions of Theorem 1, for any $T > 0$,*

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t) - \alpha \int_0^t (\tilde{X}^n(s))^+ ds \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (6.4)$$

The next lemma is a generalization of the fluid limit theorem (Theorem 5) to general patience time distributions, but with the restriction that $\rho = 1$.

Lemma 8. *Under the conditions of Theorem 1,*

$$(\bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{T}^n, \bar{X}^n, \bar{Z}^n) \Rightarrow (\bar{B}, \bar{D}, \bar{E}, \bar{T}, \bar{X}, \bar{Z}) \quad \text{as } n \rightarrow \infty, \quad (6.5)$$

where $\bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{T}^n, \bar{X}^n$ and \bar{Z}^n are fluid-scaled processes defined at the beginning of Section 5.1, and $\bar{B}(t) = \mu t, \bar{D}(t) = \mu t, \bar{E}(t) = \lambda t, \bar{T}(t) = \gamma t, \bar{X}(t) = 0$ and $\bar{Z}(t) = \gamma$ for $t \geq 0$.

Proof of Theorem 1. Using the representation (4.13) with U^n given by (6.2), one has

$$(\tilde{X}^n, \tilde{Z}^n) = \Phi(\tilde{U}^n, \tilde{V}^n),$$

where

$$\tilde{U}^n(t) = \tilde{X}^n(0) + \tilde{E}^n(t) + \sqrt{n} \left(\frac{1}{n} \lambda^n - \lambda \right) + e' \tilde{M}^n(t) - \left(\tilde{A}^n(t) - \alpha \int_0^t (\tilde{X}^n(s))^+ ds \right) \quad (6.6)$$

and \tilde{V}^n is given by (5.11). By Lemma 1, the map Φ is continuous. Thus, to prove the theorem, it suffices to prove that

$$(\tilde{U}^n, \tilde{V}^n) \Rightarrow (\tilde{U}, \tilde{V}) \quad (6.7)$$

where (\tilde{U}, \tilde{V}) is the $(K+1)$ -dimensional Brownian motion defined by (3.9) and (3.10). The convergence (6.7) follows from the proof of Lemma 5 with the following two modifications. First, the last term of \tilde{U}^n in (6.6) is

$$\left(\tilde{A}^n(t) - \alpha \int_0^t (\tilde{X}^n(s))^+ ds \right)$$

instead of

$$\tilde{G}^n \left(\int_0^t (\bar{X}^n(s))^+ ds \right)$$

in (5.10). We apply Lemma 7 to conclude that the last term in (6.6) converges to zero in distribution. Second, we use (6.5) instead of (5.1) in order to apply the random-time-change theorem to finish the proof of (6.7). \square

Remark. It follows immediately from Theorem 1 and Lemma 7 that under the conditions of Theorem 1,

$$\tilde{A}^n \Rightarrow \tilde{A} \quad \text{as } n \rightarrow \infty, \quad (6.8)$$

where

$$\tilde{A}(t) = \alpha \int_0^t (\tilde{X}(s))^+ ds \quad \text{for } t \geq 0.$$

Proof of Lemma 7. We use Theorem 1 of Dai and He (2009) to prove the lemma. In order to apply the theorem, we need only verify that the sequence of diffusion-scaled queue-length processes is stochastically bounded, that is, for any $T > 0$,

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[\sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} (X^n(t))^+ > a \right] = 0. \quad (6.9)$$

Theorem 2 of Dai and He (2009) states a comparison result: the queue length at any time in a $G/G/n+G$ queue is dominated by the queue length in the corresponding $G/G/n$ queue without abandonment. Thus, (6.9) is implied by the stochastic boundedness of the sequence of diffusion-scaled queue-length processes in the corresponding $G/Ph/n$ queues. Examining the proof of Theorem 1 in Section 5.2 for an exponential patience time distribution with rate $\alpha > 0$, one concludes that Theorem 1 holds for the corresponding $G/Ph/n$ queues without abandonment by setting $\alpha = 0$. As a consequence, the sequence of diffusion-scaled queue-length processes in the $G/Ph/n$ queues is stochastically bounded. \square

Proof of Lemma 8. The proof of the lemma follows the proof of Theorem 5 with the following two modifications. First, \bar{U} in (5.3) becomes zero in the current case because $\rho = 1$. Second, U^n has the representation (6.6) instead of (4.9). In Theorem 5, to prove $\bar{U}^n \Rightarrow 0$ as $n \rightarrow \infty$ we used the fact

$$\frac{1}{n} \sup_{0 \leq t \leq T} |\hat{G}(nt)| \Rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

which is proved in (5.6). Here, we need

$$\frac{1}{n} \sup_{0 \leq t \leq T} \left| A^n(t) - \alpha \int_0^t (X^n(s))^+ ds \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

which holds because of Lemma 7. \square

A A continuous map

Let $K \in \mathbb{N}$ be a fixed positive integer. Given functions $h_1 : \mathbb{R}^{K+1} \rightarrow \mathbb{R}$, $h_2 : \mathbb{R}^{K+1} \rightarrow \mathbb{R}^K$ and $g : \mathbb{R} \rightarrow \mathbb{R}^K$, we wish to define a map $\Upsilon : \mathbb{D}^{K+1} \rightarrow \mathbb{D}^{K+1}$. For each $y = (y_1, y_2) \in \mathbb{D}^{K+1}$ with $y_1(t) \in \mathbb{R}$ and $y_2(t) \in \mathbb{R}^K$ for $t \geq 0$, $\Upsilon(y)$ is defined to be any $x = (x_1, x_2) \in \mathbb{D}^{K+1}$ with $x_1(t) \in \mathbb{R}$ and $x_2(t) \in \mathbb{R}^K$ for $t \geq 0$ that satisfies

$$x_1(t) = y_1(t) + \int_0^t h_1(x(s)) ds, \quad (\text{A.1})$$

$$x_2(t) = y_2(t) + \int_0^t h_2(x(s)) ds + g(x_1(t)) \quad (\text{A.2})$$

for $t \geq 0$. We assume that h_1 , h_2 and g are Lipschitz continuous. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ with $d, m \in \mathbb{N}$, it is said to be *Lipschitz continuous* with Lipschitz constant $c > 0$ if

$$|f(u) - f(v)| \leq c|u - v| \quad \text{for } u, v \in \mathbb{R}^d$$

(recall that $|u| = \max_{1 \leq k \leq d} |u_k|$ denotes the maximum norm of u). The function f is said to be *positively homogeneous* if

$$f(au) = af(u) \quad \text{for any } a > 0 \text{ and } u \in \mathbb{R}^d.$$

Given $d \in \mathbb{N}$, $x \in \mathbb{D}^d$ and $T > 0$, set $\|x\|_T = \sup_{0 \leq t \leq T} |x(t)|$.

The following lemma establishes the existence and the continuity of the map Υ .

Lemma 9. *Assume that h_1 , h_2 and g are Lipschitz continuous. (a) For each $y = (y_1, y_2) \in \mathbb{D}^{K+1}$ with $y_1(t) \in \mathbb{R}$ and $y_2(t) \in \mathbb{R}^K$ for $t \geq 0$, there exists a unique $x = (x_1, x_2) \in \mathbb{D}^{K+1}$ with $x_1(t) \in \mathbb{R}$ and $x_2(t) \in \mathbb{R}^K$ for $t \geq 0$ that satisfies (A.1) and (A.2). (b) The map $\Upsilon : \mathbb{D}^{K+1} \rightarrow \mathbb{D}^{K+1}$ is Lipschitz continuous in the sense that for each $T > 0$, there exists a constant $C_T > 0$ such that*

$$\|\Upsilon(y) - \Upsilon(\tilde{y})\|_T \leq C_T \|y - \tilde{y}\|_T \quad \text{for any } y, \tilde{y} \in \mathbb{D}^{K+1}.$$

(c) *The map Υ is continuous when the domain \mathbb{D}^{K+1} and the range \mathbb{D}^{K+1} are both endowed with the Skorohod J_1 topology. (d) If, in addition, h_1 , h_2 and g are assumed to be positively homogeneous, then the map Υ is positively homogeneous in the sense that*

$$\Upsilon(ay) = a\Upsilon(y) \quad \text{for each } a > 0 \text{ and each } y \in \mathbb{D}^{K+1}.$$

Proof. Assume that h_1 , h_2 and g are Lipschitz continuous with Lipschitz constant $c > 0$. Let $y = (y_1, y_2) \in \mathbb{D}^{K+1}$ be given. Let $T > 0$ be fixed for the moment. Define $x^0 = y$ and for each $n \in \mathbb{Z}_+$, let $x^{n+1} = (x_1^{n+1}, x_2^{n+1})$ be defined via

$$x_1^{n+1}(t) = y_1(t) + \int_0^t h_1(x^n(s)) ds,$$

$$x_2^{n+1}(t) = y_2(t) + \int_0^t h_2(x^n(s)) ds + g(x_1^{n+1}(t))$$

for $t \in [0, T]$. Setting

$$X^{(n)}(t) = \|x^{n+1} - x^n\|_t,$$

because

$$\begin{aligned} x_2^{n+1}(t) - x_2^n(t) &= \int_0^t (h_2(x^n(s)) - h_2(x^{n-1}(s))) ds \\ &\quad + g\left(y_1(t) + \int_0^t h_1(x^n(s)) ds\right) - g\left(y_1(t) + \int_0^t h_1(x^{n-1}(s)) ds\right) \end{aligned}$$

for $t \in [0, T]$, one has

$$X^{(n+1)}(t) \leq (c + c^2) \int_0^t X^{(n)}(s) ds \quad \text{for } t \in [0, T].$$

Then, by Lemma 11.3 in Mandelbaum et al. (1998),

$$X^{n+1}(t) \leq (c + c^2) \frac{T^n}{n!} \sup_{0 \leq s \leq t} X^{(0)}(s) \quad \text{for } t \in [0, T].$$

Therefore, similar to (11.22) in Mandelbaum et al. (1998), $\{x^n, n \in \mathbb{N}\}$ is a Cauchy sequence under the uniform norm $\|\cdot\|_T$. Since $(\mathbb{D}([0, T], \mathbb{R}^{K+1}), \|\cdot\|_T)$ is a complete metric space (being a closed subset of the Banach space of bounded functions defined from $[0, T]$ into \mathbb{R}^{K+1} and endowed with the uniform norm), $\{x^n, n \in \mathbb{N}\}$ has a limit x that is in $\mathbb{D}([0, T], \mathbb{R}^{K+1})$. One can check that x satisfies (A.1) and (A.2) for $t \in [0, T]$. This proves the existence of the map Υ from $\mathbb{D}([0, T], \mathbb{R}^{K+1})$ to $\mathbb{D}([0, T], \mathbb{R}^{K+1})$.

Now we prove that the map from $\mathbb{D}([0, T], \mathbb{R}^{K+1})$ to $\mathbb{D}([0, T], \mathbb{R}^{K+1})$ is Lipschitz continuous with respect to the uniform norm. Assume that $y, \tilde{y} \in \mathbb{D}([0, T], \mathbb{R}^{K+1})$. Let $\Upsilon(y)$ be any solution x such that x and y satisfy (A.1) and (A.2) on $[0, T]$. Similarly, let $\Upsilon(\tilde{y})$ be any solution associated with \tilde{y} . Setting $x = (x_1, x_2) = \Upsilon(y)$ and $\tilde{x} = (\tilde{x}_1, \tilde{x}_2) = \Upsilon(\tilde{y})$, then for any $t \in [0, T]$,

$$\begin{aligned} |x_1(t) - \tilde{x}_1(t)| &\leq |y(t) - \tilde{y}(t)| + c \int_0^t |\Upsilon(y)(s) - \Upsilon(\tilde{y})(s)| ds, \\ |x_2(t) - \tilde{x}_2(t)| &\leq (1 + c)|y(t) - \tilde{y}(t)| + (c + c^2) \int_0^t |\Upsilon(y)(s) - \Upsilon(\tilde{y})(s)| ds. \end{aligned}$$

Hence,

$$|\Upsilon(y)(t) - \Upsilon(\tilde{y})(t)| \leq (1 + c)|y(t) - \tilde{y}(t)| + (c + c^2) \int_0^t |\Upsilon(y)(s) - \Upsilon(\tilde{y})(s)| ds \quad \text{for } t \in [0, T].$$

By Corollary 11.2 in Mandelbaum et al. (1998)

$$\|\Upsilon(y) - \Upsilon(\tilde{y})\|_T \leq (1 + c)\|y - \tilde{y}\|_T \exp((c + c^2)T).$$

Hence, Υ is Lipschitz continuous, which implies part (b) of the lemma. The Lipschitz continuity of Υ as a map from $\mathbb{D}([0, T], \mathbb{R}^{K+1})$ to $\mathbb{D}([0, T], \mathbb{R}^{K+1})$ shows that it is well defined on $[0, T]$. Since $T > 0$ is arbitrary, Υ as a map from \mathbb{D}^{K+1} to \mathbb{D}^{K+1} is well defined. This proves part (a) of the lemma.

Next we prove the continuity of Υ provided that \mathbb{D}^{K+1} is endowed with the Skorohod J_1 topology (see, for example, Section 3 of Whitt (2002)). Consider a sequence $\{y^n, n \in \mathbb{N}\}$ and y in \mathbb{D}^{K+1} such that $y^n \rightarrow y$ as $n \rightarrow \infty$. Let $x^n = (x_1^n, x_2^n) = \Upsilon(y^n)$ and $x = (x_1, x_2) = \Upsilon(y)$. Note that since $x \in \mathbb{D}^{K+1}$ there exists $M > 0$ such that

$$\|\Upsilon(y)\|_T < M. \quad (\text{A.3})$$

Let Λ be the set of strictly increasing functions $\lambda : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $\lambda(0) = 0$, $\lim_{t \rightarrow \infty} \lambda(t) = \infty$, and

$$\gamma(\lambda) = \sup_{0 \leq s < t} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right| < \infty.$$

Since $y^n \rightarrow y$ as $n \rightarrow \infty$ in the J_1 topology on \mathbb{D}^{K+1} , it follows from Proposition 3.5.3 of Ethier and Kurtz (1986) that there exists a sequence $\{\lambda^n, n \in \mathbb{N}\} \subset \Lambda$ such that

$$\lim_{n \rightarrow \infty} \gamma(\lambda^n) = 0 \quad (\text{A.4})$$

and for each $T > 0$

$$\lim_{n \rightarrow \infty} \|y^n(\cdot) - y(\lambda^n(\cdot))\|_T = 0. \quad (\text{A.5})$$

For each $\lambda^n \in \Lambda$, $\lambda^n(t)$ is Lipschitz continuous in t . Hence, it is differentiable almost everywhere in t with respect to the Lebesgue measure. Furthermore, it follows from (3.5.5) of Ethier and Kurtz (1986) that when λ^n is differential at time t , its derivative $\dot{\lambda}^n(t)$ satisfies

$$|\dot{\lambda}^n(t) - 1| \leq \gamma(\lambda^n). \quad (\text{A.6})$$

Note that, for $i = 1, 2$

$$\int_0^{\lambda^n(t)} h_i(x(s)) ds = \int_0^t h_i(x(\lambda^n(s))) \dot{\lambda}^n(s) ds. \quad (\text{A.7})$$

By (A.1) and (A.7)

$$\begin{aligned} x_1(\lambda^n(t)) &= y_1(\lambda^n(t)) + \int_0^{\lambda^n(t)} h_1(x(s)) ds \\ &= y_1(\lambda^n(t)) + \int_0^t h_1(x(\lambda^n(s))) \dot{\lambda}^n(s) ds \\ &= y_1(\lambda^n(t)) + \int_0^t h_1(x(\lambda^n(s))) ds - \int_0^t h_1(x(\lambda^n(s))) (1 - \dot{\lambda}^n(s)) ds. \end{aligned} \quad (\text{A.8})$$

Similarly, by (A.2) and (A.7)

$$\begin{aligned} x_2(\lambda^n(t)) &= y_2(\lambda^n(t)) + \int_0^t h_2(x(\lambda^n(s))) ds - \int_0^t h_2(x(\lambda^n(s))) (1 - \dot{\lambda}^n(s)) ds \\ &\quad + g(x_1(\lambda^n(t))). \end{aligned} \tag{A.9}$$

By (A.1) and (A.8),

$$\begin{aligned} &|x_1^n(t) - x_1(\lambda^n(t))| \\ &\leq |y_1^n(t) - y_1(\lambda^n(t))| + \int_0^t |h_1(x^n(s)) - h_1(x(\lambda^n(s)))| ds \\ &\quad + \int_0^t |h_1(x(\lambda^n(s))) - h_1(0)| |1 - \dot{\lambda}^n(s)| ds + \int_0^t |h_1(0)| |1 - \dot{\lambda}^n(s)| ds \\ &\leq |y^n(t) - y(\lambda^n(t))| + c \int_0^t |x^n(s) - x(\lambda^n(s))| ds \\ &\quad + c \int_0^t |x(\lambda^n(s))| |1 - \dot{\lambda}^n(s)| ds + |h_1(0)| \int_0^t |1 - \dot{\lambda}^n(s)| ds. \end{aligned} \tag{A.10}$$

By (A.2), (A.9) and (A.10)

$$\begin{aligned} &|x_2^n(t) - x_2(\lambda^n(t))| \\ &\leq |y_2^n(t) - y_2(\lambda^n(t))| + \int_0^t |h_2(x^n(s)) - h_2(x(\lambda^n(s)))| ds + |g(x_1^n(t)) - g(x_1(\lambda^n(t)))| \\ &\quad + \int_0^t |h_2(x(\lambda^n(s))) - h_2(0)| |1 - \dot{\lambda}^n(s)| ds + \int_0^t |h_2(0)| |1 - \dot{\lambda}^n(s)| ds \\ &\leq |y^n(t) - y(\lambda^n(t))| + c \int_0^t |x^n(s) - x(\lambda^n(s))| ds + c |x_1^n(t) - x_1(\lambda^n(t))| \\ &\quad + c \int_0^t |x(\lambda^n(s))| |1 - \dot{\lambda}^n(s)| ds + |h_2(0)| \int_0^t |1 - \dot{\lambda}^n(s)| ds \\ &\leq (1 + c) |y^n(t) - y(\lambda^n(t))| + (c + c^2) \int_0^t |x^n(s) - x(\lambda^n(s))| ds \\ &\quad + (c + c^2) \int_0^t |x(\lambda^n(s))| |1 - \dot{\lambda}^n(s)| ds + (|h_2(0)| + c |h_1(0)|) \int_0^t |1 - \dot{\lambda}^n(s)| ds. \end{aligned} \tag{A.11}$$

Then (A.10) and (A.11) yield

$$\begin{aligned}
& |\Upsilon(y^n)(t) - \Upsilon(y)(\lambda^n(t))| \\
& \leq (1+c)|y^n(t) - y(\lambda^n(t))| + (c+c^2) \int_0^t |\Upsilon(y^n)(s)ds - \Upsilon(y)(\lambda^n(s))| ds \\
& \quad + (c+c^2) \int_0^t |1 - \dot{\lambda}^n(s)| |\Upsilon(y)(\lambda^n(s))| ds + (|h_2(0)| + c|h_1(0)|) \int_0^t |1 - \dot{\lambda}^n(s)| ds.
\end{aligned} \tag{A.12}$$

It follows from (A.3), (A.4), (A.6) and the dominated convergence theorem that

$$\int_0^t |1 - \dot{\lambda}^n(s)| |\Upsilon(y)(\lambda^n(s))| ds \rightarrow 0 \quad \text{as } n \rightarrow \infty. \tag{A.13}$$

Given $\delta > 0$, by (A.4), (A.6) and (A.13), for n large enough

$$(c+c^2) \int_0^T |1 - \dot{\lambda}^n(s)| |\Upsilon(y)(\lambda^n(s))| ds + (|h_2(0)| + c|h_1(0)|) \int_0^T |1 - \dot{\lambda}^n(s)| ds < \frac{\delta}{2},$$

and by (A.5)

$$(1+c) \|y^n(\cdot) - y(\lambda^n(\cdot))\|_T < \frac{\delta}{2}.$$

By Corollary 11.2 in Mandelbaum et al. (1998) and (A.12),

$$\|\Upsilon(y^n)(\cdot) - \Upsilon(y)(\lambda^n(\cdot))\|_T \leq \delta \exp((c+c^2)T)$$

for large enough n . Thus, for each $T > 0$,

$$\lim_{n \rightarrow \infty} \|\Upsilon(y^n)(\cdot) - \Upsilon(y)(\lambda^n(\cdot))\|_T = 0.$$

Hence, $\Upsilon(y^n) \rightarrow \Upsilon(y)$ as $n \rightarrow \infty$ in \mathbb{D}^{K+1} in the J_1 topology. This implies part (c) of the lemma.

To prove part (d) of the lemma, for $y \in \mathbb{D}^{K+1}$, assume that x and y satisfy (A.1) and (A.2). Then, for $a > 0$, one can check that ax and ay also satisfy (A.1) and (A.2) because of the positive homogeneity of h_1 , h_2 and g . Therefore, $\Upsilon(ay) = a\Upsilon(y)$. \square

B Proofs of Lemmas 2–3 and Theorem 3

This section is devoted to proving Lemmas 2–3 and Theorem 3. We first present two lemmas.

The first lemma is an immediate result by Proposition 4 of Dai and He (2009). It proves the stochastic boundedness of the virtual waiting time process (see the paragraph prior to Theorem 3 for its definition) after proper scaling.

Lemma 10. *Under the conditions of Theorem 1, for any $T > 0$,*

$$\lim_{a \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P} \left[\sup_{0 \leq t \leq T} \sqrt{n} W^n(t) > a \right] = 0, \quad (\text{B.1})$$

which implies

$$W^n \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (\text{B.2})$$

For $t \geq 0$, let

$$\zeta^n(t) = \inf\{s \geq 0 : s + W^n(s) > t\}. \quad (\text{B.3})$$

Since $s + W^n(s) \leq t$ for all $s < \zeta^n(t)$, each customer arriving before time $\zeta^n(t)$ cannot be waiting in queue at time t (see Lemmas 13 and 14 of Dai and He (2009) for a detailed explanation); similarly, since $s + W^n(s) > t$ for all $s > \zeta^n(t)$, a customer who arrives after time $\zeta^n(t)$ cannot be in service at t . So $\zeta^n(t)$ is a crucial epoch with respect to the queue length at time t . The next lemma concerns the process $\zeta^n = \{\zeta^n(t), t \geq 0\}$.

Lemma 11. *Under the conditions of Theorem 1, $\zeta^n \in \mathbb{D}$ is nondecreasing for each $n \in \mathbb{N}$, and*

$$\zeta^n \Rightarrow \zeta \quad \text{as } n \rightarrow \infty,$$

where $\zeta(t) = t$ for $t \geq 0$ is the identity function on \mathbb{R}_+ .

Proof. First note that

$$\zeta^n(t) + W^n(\zeta^n(t)) \geq t \quad \text{for } t \geq 0, \quad (\text{B.4})$$

because W^n is right-continuous.

Next, we prove that ζ^n is nondecreasing in t . Suppose on the contrary that for some $0 \leq s < t$, we have $\zeta^n(t) < \zeta^n(s)$. This implies by (B.3) that for any $\zeta^n(t) < u < \zeta^n(s)$,

$$t < u + W^n(u) \leq s,$$

leading to a contradiction.

Now we prove that $\zeta^n \in \mathbb{D}$, i.e., ζ^n is right-continuous on $[0, \infty)$ and has left limits on $(0, \infty)$. Since $\zeta^n(t) \leq t$ by (B.3) and ζ^n is nondecreasing, $\zeta^n(t-)$ exists for each $t > 0$; therefore, ζ^n has left limits on $(0, \infty)$. To prove right-continuity, fix $\varepsilon > 0$ and $t \geq 0$. We have

$$\zeta^n(t) + \varepsilon + W^n(\zeta^n(t) + \varepsilon) > t + \delta \quad \text{for some } \delta > 0,$$

so that $\zeta^n(t + \delta') \leq \zeta^n(t + \delta) \leq \zeta^n(t) + \varepsilon$ for $0 < \delta' \leq \delta$. Hence, ζ^n is right-continuous at t , proving $\zeta^n \in \mathbb{D}$.

Finally, we prove the convergence. By (B.4) and the fact $\zeta^n(t) \leq t$, for any $T > 0$,

$$\sup_{0 \leq t \leq T} |t - \zeta^n(t)| \leq \sup_{0 \leq t \leq T} W^n(\zeta^n(t)) \leq \sup_{0 \leq t \leq T} W^n(t).$$

Then $\zeta^n \Rightarrow \zeta$ as $n \rightarrow \infty$ follows from (B.2). □

Proof of Lemma 2. Fix $T > 0$, and restrict $t \in [0, T]$. Since each customer arriving before time $\zeta^n(t)$ will either have entered service or abandoned the system by time t , we have $(X^n(t))^+ \leq E^n(t) - E^n(\zeta^n(t)) + \Delta^n(\zeta^n(t))$ where $\Delta^n(t) = E^n(t) - E^n(t-)$ is the number of customers who arrive (exactly) at time t . Because $\zeta^n(t) \leq t$ by (B.3), we have

$$\sup_{0 \leq t \leq T} \Delta^n(\zeta^n(t)) \leq \|\Delta^n\|_T,$$

for $\|\Delta^n\|_T = \sup_{0 \leq t \leq T} \Delta^n(t)$; thus,

$$(X^n(t))^+ \leq E^n(t) - E^n(\zeta^n(t)) + \|\Delta^n\|_T. \quad (\text{B.5})$$

Similarly, because a customer who arrives during $(\zeta^n(t), t]$ will either be waiting in queue at time t or has abandoned the system by t , one has

$$(X^n(t))^+ \geq E^n(t) - E^n(\zeta^n(t)) - (A^n(t) - A^n(\zeta^n(t))). \quad (\text{B.6})$$

Let $\{\psi^0(i), i \in \mathbb{N}\}$ be a sequence of iid K -dimensional random vectors such that for $k = 1, \dots, K$, the probability that $\psi^0(i) = e^k$ is p_k ; it is used to indicate the initial service phase of each customer (see the first paragraph in Section 4.2). Write

$$\Psi^0(N) = \sum_{i=1}^N \psi^0(i) \quad \text{and} \quad \hat{\Psi}^0(N) = \Psi^0(N) - pN.$$

Because the customers who arrive before time $\zeta^n(t)$ cannot be waiting in queue at time t (they have either abandoned the system or started service), for $k = 1, \dots, K$,

$$\begin{aligned} Q_k^n(t) &\leq \Psi_k^0((X^n(0))^+ + E^n(t)) - \Psi_k^0((X^n(0))^+ + E^n(\zeta^n(t)) - \|\Delta^n\|_T) \\ &= \hat{\Psi}_k^0((X^n(0))^+ + E^n(t)) - \hat{\Psi}_k^0((X^n(0))^+ + E^n(\zeta^n(t)) - \|\Delta^n\|_T) \\ &\quad + p_k(E^n(t) - E^n(\zeta^n(t)) + \|\Delta^n\|_T). \end{aligned} \quad (\text{B.7})$$

Similarly, the customers who arrive during $(\zeta^n(t), t]$ cannot get into service by time t . Then

$$\begin{aligned} Q_k^n(t) + (A^n(t) - A^n(\zeta^n(t))) &\geq \Psi_k^0((X^n(0))^+ + E^n(t)) - \Psi_k^0((X^n(0))^+ + E^n(\zeta^n(t))) \\ &= \hat{\Psi}_k^0((X^n(0))^+ + E^n(t)) - \hat{\Psi}_k^0((X^n(0))^+ + E^n(\zeta^n(t))) \\ &\quad + p_k(E^n(t) - E^n(\zeta^n(t))). \end{aligned} \quad (\text{B.8})$$

Combining (B.5)–(B.8), we have

$$\Lambda_k^n(t) \leq Q_k^n(t) - p_k(X^n(t))^+ \leq \Pi_k^n(t), \quad (\text{B.9})$$

where

$$\begin{aligned} \Lambda_k^n(t) &= \hat{\Psi}_k^0((X^n(0))^+ + E^n(t)) - \hat{\Psi}_k^0((X^n(0))^+ + E^n(\zeta^n(t))) \\ &\quad - (A^n(t) - A^n(\zeta^n(t))) - p_k \|\Delta^n\|_T, \\ \Pi_k^n(t) &= \hat{\Psi}_k^0((X^n(0))^+ + E^n(t)) - \hat{\Psi}_k^0((X^n(0))^+ + E^n(\zeta^n(t)) - \|\Delta^n\|_T) \\ &\quad + p_k(\|\Delta^n\|_T + A^n(t) - A^n(\zeta^n(t))). \end{aligned}$$

Let $\tilde{\Psi}^{0,n}(t) = \tilde{\Psi}^0(\lfloor nt \rfloor)/\sqrt{n}$, $\|\tilde{\Delta}^n\|_T = \|\Delta^n\|_T/\sqrt{n}$, and $\|\bar{\Delta}^n\|_T = \|\Delta^n\|_T/n$. Rewriting (B.9) using diffusion scaling one has

$$\tilde{\Lambda}_k^n(t) \leq \frac{1}{\sqrt{n}} (Q_k^n(t) - p_k(X^n(t))^+) \leq \tilde{\Pi}_k^n(t), \quad (\text{B.10})$$

where

$$\begin{aligned} \tilde{\Lambda}_k^n(t) &= \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(t)) - \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(\zeta^n(t))) \\ &\quad - (\tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t))) - p_k \|\tilde{\Delta}^n\|_T, \\ \tilde{\Pi}_k^n(t) &= \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(t)) - \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(\zeta^n(t)) - \|\bar{\Delta}^n\|_T) \\ &\quad + p_k (\|\tilde{\Delta}^n\|_T + \tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t))). \end{aligned}$$

Next, we show that $\tilde{\Lambda}_k^n \Rightarrow 0$ and $\tilde{\Pi}_k^n \Rightarrow 0$ as $n \rightarrow \infty$, which, together with (B.10), will lead to (3.20). Using (2.2), we have

$$\|\tilde{\Delta}^n\|_T \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (\text{B.11})$$

Lemma 11, (6.8), Theorem 3.9 in Billingsley (1999) and the random-time-change theorem (see the lemma on page 151 of Billingsley (1999)) yield

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t)) \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (\text{B.12})$$

By Theorem 5, Lemma 11 and the random-time-change theorem,

$$(\bar{X}^n(0))^+ \Rightarrow 0 \quad \text{and} \quad \bar{E}^n(\zeta^n(\cdot)) \Rightarrow \bar{E} \quad \text{as } n \rightarrow \infty. \quad (\text{B.13})$$

Since $\tilde{\Psi}^{0,n} \Rightarrow \tilde{\Psi}^0$ where $\tilde{\Psi}^0$ is a K -dimensional Brownian motion, by Theorem 5, (B.11), (B.13) and the random-time-change theorem

$$\sup_{0 \leq t \leq T} \left| \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(t)) - \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(\zeta^n(t)) - \|\bar{\Delta}^n\|_T) \right| \Rightarrow 0, \quad (\text{B.14})$$

$$\sup_{0 \leq t \leq T} \left| \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(t)) - \tilde{\Psi}_k^{0,n}((\bar{X}^n(0))^+ + \bar{E}^n(\zeta^n(t))) \right| \Rightarrow 0 \quad (\text{B.15})$$

as $n \rightarrow \infty$. We deduce from (B.11)–(B.15) that $\tilde{\Lambda}_k^n \Rightarrow 0$ and $\tilde{\Pi}_k^n \Rightarrow 0$ as $n \rightarrow \infty$. \square

Proof of Theorem 3. Since all customers arriving prior to time $t \geq 0$ will have either got into service or abandoned the system by time $t + W^n(t)$ (see Lemmas 13 and 14 of Dai and He (2009)), then

$$(X^n(t + W^n(t)))^+ \leq E^n(t + W^n(t)) - E^n(t).$$

For a customer who arrives during $(t, t + W^n(t)]$, he can possibly be waiting in queue at time $t + W^n(t)$, or have abandoned the system by $t + W^n(t)$, or starts his service (exactly) at $t + W^n(t)$. Therefore,

$$E^n(t + W^n(t)) - E^n(t) \leq (X^n(t + W^n(t)))^+ + A^n(t + W^n(t)) - A^n(t) + \Delta_D^n(t + W^n(t)),$$

where $\Delta_D^n(t) = D^n(t) - D^n(t-)$ is the number of service completions (exactly) at time t . Then by (2.3) and (6.3),

$$\begin{aligned} 0 &\leq \frac{1}{\sqrt{n}} \lambda^n W^n(t) - (\tilde{X}^n(t + W^n(t)))^+ + \tilde{E}^n(t + W^n(t)) - \tilde{E}^n(t) \\ &\leq \tilde{A}^n(t + W^n(t)) - \tilde{A}^n(t) + \tilde{\Delta}_D^n(t + W^n(t)), \end{aligned}$$

where $\tilde{\Delta}_D^n(t) = \Delta_D^n(t)/\sqrt{n}$. This leads to

$$\begin{aligned} \left| \mu \sqrt{n} W^n(t) - (\tilde{X}^n(t + W^n(t)))^+ \right| &\leq \left| \left(\frac{1}{n} \lambda^n - \mu \right) \sqrt{n} W^n(t) \right| + \left| \tilde{E}^n(t + W^n(t)) - \tilde{E}^n(t) \right| \\ &\quad + \left| \tilde{A}^n(t + W^n(t)) - \tilde{A}^n(t) \right| + \tilde{\Delta}_D^n(t + W^n(t)). \end{aligned} \tag{B.16}$$

Next we show that all terms on the right-hand side of (B.16) converge weakly to zero as $n \rightarrow \infty$. Using (2.1) and (B.1), we get

$$\left| \left(\frac{1}{n} \lambda^n - \mu \right) \sqrt{n} W^n \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \tag{B.17}$$

For any $T > 0$, by (2.2) and (B.2),

$$\sup_{0 \leq t \leq T} \left| \tilde{E}^n(t + W^n(t)) - \tilde{E}^n(t) \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \tag{B.18}$$

By (6.8) and (B.2),

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t + W^n(t)) - \tilde{A}^n(t) \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \tag{B.19}$$

Set $\tilde{D}^n(t) = (D^n(t) - n\mu t)/\sqrt{n}$. It follows from (4.8) that $\tilde{D}^n \Rightarrow \tilde{D}$ as $n \rightarrow \infty$, where

$$\tilde{D}(t) = -e' \tilde{M}(t) + e' R \int_0^t \tilde{Z}(s) ds.$$

Since \tilde{D} is continuous almost surely, using (B.2) again, we have

$$\sup_{0 \leq t \leq T} \left| \tilde{\Delta}_D^n(t + W^n(t)) \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \tag{B.20}$$

Combining (B.16)–(B.20), we deduce that

$$\sup_{0 \leq t \leq T} \left| \mu \sqrt{n} W^n(t) - (\tilde{X}^n(t + W^n(t)))^+ \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (\text{B.21})$$

By (3.12), the process \tilde{X} is continuous almost surely; then so is $(\tilde{X})^+$. Because $s + W^n(s) \leq t + W^n(t)$ for $0 \leq s \leq t$ (see Lemma 13 of Dai and He (2009)) and the process $(\tilde{X})^+$ is continuous almost surely, we have

$$(\tilde{X}^n(\cdot + W^n(\cdot)))^+ \Rightarrow (\tilde{X})^+ \quad \text{as } n \rightarrow \infty \quad (\text{B.22})$$

by (B.2) and the random-time-change theorem. By (B.21), (B.22) and the convergence-together theorem (see Theorem 3.1 of Billingsley (1999)), $\sqrt{n}W^n \Rightarrow (\tilde{X})^+/\mu$ as $n \rightarrow \infty$. \square

Proof of Lemma 3. Recall that any customer who is waiting in queue at time $t \geq 0$ must arrive at the system during $[\zeta^n(t), t]$ (see (B.3) and the discussion therein), and must leave the queue (either goes into service or abandons the system) by time $t + W^n(t)$ (see Lemmas 13 and 14 of Dai and He (2009)). This implies

$$A_Q^n(t) \leq A^n(t + W^n(t)) - A^n(\zeta^n(t)-).$$

It follows that for any $T > 0$,

$$\sup_{0 \leq t \leq T} \tilde{A}_Q^n(t) \leq \sup_{0 \leq t \leq T} \left| \tilde{A}^n(t + W^n(t)) - \tilde{A}^n(\zeta^n(t)) \right| + \sup_{0 \leq t \leq T} \left| \tilde{A}^n(\zeta^n(t)) - \tilde{A}^n(\zeta^n(t)-) \right|.$$

By (B.12) and (B.19),

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t + W^n(t)) - \tilde{A}^n(\zeta^n(t)) \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

By (6.8) and the fact $\zeta^n(t) \leq t$,

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(\zeta^n(t)) - \tilde{A}^n(\zeta^n(t)-) \right| \leq \sup_{0 \leq t \leq T} \left| \tilde{A}^n(t) - \tilde{A}^n(t-) \right| \Rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Hence, (3.25) holds. \square

Acknowledgement

This research is supported in part by NSF grants CMMI-0727400 and CMMI-0825840, and by an IBM Faculty Award.

References

- BILLINGSLEY, P. (1999). *Convergence of Probability Measures*. 2nd ed. Wiley, New York.
- BOROVKOV, A. (1967). On limit laws for service processes in multi-channel systems. *Siberian Mathematical Journal*, **8** 746–763.
- DAI, J. G. and HE, S. (2009). Customer abandonment in many-server queues. Preprint, URL http://www2.isye.gatech.edu/~dai/publications/draft_daiHe09.pdf.
- DAI, J. G. and TEZCAN, T. (2005). State space collapse in many-server diffusion limits of parallel server systems. Preprint, URL http://www2.isye.gatech.edu/~dai/publications/draft_daiTezcan05.pdf.
- DAI, J. G. and TEZCAN, T. (2008). Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems*, **59** 95–134.
- ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.
- GAMARNIK, D. and MOMČILOVIĆ, P. (2008). Steady-state analysis of a multi-server queue in the Halfin-Whitt regime. *Advances in Applied Probability*, **40** 548–577.
- GARNETT, O., MANDELBAUM, A. and REIMAN, M. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management*, **4** 208–227.
- HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, **29** 567–588.
- IGLEHART, D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, **2** 150–177.
- JELENKOVIĆ, P., MANDELBAUM, A. and MOMČILOVIĆ, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Systems*, **47** 53–69.
- JOHNSON, D. P. (1983). *Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks*. Ph.D. thesis, University of Wisconsin.
- KANG, W. N. and RAMANAN, K. (2008). Fluid limits of many-server queues with reneging. Preprint, URL http://www.math.cmu.edu/users/kramanan/research/sub_reneging.pdf.
- KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

- KASPI, H. and RAMANAN, K. (2007). Law of large numbers limits for many server queues. Preprint, URL http://www.math.cmu.edu/users/kramanan/research/Many_server_fluid.pdf.
- KIEFER, J. and WOLFOWITZ, J. (1955). On the theory of queues with many servers. *Transactions of the American Mathematical Society*, **78** 1–18.
- LATOUCHE, G. and RAMASWAMI, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Probability, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
- MANDELBAUM, A., MASSEY, W. A. and REIMAN, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Systems*, **30** 149–201.
- MANDELBAUM, A. and MOMČILOVIĆ, P. (2009). Queues with many servers and impatient customers. Preprint, URL <http://http://iew3.technion.ac.il/serveng/References/MM0309.pdf>.
- MANDELBAUM, A. and PATS, G. (1998). State-dependent stochastic networks. part I: approximations and applications with continuous diffusion limits. *Annals of Applied Probability*, **8** 569–646.
- PANG, G., TALREJA, R. and WHITT, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, **4** 193–267.
- PUHALSKII, A. A. and REED, J. E. (2008). On many-server queues in heavy traffic. Preprint, URL http://pages.stern.nyu.edu/~jreed/Papers/puhalskii_reed.pdf.
- PUHALSKII, A. A. and REIMAN, M. I. (2000). The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Advances in Applied Probability*, **32** 564–595. Correction: **36**, 971 (2004).
- REED, J. E. (2007). The $G/GI/N$ queue in the Halfin-Whitt regime I: infinite server queue system equations. *Annals of Applied Probability*. To appear.
- REIMAN, M. I. (1984). Open queueing networks in heavy traffic. *Mathematics of Operations Research*, **9** 441–458.
- ROGERS, L. C. G. and WILLIAMS, D. (2000). *Diffusions, Markov processes, and Martingales Volume 2: Itô Calculus*. 2nd ed. Cambridge University Press, Cambridge, UK.
- STONE, C. (1963). Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois Journal of Mathematics*, **7** 638–660.

- TEZCAN, T. (2006). *State Space Collapse in Many Server Diffusion Limits of Parallel Server Systems and Applications*. Ph.D. thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology.
- TEZCAN, T. and DAI, J. G. (2008). Dynamic control of N-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. *Operations Research*. To appear.
- WHITT, W. (2002). *Stochastic-process Limits*. Springer, New York.
- WHITT, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, **50** 1449–1461.
- WHITT, W. (2005). Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research*, **30** 1–27.
- WHITT, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research*, **54** 37–54.
- WHITT, W. (2007). Proofs of the martingale FCLT. *Probability Surveys*, **4** 268–302.
- ZHANG, J. (2009). *Limited Processor Sharing Queues and Multi-server Queues*. Ph.D. thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology.