Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

On the Analysis of G-Queues Under Heavy Traffic

Saul C. Leite and Marcelo D. Fragoso

***Abstract*— It is a *fait accompli* that heavy traffic analysis is a very powerful technique, which allows us to model the number of customers of a queueing system via a reflected diffusion (or a reflected Lévy process). In this paper, we derive heavy traffic type theorems for queueing systems that receive either Markov modulated or independent positive and negative arrivals with general service and inter-arrival time processes.**

Keywords: Queueing, Heavy Traffic Analysis, G-Queues.

I. Introduction

Consider a queueing system with two types of arrivals, negative and positive. The positive arrivals corresponds to regular customers that enter the system in order to receive service. A negative customer is a signal to the system indicating that it must remove a regular customer from the queue. These queueing systems are called G-queues. Since their introduction in [8], these models have been extensively studied (e.g., [6], [3], [11], [16], [9], [7], [12], [13], [2], [5]) and are motivated by a series of practical applications. For example, a negative arrival may represent a signal to delete some transaction in a distributed database system [8], packet loss in Internet traffic [6] or inhibitory signals in mathematical models of neurons [8].

In this paper, we derive a model for G-queues under heavy traffic. The main idea of heavy traffic analysis is to approximate the stochastic processes that describe a queueing system by a reflected diffusion, under an appropriate time and space scaling. These approximations simplify the models considerably. Heavy traffic analysis can be employed in situations where the rate of arrivals into the system is close to the rate of departures (hence, the name heavy traffic). However, they are known to give good estimates even for systems under only moderate traffic [14]. A common application is modelling computer systems (e.g., [1]). For a complete account on the subject see [14].

To our knowledge, queues with negative arrivals have not yet been treated under heavy traffic analysis. This kind of model help us understand the general behavior of such queues for *arbitrary inter-arrival and service time distributions* (usually, G-queues are treated only for arrival or departure processes having some kind of Markovian structure) in addition to clarifying the interactions among the model's parameter. Also, it describes the transient evolution of these queues with a simple time-dependent equation, which is otherwise difficult to obtain in such a convenient form.

The layout of the paper is as follows: in the next section, we will introduce the notation and assumptions that will be used throughout the paper. Next, in section III, we will state the heavy traffic theorems for the *number of customers* and *workload process*. In section IV, we will apply these results to approximate some queueing systems and compare it against a computer simulation.

National Laboratory for Scientific Computing (LNCC), Av. Getilio Vargas 333, 25651-075, Quitandinha, Petropolis, RJ, Brazil. Contact email: frag@lncc.br

II. Notation and Assumptions

For this article, we will restrict ourselves to the first come first served (FCFS) queue discipline. Also, we will suppose that negative arrivals will remove customers from the end of the queue. This discipline is usually called RCT discipline (removal of customers from the tail of the queue) [8]. In addition, we assume that a negative customer may only remove positive customers if they are not being served. This kind of customer removal discipline was called "RCT-immune servicing" in [7].

As it is common in heavy traffic models, we consider a "sequence" of queues indexed by the parameter n. When this parameter grows larger, the difference between the rate of customers leaving the queue and the rate of regular arrival gets smaller, tending to what is called a "heavy traffic situation." This is enforced by equation (1).

The notation used here is mainly the one in [14]. Let $\{\Delta_l^{a,n}\}$ and $\{\Delta_l^{r,n}\}$ be stochastic processes denoting the inter-arrival times of positive and negative customers into the queue, respectively. The service times are denoted by $\{\Delta_l^{d,n}\}$.

Assumption 2.1: The random variables $\Delta_l^{\alpha,n}$ are mutually independent for each l and n, and they are identically distributed with means $0 < \bar{\Delta}^{\alpha,n} < \infty$, $\alpha = a, r, d$, for each l. Let $\sigma^{\alpha,n}$ be the coefficient of variation of $\Delta_l^{\alpha,n}$. There are positive constants σ^α and $\bar{\Delta}^\alpha$ such that $\sigma^{\alpha,n} \to \sigma^\alpha$ and $\bar{\Delta}^{\alpha,n} \to \bar{\Delta}^\alpha$, as $n \to \infty$, for $\alpha = a, d, r$. Also, $\{\Delta_l^{a,n}\}$, $\{\Delta_l^{r,n}\}$ and $\{\Delta_l^{d,n}\}$ are independent, and $\{|\Delta^{a,n}|^2, |\Delta^{d,n}|^2, |\Delta^{r,n}|^2; n\}$ is uniformly integrable.

The assumption on the uniform integrability can be replaced by supposing that $\Delta_l^{\alpha,n}$ converges weakly to a stochastic process (as $n \to \infty$) that has mean $\bar{\Delta}^\alpha$ and coefficient of variation σ^α (e.g, Theorem 5.4 in [4]).

Also, we suppose that there exists a constant $b \in \mathbb{R}$ such that

$$\sqrt{n}\left(\frac{1}{\bar{\Delta}^{a,n}} - \frac{1}{\bar{\Delta}^{r,n}} - \frac{1}{\bar{\Delta}^{d,n}}\right) \triangleq b_n \to b. \tag{1}$$

This condition is usually referred to as the "heavy traffic condition." For later use, define $\lambda^{\alpha,n} \triangleq 1/\bar{\Delta}^{\alpha,n}$, $\alpha = a, r, d$.

Let $S^{a,n}(t)$ (resp., $S^{r,n}(t)$) denote $1/n$ times the number of positive (resp., negative) arrivals to the system by time nt, and $S^{d,n}(t)$ denote $1/n$ times the number of service

completions by time nt. Observe that we may write, for $\alpha = a, r$,

$$S^{\alpha,n}(t) = \frac{1}{n}\max\left\{m \in \mathbb{N}_0 : \sum_{l=1}^{m} \Delta_l^{\alpha,n} \leq nt\right\}, \quad (2)$$

$$S^{d,n}(t) = \frac{1}{n}\max\left\{m \in \mathbb{N}_0 : \sum_{l=1}^{m} \Delta_l^{d,n} \leq nt - T^n(t)\right\} \quad (3)$$

where $T^n(t)$ is defined as the total server idle time by time nt.

Observe that the process $x^n(t)$, denoting $1/\sqrt{n}$ times the number of customers in the system by time nt, can be written as

$$\begin{aligned} x^n(t) &= x^n(0) + \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{a,n}(t)} 1 - \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{r,n}(t)} 1 \\ &\quad - \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{d,n}(t)} 1 + z_0^n(t), \end{aligned} \quad (4)$$

where $x^n(0)$ is $1/\sqrt{n}$ times the number of customers in the system at time zero, and $z_0^n(t)$ is $1/\sqrt{n}$ times the number of negative customers that arrived when the queue was empty by time nt. Notice that $z_0^n(t)$ may increase only when $x^n(t) \leq 1/\sqrt{n}$, since it will increase when a negative customer arrives and finds an empty system and when it finds an empty queue but one customer at service.

III. Heavy Traffic Theorems

In this section we will prove the main theorems of the paper. We will begin considering the queue lenght process. Next, we will derive a heavy traffic model for the workload process. And finally, we present the result for a G-queue with arrivals modulated by a Markov chain. It is perhaps noteworthy that weak convergence here refers to convergence in distribution, as in [4].

A. Number of Customers in Queue

Theorem 3.1: Suppose that $x^n(0)$ converges weakly to $x(0)$, and is independent of the inter-arrival and service time processes. With the assumptions described in the previous section, the process $x^n(\cdot)$ converges weakly to the process $x(\cdot)$ taking values

$$\begin{aligned} x(t) &= x(0) + w^a(\lambda^a t) - w^r(\lambda^r t) - w^d(\lambda^d t) \\ &\quad + bt + z(t), \end{aligned} \quad (5)$$

where $w^a(\cdot)$, $w^r(\cdot)$ and $w^d(\cdot)$ are independent Wiener processes with variances $(\sigma^a)^2$, $(\sigma^r)^2$, and $(\sigma^d)^2$, respectively, b is the constant in (1), and $z(\cdot)$ is the reflection process (i.e., $z(0) = 0$, $z(\cdot)$ is nondecreasing and can increase only at t where $x^n(t) = 0$).

Proof: The proof of this theorem is inspired by the proof of Theorem 5.1.1 in [14]. Observe that we may write

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{\alpha,n}(t)} 1 &= \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{\alpha,n}(t)} \left(1 - \frac{\Delta_l^{\alpha,n}}{\bar{\Delta}^{\alpha,n}}\right) \\ &\quad + \frac{1}{\bar{\Delta}^{\alpha,n}\sqrt{n}} \sum_{l=1}^{nS^{\alpha,n}(t)} \Delta_l^{\alpha,n}, \end{aligned} \quad (6)$$

for $\alpha = a, r, d$. Define $w^{\alpha,n}(\cdot)$ as

$$w^{\alpha,n}(t) \triangleq \frac{1}{\sqrt{n}} \sum_{l=1}^{nt} \left(1 - \frac{\Delta_l^{\alpha,n}}{\bar{\Delta}^{\alpha,n}}\right). \quad (7)$$

Now an application of Theorem 2.8.6 in [14], which is an extension of Donsker's Theorem or Functional Central Limit Theorem, tells us that $w^{\alpha,n}(\cdot)$ converges weakly to the Wiener process with variance

$$\lim_n \mathbb{E}\left[\left(1 - \frac{\Delta_l^{\alpha,n}}{\bar{\Delta}^{\alpha,n}}\right)^2\right] = (\sigma^\alpha)^2.$$

Using Theorem 1.1 (in the appendix) and equation (2), we observe that $S^{\alpha,n}(t)$ converges weakly to the process that takes values $\lambda^\alpha t$, $\alpha = a, r$. Similarly, using equation (3), $S^{d,n}(t)$ is tight and any weak-sense limit has continuous sample paths by the first part of Theorem 1.1. Hence, since the Wiener process has almost certainly continuous sample paths, $w^{\alpha,n}(S^{\alpha,n}(\cdot))$ converges weakly to $w^\alpha(\lambda^a \cdot)$, $\alpha = a, r$, and $w^{d,n}(S^{d,n}(\cdot))$ is asymptotically continuous.

Observe that, for $\alpha = a, r$,

$$\frac{1}{\bar{\Delta}^{\alpha,n}\sqrt{n}} \sum_{l=1}^{nS^{\alpha,n}(t)} \Delta_l^{\alpha,n} = \frac{nt}{\bar{\Delta}^{\alpha,n}\sqrt{n}} + \epsilon_\alpha^n(t)$$

where $\epsilon_\alpha^n(t)$ denotes $1/(\bar{\Delta}^{\alpha,n}\sqrt{n})$ times the time since last arrival, which will be negligible as $n \to \infty$. For $\alpha = d$, we have to account for server idle time

$$\frac{1}{\bar{\Delta}^{d,n}\sqrt{n}} \sum_{l=1}^{nS^{d,n}(t)} \Delta_l^{d,n} = \frac{nt - T^n(t)}{\bar{\Delta}^{d,n}\sqrt{n}} + \epsilon_d^n(t)$$

where $\epsilon_d^n(t)$ is a negligible error.

Using expansion (4), we can write the following

$$\begin{aligned} x^n(t) &= x^n(0) + w^{a,n}(S^{a,n}(t)) - w^{r,n}(S^{r,n}(t)) \\ &\quad - w^{d,n}(S^{d,n}(t)) + \sqrt{n}\left(\frac{1}{\bar{\Delta}^{a,n}} - \frac{1}{\bar{\Delta}^{r,n}} - \frac{1}{\bar{\Delta}^{d,n}}\right) t \\ &\quad + z^n(t) + \epsilon^n(t), \end{aligned}$$

where $z^n(t) = z_0^n(t) + T^n(t)/(\sqrt{n}\bar{\Delta}^{d,n})$, and $\epsilon^n(t)$ is the sum of the negligible error terms.

The fact that $\{z^n(\cdot)\}$ is tight and converges weakly to the reflection term follows from Theorem 3.6.1 in [14], since $z^n(0) = 0$, $z_0^n(t)$ is non-decreasing and has jump sizes of $1/\sqrt{n}$, and $z^n(t)$ may increase only at the times when $x^n(t) \leq 1/\sqrt{n}$. This implies that $\{T^n(\cdot)/\sqrt{n}\}$ is tight and $S^{d,n}(\cdot)$ converges weakly to the process taking values $\lambda^d t$, by Theorem 1.1. ■

B. Workload Process

Define the workload process as the total time that is required to complete all work in the system. In other words, the workload at time t is the sum of all service times for all customers present in the system by this time. Following the usual scaling, let $wl^n(t)$ be $1/\sqrt{n}$ times the workload at time nt.

Suppose that upon the arrival of a negative customer, the system has to remove a fixed amount of work $\bar{W}$. This time, we change the model slightly and we do not care if a negative customer removes a customer at service. Similar to what was done in (4), we write

$$\begin{aligned} wl^n(t) &= wl^n(0) + \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{a,n}(t)} \Delta_l^{d,n} - \sqrt{n}t \\ &\quad - \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{r,n}(t)} \bar{W} + z^n(t), \end{aligned} \tag{8}$$

where $wl^n(0)$ is the initial scaled amount of work, $z^n(t) = z_0^n(t) + T^n(t)/\sqrt{n}$, and the process $T^n(t)$ is defined as the total server idle time by time nt. $z_0^n(\cdot)$ is added to maintain $wl^n(\cdot)$ positive, it will increase at the arrival of negative customers that find $wl^n(t) < \bar{W}/\sqrt{n}$.

Redefine b_n and b as the following

$$\sqrt{n}\bar{\Delta}^{d,n}\left(\frac{1}{\bar{\Delta}^{a,n}} - \frac{1}{\bar{\Delta}^{d,n}} - \frac{\bar{W}}{\bar{\Delta}^{r,n}\bar{\Delta}^{d,n}}\right) \triangleq b_n \to b \in \mathbb{R} \tag{9}$$

Then we arrive at the following result:

Theorem 3.2: Suppose that $wl^n(0)$ converges weakly to $wl(0)$, and is independent of the inter-arrival and service time processes. With assumption (9) and the ones described in the previous section, the process $wl^n(\cdot)$ converges weakly to the process $wl(\cdot)$, defined as

$$\begin{aligned} wl(t) &= wl(0) + \bar{\Delta}^d w^a(\lambda^a t) - \bar{W} w^r(\lambda^r t) \\ &\quad - \bar{\Delta}^d w^d(\lambda^a t) + bt + z(t) \end{aligned} \tag{10}$$

where $w^a(\cdot)$, $w^r(\cdot)$ and $w^d(\cdot)$ are independent Wiener processes with variances $(\sigma^a)^2$, $(\sigma^r)^2$, and $(\sigma^d)^2$, respectively, b is the constant in (9), and $z(\cdot)$ is the reflection process.

Proof: This proof is similar to the one of Theorem 3.1. For details see [15]. ∎

It is interesting to notice that if we set $\bar{W} = \bar{\Delta}^d$ and assume zero initial conditions, the limit expressions (5) and (10) differ only by the scale factor $\bar{\Delta}^d$ and the time scale of the Wiener process $w^d(\cdot)$. In order to compare the models of section III-A and of this section which differ only in the way that a negative customer removes work from queue, we will consider the following theorem. It says that the limit workload process of the two models are different only in the time scale of the Wiener process $w^d(\cdot)$, if $\bar{W} = \bar{\Delta}^d$.

Theorem 3.3: Let $\tilde{w}l^n(\cdot)$ denote $1/\sqrt{n}$ times the workload process at time nt for the model of section III-A. Using assumptions of Theorem 3.1, the difference $\bar{\Delta}^{d,n}x^n(t) - \tilde{w}l^n(t)$ converges weakly to the zero process.

Proof: We will follow the approach of Theorem 5.3.3 in [14]. Let $I^n(t) \triangleq$ number of customers that left the system by time nt (by either the arrival of a negative customer or service completion). Observe that $I^n(t)$ is bounded by $n(S^{d,n}(t)+S^{r,n}(t))$. Also, the processes $S^{d,n}(t)$ and $S^{r,n}(t)$ are bounded with high probability on any interval $[0,T]$, by Theorem 1.1. Re-index $\{\Delta_l^{d,n}; l\}$ by the order that the customers leave the system. Now, we may write

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{l=I^n(t)+2}^{I^n(t)+\sqrt{n}x^n(t)} \Delta_l^{d,n} &\le \tilde{w}l^n(t) \\ &\le \frac{1}{\sqrt{n}} \sum_{l=I^n(t)+1}^{I^n(t)+\sqrt{n}x^n(t)} \Delta_l^{d,n}. \end{aligned} \tag{11}$$

Observe that

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{l=I^n(t)+1}^{I^n(t)+\sqrt{n}x^n(t)} \Delta_l^{d,n} = \\ &\frac{1}{\sqrt{n}} \sum_{l=I^n(t)+1}^{I^n(t)+\sqrt{n}x^n(t)} (\Delta_l^{d,n} - \bar{\Delta}^{d,n}) + \bar{\Delta}^{d,n}x^n(t). \end{aligned} \tag{12}$$

Since $\{x^n(\cdot)\}$ is tight, $\{x^n(\cdot)/\sqrt{n}\}$ converges to the zero process. By the bound on $S^{d,n}(t) + S^{r,n}(t)$, $I^n(t)/n$ is also bounded with high probability in any $[0,T]$. Hence, since $\{n^{-1/2}\sum_{l=1}^{nt}(\Delta_l^{d,n} - \bar{\Delta}^{d,n})\}$ is tight and asymptotically continuous, the right hand sum of (12) converges weakly to the zero process. The same can be done for the left hand sum of (11). ∎

C. Markov Modulated Arrivals

In this section we will consider an extension of a model defined in [8]. Markov modulated G-queues have also been studied recently, for instance [6], [16], [7], motivated by the fact that arrival streams in Internet traffic are bursty and often correlated.

Suppose that there is only one arrival stream that feeds the system with positive and negative customers. The decision of whether an arrival will be positive or negative depends on the state of a Markov chain. This chain can also interfere in the distribution of $\Delta_l^{a,n}$. Similar to the model in section III, we assume that a negative customer may not remove a client from service. We formalize this model with the following assumptions and definitions.

Let us define $\mathbb{I}_l^{+,n}$ (resp., $\mathbb{I}_l^{-,n}$) as the indicator function of the event that the lth arrival is positive (resp., negative). Let $\{M_l^n; l\}$ be a stationary time-homogeneous irreducible Markov chain for each n on the state space $\mathcal{S}$. We assume that any p-step conditional probability converges geometrically to the stationary distribution $(\pi_k^n, k \in \mathcal{S})$. That is, we assume that there exists a constant C_n and $0 < \epsilon_n < 1$ such that

$$\sum_k \left| \sum_z \mathbb{P}\left(M_p^n = k \middle| M_0^n = z\right) \mu(z) - \pi_k^n \right| \le C_n(1-\epsilon_n)^p,$$

for any initial distribution μ. Also, assume that $\pi_k^n \to \pi_k$ for each $k \in \mathcal{S}$ as $n \to \infty$. A Markov chain has geometrically convergent transition probabilities if it, for example, satisfies

Doeblin's condition (e.g., [18]). For instance, it is sufficient if we suppose that $\mathcal{S}$ is finite and the chain is aperiodic.

Assume that the processes $\{M_l^n; l\}$, $\{\Delta_l^{a,n}; l\}$, and $\{\mathbb{I}_l^{\alpha,n}; l, \alpha = +,-\}$ are independent of $\{\Delta_l^{d,n}; l\}$. Also, conditioned on the modulating states, the random variables $\{\Delta_l^{a,n}, \mathbb{I}_l^{\alpha,n}; l, \alpha = +,-\}$ are mutually independent.

Let $w^{d,n}(\cdot)$, defined as in (7), converge weakly to a Wiener process with variance σ_d^2, where $\bar{\Delta}^{d,n}$ is a constant such that $\bar{\Delta}^{d,n} \to \bar{\Delta}^d \in \mathbb{R}_{>0}$. This condition is attended if, for example, we use assumption (2.1) for $\Delta_l^{d,n}$.

Define $\mathcal{F}_l^{a,n}$ as the minimal σ-algebra that measures all processes up to the time of the lth arrival, not including $\Delta_l^{a,n}$ and $\mathbb{I}_l^{\alpha,n}$, $\alpha = +,-$. Suppose that the distribution of $\Delta_l^{a,n}$ given $\mathcal{F}_l^{a,n}$ and the event $\{M_l^n = k\}$ does not depend on l or $\mathcal{F}_l^{a,n}$ and converges weakly as $n \to \infty$. Also, let

$$\begin{aligned}\bar{\Delta}^{a,n}(k) &\triangleq \mathbb{E}\left[\Delta_l^{a,n} \middle| \mathcal{F}_l^{a,n}, M_l^n = k\right] \\ &= \mathbb{E}\left[\Delta_l^{a,n} \middle| M_l^n = k\right] \to \bar{\Delta}^a(k) \in \mathbb{R}_{>0}, \\ v^{a,n}(k) &\triangleq \mathbb{E}\left[(\Delta_l^{a,n})^2 \middle| \mathcal{F}_l^{a,n}, M_l^n = k\right] \\ &= \mathbb{E}\left[(\Delta_l^{a,n})^2 \middle| M_l^n = k\right] \to v^a(k) \in \mathbb{R}_{>0},\end{aligned}$$

when $n \to \infty$, for each value of k, and

$$\sup_n \sup_{k \in \mathcal{S}} |v^{a,n}(k)| < \infty \qquad \sup_n \sup_{k \in \mathcal{S}} |\bar{\Delta}^{a,n}(k)| < \infty.$$

In addition, let $\{|\Delta_l^{a,n}|^2; n, l\}$ be uniformly integrable.

Suppose that for each value $k \in \mathcal{S}$, there are constants $q^\alpha(k) \in \mathbb{R}_{>0}$ such that

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_l^{\alpha,n} \middle| \mathcal{F}_l^{a,n}, M_l^n = k\right] &= \mathbb{E}\left[\mathbb{I}_l^{\alpha,n} \middle| M_l^n = k\right] \\ &= q^{\alpha,n}(k) \to q^\alpha(k)\end{aligned}$$

for $\alpha = +,-$, as $n \to \infty$. Note that $(q^{+,n}(k) + q^{-,n}(k))$ and $(q^+(k) + q^-(k))$ can take values less than 1. This is done to account for problems where there is a possibility that an arrival might not get to the queue when the modulating Markov chain is at state k.

We will also need the following heavy traffic condition

$$\sqrt{n}\left(\frac{\bar{q}^{+,n} - \bar{q}^{-,n}}{\bar{\Delta}^{a,n}} - \frac{1}{\bar{\Delta}^{d,n}}\right) \triangleq b_n \to b \in \mathbb{R}. \tag{13}$$

Similar to what was done in (4) and (8), we can write the process $x^n(t)$ as

$$\begin{aligned}x^n(t) = x^n(0) &+ \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{a,n}(t)} \left(\mathbb{I}_l^{+,n} - \mathbb{I}_l^{-,n}\right) \\ &- \frac{1}{\sqrt{n}} \sum_{l=1}^{nS^{d,n}(t)} 1 + z_0^n(t),\end{aligned} \tag{14}$$

where $z_0^n(t)$ is the total number of negative customers that arrived when the queue was empty by time nt.

Define the following constants

$$\bar{q}^{\alpha,n} \triangleq \sum_k q^{\alpha,n}(k)\pi_k^n \qquad \bar{\Delta}^{a,n} \triangleq \sum_k \bar{\Delta}^{a,n}(k)\pi_k^n$$

$$h_a^{\alpha,n} \triangleq \sum_k \bar{\Delta}^{a,n}(k) q^{\alpha,n}(k)\pi_k^n \qquad \bar{v}^{a,n} \triangleq \sum_k v^{a,n}(k)\pi_k^n$$

and denote by $\bar{q}^\alpha$, $\bar{\Delta}^a$, h_a^α, and $\bar{v}^a$ their respective limits as $n \to \infty$. Also, define the matrices

$$\Sigma_0 \triangleq \begin{pmatrix} \bar{q}^+ - (\bar{q}^+)^2 & -\bar{q}^+\bar{q}^- & h_a^+ - \bar{q}^+\bar{\Delta}^a \\ -\bar{q}^+\bar{q}^- & \bar{q}^- - (\bar{q}^-)^2 & h_a^- - \bar{q}^-\bar{\Delta}^a \\ h_a^+ - \bar{q}^+\bar{\Delta}^a & h_a^- - \bar{q}^-\bar{\Delta}^a & \bar{v}^a - (\bar{\Delta}^a)^2 \end{pmatrix}, \tag{15}$$

$$\Sigma_1 = \lim_n \sum_{u=1}^{\infty} \sum_{k,m} \zeta^n(k)\zeta^n(m)' D^{n,k,m}(u),$$

where $D^{n,k,m}(u) \triangleq [\mathbb{P}(M_u^n = m | M_0^n = k) - \pi_m^n]\pi_k^n$, and $\zeta^n(k) \triangleq (q^{+,n}(k), q^{-,n}(k), \bar{\Delta}^{a,n}(k))$.

It is perhaps noteworthy here that the above assumptions are standard in the framework of heavy traffic analysis [14].

Theorem 3.4: Suppose that $x^n(0)$ converges weakly to $x(0)$ and is independent of the inter-arrival and service time processes. With the assumptions above, the process $x^n(\cdot)$ converges weakly to the process $x(\cdot)$ taking values

$$\begin{aligned}x(t) = x(0) &- \lambda^a(\bar{q}^+ - \bar{q}^-)w^a(\lambda^a t) \\ &+ w^+(\lambda^a t) - w^-(\lambda^a t) - w^d(\lambda^d t) \\ &+ bt + z(t)\end{aligned} \tag{16}$$

where $w^d(\cdot)$ is a Wiener processes with variance $(\sigma^d)^2$, $\tilde{w}^a(\cdot) \triangleq (w^+(\cdot), w^-(\cdot), w^a(\cdot))$ is a Wiener process with covariance matrix $\Sigma \triangleq \Sigma_0 + 2\Sigma_1$, b is the constant in assumption (13), and $z(\cdot)$ is the reflection process.

Proof: The proof is based on the ideas of Theorem 5.5.1 in [14]. For details see [15]. ■

Let us now study the heavy traffic assumption (13) based on the result in [8], where it was considered a stability condition for a G-queue with Markov type decision for positive or negative arrival. Put in our notation, the scenario treated in [8] was the following: $\{M_l^n; l\}$ is a Markov chain with state space $\mathcal{S} = \{0, 1\}$, the chain is identical for each n, hence we drop this superscript. The transition matrix is given by

$$\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

where $p \in (0, 1)$, and define $q = 1 - p$. The decision of whether a customer was negative or positive was based on the current state of the chain, if it were 1 the arrival was positive, and if it were 0 the arrival was negative. Therefore, we have

$$q^+(0) = 0,\ q^+(1) = 1,\ q^-(0) = 1,\ q^-(1) = 0.$$

The stationary distribution can be easily computed to be $\pi = (q/(1+q), 1/(1+q))$. Hence, we have that $\bar{q}^+ = 1/(1+q)$ and $\bar{q}^- = q/(1+q)$. The inter-arrival time was independent of the modulating Markov chain, therefore $\bar{\Delta}^{a,n}(k)$ is just the mean inter-arrival time, for any k.

The result in [8] is that $\{N^n(t)\}$, where $N^n(t)$ is the number of customer in the queue by real time t, is a positive recurrent regenerative process if and only if

$$\frac{1}{\bar{\Delta}^{a,n}} < \frac{1}{\bar{\Delta}^{d,n}}\left(\frac{1+q}{1-q}\right).$$

Notice that there was a typographical error in [8], p is written in place of q. Now, assumption (13) tells us that, $\bar{\Delta}^{d,n}(\bar{q}^+ - \bar{q}^-)/\bar{\Delta}^{a,n} \uparrow 1$, or for this particular case,

$$\frac{\bar{\Delta}^{d,n}}{\bar{\Delta}^{a,n}}\left(\frac{1-q}{1+q}\right) \uparrow 1.$$

Hence, as $n \to \infty$ the system approaches the limit of stability, which characterizes the heavy traffic scenario. We believe that the stability condition $\bar{\Delta}^{d,n}(\bar{q}^+ - \bar{q}^-)/\bar{\Delta}^{a,n} < 1$ is valid for the larger class of Markov modulated G-queues, which is treated in this paper.

IV. Numerical Results

In applications, one uses the heavy traffic approximation above in the following way. First, one chooses $n \in \mathbb{N}$ such that b_n is of moderate size. Then one may approximate the number of customers (or the workload) at time t, denoted by $N(t)$, using $N(nt) \sim \sqrt{n}x(t; b_n)$, where $x(t; b_n)$ is the limit expression for the queueing process (i.e., equations (5), (10), (16)) with the drift constant b set to b_n.

In order to illustrate the approximation, we have calculated the mean number of customers and mean workload under steady-state in two different scenarios, using the heavy traffic approximation and a computer simulation for varying values of ρ. We can calculate the expected value of any ergodic reflected Brownian motion $x(t) = \sigma w(t) + bt + z(t)$, $x(t) \in \mathbb{R}_+$, under steady state using $\sigma^2/2|b|$ (see for instance [17]).

Let us first consider the non-Markov modulated models, where $\rho \triangleq \lambda^a/(\lambda^r + \lambda^d)$. For convenience, we set $\lambda^a = \rho$, $(\lambda^r + \lambda^d) = 1$, and assume that the distribution of the inter-arrival and service times are hyper-exponentially distributed for all examples. For the first problem, we set the squared coefficient of variation to 1, 1.5 and 2 for the positive customer inter-arrival time, negative customer inter-arrival time and service time distributions, respectively. Also, we set $\lambda^r = 1/4$ and $\lambda^d = 3/4$. For the second problem, we set $((\sigma^a)^2, (\sigma^r)^2, (\sigma^d)^2) = (2, 1.5, 1)$ with unchanged mean values. For the workload model we set $\bar{W} = \bar{\Delta}^d$ (the workload model used is the one in Theorem 3.2). Hence, we have that $\sigma^2 = \lambda^a(\sigma^a)^2 + \lambda^d(\sigma^d)^2 + \lambda^r(\sigma^r)^2$ and $b = \lambda^a - \lambda^d - \lambda^r = \rho - 1$, for the queue length, and $\sigma^2 = \left(\lambda^a(\sigma^a)^2 + \lambda^a(\sigma^d)^2 + \lambda^r(\sigma^r)^2\right)/(\lambda^d)^2$ and $b = \left(\lambda^a - \lambda^d - \lambda^r\right)/\lambda^d$, for the workload process. The results are given in Figure (1).

For the Markov modulated model, we consider the scenario of [8] (but with the mean for the arrival stream dependent on the Markov chain) in two different settings. As it was discussed in section III-C, the traffic intensity is defined to be $\rho \triangleq \lambda^a(1-q)/\lambda^d(1+q)$. For simplicity, let $\lambda^d = \lambda^a(1-q)/\rho(1+q)$, and let the distribution for the inter-arrival and service times be hyper-exponentially distributed. For the first setting we let $(\bar{\Delta}^a(1), \bar{\Delta}^a(0), (\sigma^a)^2, (\sigma^d)^2, q) = (2, 1, 2, 1, .1)$, and, for the second, $(\bar{\Delta}^a(1), \bar{\Delta}^a(0), (\sigma^a)^2, (\sigma^d)^2, q) = (1, 2, 1, 2, .9)$. For this example, the matrix Σ_1 is simply

$$\Sigma_1 = \frac{-q^2}{(1+q)^3}\sum_{k,m}(-1)^k(-1)^m\zeta(k)\zeta(m)'.$$

In order to calculate the mean number of customers under steady state, let $\tilde{\Sigma}$ be defined as $\Sigma = \Sigma_0 + 2\Sigma_1$ but with the elements of last row and column multiplied by $(\lambda^a(\bar{q}^+ - \bar{q}^-))^2$. Let A be a matrix such that $\tilde{\Sigma} = AA'$, and define $B = \nu A$, where ν is the row vector $(1, -1, -1)$. Then $\sigma^2 = BB'\lambda^a + (\sigma^d)^2\lambda^d$ and $b = \lambda^a(\bar{q}^+ - \bar{q}^-) - \lambda^d$. The results are also in Figure (1).

V. Conclusion

We have presented heavy traffic models of a queue that receives both positive and negative type of customers. Also, we apply this model to some examples to calculate the mean queue length and workload. The results seem to indicate that the approximations work well even for relative small values of ρ.

VI. Acknowledgments

Research partially supported by the Brazilian National Research Council-CNPq, under the Grants Nos. 140687/2005-0, 301740/2007-0 and 470527/2007-2, and FAPERJ, Grant No. E-26/100.579/2007.

References

[1] E. Altman and H.J. Kushner. Admission control for combined guaranteed performance and best effort communications systems under heavy traffic. *SIAM Journal on Control and Optimization*, 37(6):1780–1807, 1999.

[2] J.R. Artalejo and A. Gómez-Corral. On a single server queue with negative arrivals and request repeated. *Journal of Applied Probability*, 36:907–918, 1999.

[3] I. Atencia and P. Moreno. A single-server g-queue in discrete-time with geometrical arrival and service process. *Performance Evaluation*, 59(1):85–97, 2005.

[4] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, 1968.

[5] R.J. Boucherie and O.J. Boxma. The workload in the $m/g/1$ queue with work removal. *Probability in the Engineering and Informational Sciences*, 10:1–20, 1996.

[6] R. Chakka and T.V. Do. The $mm \sum_{k=1}^{k} cpp_k/ge/c/l$ g-queue with heterogeneous servers: Steady state solution and an application to performance evaluation. *Performance Evaluation*, 64(3):191–209, 2007.

[7] R. Chakka and P.G. Harrison. A markov modulated multi-server queue with negative customers—the mm cpp/ge/c/l g-queue. *Acta Informatica*, 37(11-12):881–919, 2001.

[8] E. Gelenbe, P. Glynn, and K. Sigman. Queues with negative arrivals. *Journal of Applied Probability*, 28(1):245–250, 1991.

[9] A. Gómez-Corral. On a tandem g-network with blocking. *Advances in Applied Probability*, 34:626–661, 2002.

[10] D. Gross and C.M. Harris. *Fundamentals of queueing theory (3nd ed.)*. John Wiley & Sons, New York, 1998.

[11] P.G. Harrison. Compositional reversed markov processes, with applications to g-networks. *Performance Evaluation*, 57(3):379–408, 2004.

[12] P.G. Harrison and E. Pitel. The m/g/1 queue with negative customers. *Advances in Applied Probability*, 28:540–566, 1996.

[13] G. Jain and K. Sigman. A pollaczek-khintchine formula for $m/g/1$ queues with disasters. *Journal of Applied Probability*, 33:1191–1200, 1996.

[14] H.J. Kushner. *Heavy traffic Analysis of Controlled Queueing and Communication Networks*. Springer-Verlag, New York, 2001.

[15] S.C. Leite and M.D Fragoso. On the analysis of g-queues under heavy traffic. LNCC Internal Report No. 34/2007, 2007.

[16] Y.W. Shin and B.D. Choi. A queue with positive and negative arrivals governed by a markov chain. *Probability in the Engineering and Informational Sciences*, 17(4):487–501, 2003.

[17] K. Sigman. Lecture notes on queueing theory. (url: http://www.columbia.edu/~ks20/6704-04/6704-04.html) Columbia University, New York, 2004.

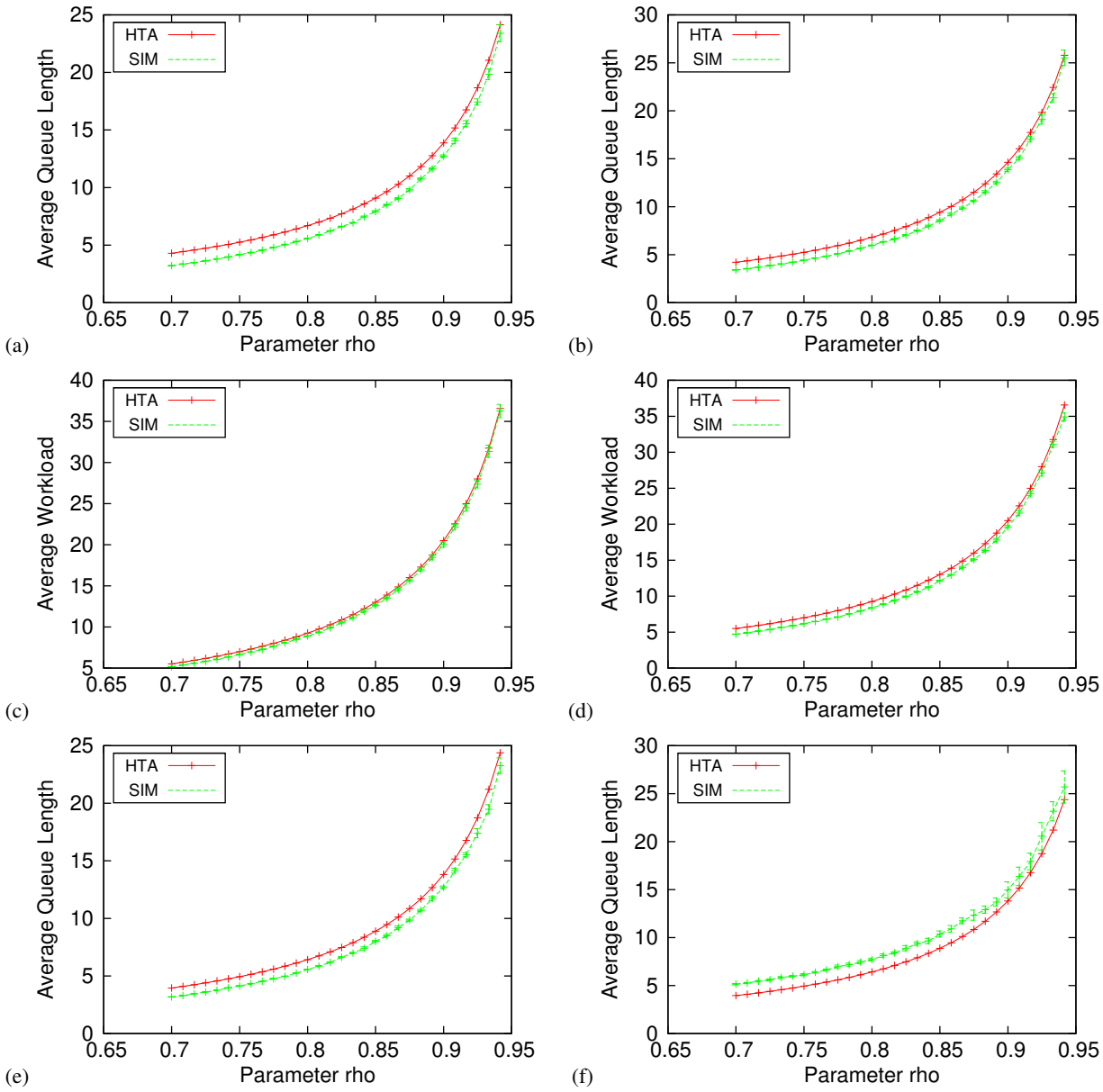

Fig. 1. Plot of mean number of customers in the queue and mean workload in two different scenarios for varying values of ρ. These values were computed using a computer simulation (SIM) and the heavy traffic approximation (HTA). For the simulation, it is also plotted the 95% t confidence interval (see [10], pg. 392). Figures (a) and (b) (resp., (c) and (d)) depict the average number of customer (resp., workload) for the first and second examples. Figures (e) and (f) show that average number of customers for the Markov modulated model.

[18] D.W. Stroock. *An Introduction to Markov Process*. Springer-Verlag, New York, 2005.

APPENDIX

The following theorems is shown here to facilitate referencing. The theorem is a result which is part of Theorem 5.1.1 in [14].

Theorem 1.1: Consider a set $\{\xi_l^n, l < \infty\}$, where ξ_l^n takes positive values, such that the set $\{h^n(\cdot)\}$, with process defined by

$$h^n(t) \triangleq \frac{1}{\sqrt{n}} \sum_{l=1}^{nt} \left(\xi_l^n - \bar{\xi}^n\right), \tag{17}$$

is tight, where $\bar{\xi}^n \in \mathbb{R}_{>0}$ and $\bar{\xi}^n \to \bar{\xi} \in \mathbb{R}_{>0}$ as $n \to \infty$. Let $\mathcal{J}^n(\cdot)$ be a nondecreasing process such that $\mathcal{J}^n(0) = 0$ and for every $t \in \mathbb{R}^+$, $\mathcal{J}^n(t) \leq nt$. Then $\{N^n(\cdot)\}$, where

$$N^n(t) \triangleq \frac{1}{n} \max \left\{ m \in \mathbb{N}_0 : \sum_{l=1}^{m} \xi_l^n \leq nt - \mathcal{J}^n(t) \right\},$$

is tight and any weak-sense limit has Lipschitz continuous sample paths, with Lipschitz constant no greater than $1/\bar{\xi}$. If in addition $\{\mathcal{J}^n(\cdot)/\sqrt{n}\}$ is tight, the process $N^n(\cdot)$ converges weakly to a process $N(\cdot)$ taking values $N(t) \triangleq t/\bar{\xi}$.

Proof: See [15]. ■