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

Diffusion Approximation of State Dependent G-Networks Under Heavy Traffic

Saul C. Leite and Marcelo D. Fragoso

Abstract— **This paper is concerned with the characterization of weak-sense limits of state-dependent G-network under heavy traffic. It is shown that, for a certain class of networks (which includes a two layer feedforward network and two queues in tandem), it is possible to approximate the number of customers in the queue by a reflected stochastic differential equation. The benefits of such an approach are that it describes the transient evolution of these queues and allows the introduction of controls, *inter alia*. We illustrate the application of the results with numerical experiments.**

Keywords: Queueing Theory, G-Networks, Heavy Traffic Type Result

I. Introduction

Queueing systems that receive signals, in addition to customers, are called G-networks and were first introduced by Gelenbe in [20]. Signals may come from outside or from other queues within the network and cause different types of effects on the receiving queue. A common type of signal, which is called "negative customer," forces the receiving queue to remove a customer from the waiting line. Other examples of signals include: "triggers," which moves a customer from one queue to another [21]; "disasters," which completely cleans the waiting line of the receiving queue [14]; and "resets," which sets the length of the receiving queue to a random value distributed according to the stationary distribution for that queue [24]. Thus, every queue in the system may exert some sort of control over the network through the signals. These models have been extensively studied (some examples include [49], [13], [32], [45], [34], [31], [15], [35], [16], [29]) and are motivated by a series of practical applications. One of the most successful applications, which was also the initial motivation for G-networks, is neural network modelling [22], [30], [19]. In this context, each queue represents a neuron and positive and negative customers are interpreted as excitatory and inhibitory signals, respectively (see also, [25], [7], [6]). Other applications include, computer networks with virus infection, load balancing networks, and synchronization signalling in parallel computation (see [5] for an extensive list of references). Another more recent application is modelling genetic regulatory systems [23], [3].

Although G-networks generally have some pleasing mathematical properties, such as product-form stationary distributions, the transient evolution of these systems are not easily (or conveniently) described and is rarely treated. The interaction among several different queues and the discrete nature of the system contribute to making it a complex problem and often the only resource available are simulations, which are time consuming and computationally expensive. Moreover, problems such as the optimum choice of signal or customer scheduling are impractical in this setting. Thus, a mathematical model is sought, even if approximative, that can give a reasonable degree of accuracy.

There exists two common types of approximations that describe the transient evolution of queueing networks: fluid and diffusion (or heavy traffic) approximations. Usually, fluid models describes the dynamics of the system "average" by a differential equation. Diffusion approximations differs from the fluid model in the fact that the "randomness" usually found in queueing systems is not averaged out and it appears in the model as a Wiener process (or in some cases as an Itô integral). Hence, diffusion approximations are more faithful to the dynamics of the system when compared to fluid approximations. However, this comes with the addition of the heavy traffic assumption, which requires the rate of customers entering a queue to be close to the rate of customers leaving this queue. This is a common scenario in many applications of interest, most notably in modern computer systems.

The problem of describing the transient evolution of a queueing network with negative customers has been dealt with in some recent works using fluid approximation [33], [4]. In the latter article, transient evolution of a state-dependent network with negative customers was considered using a fluid approximation together with a heavy traffic assumption. However, as discussed in the above paragraph, diffusion approximations are more suited for systems under this condition. To our knowledge, G-networks have not yet been treated under a diffusion analysis. Such an approximation is useful in practical problems in which G-networks are applicable. For example, one could use the heavy traffic approximation to construct a stochastic optimal control problem for synchronization of signals in parallel computer systems. In addition, the diffusion model can help us gain insights into the connections among some of the model parameters and the general behavior of queueing networks with signals.

Diffusion approximations for queueing systems have been studied since the pioneering works of Kingman [39], Prohorov [47] and Borovkov [9], [10] in the early 60's. Other early papers on the subject include [48], [18], [50], [37], [17], to cite a few. One of the interesting aspects of diffusion approximations is that they offer a "macroscopic view" [51]

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

of the complex interactions that are present in queueing networks, and synthesize the general behavior of the system in a simple time dependent equation. In addition, the approximation has been observed to give a good estimate for systems under heavy, or only moderately heavy, traffic [40]. Hence, it is no surprise that it has been successfully applied to several practical problems, most of them in computer systems, where heavy traffic is common. Some example include ([2], [12], [1], [27], [26], [42], [28]).

In this paper, we will consider G-networks with *state-dependent* arrival rates (of customers and signals), service rates and routing probabilities. Besides [4], discussed above, other results regarding state dependent G-queues and networks treat the system under stationary regime [8], [36], [16]. As mentioned previously, the benefits of the diffusion approximation is that it describes the transient evolution of these networks via a stochastic model. In addition, the state-dependence allows for the introduction of feedback controls [40]. We also consider that the network is *under heavy traffic*, in the sense that every queue in the system is operating at nearly maximum capacity. Under this condition, it will be shown that the number of customers in each queue in the network can be approximated by a *reflected stochastic differential equation*. The model is an adaptation of the models presented in [40], [46] with the introduction of negative customers. The result presented here is for a certain class of queueing networks which satisfy assumption 3.2a, which will be presented later in the development of the model. Two examples of networks are given which satisfy this condition: two queues in tandem and a two layer feedforward network.

The layout of the paper is as follows: in the following section the queueing model treated here will be described in more details. Next, in section III, the heavy traffic theorem for the number of customers will be stated and proved. In section IV, it will be shown two examples of networks which satisfy condition 3.2a. Finally, in section V, we illustrate the application of the model with a numerical example.

II. Queueing Model

We will restrict ourselves to queues with one server, first come first served (FCFS) service discipline, and signals of the "negative customer" type. Hence, any queue that receives this signal is forced to remove a customer from the system. If the queue is empty, the negative customer will have no effect on the system. Although being denominated a "customer," this signal does not receive service and leaves the receiving queue immediately after its arrival. Signals coming from within the network are regular customer that have finished work at a queue and were routed as a negative customer.

The queue length process, X_i, for a network of K queues takes the form

$$\begin{aligned} X_i(t) &= X_i(0)+A_i(t)-D_i(t)-S_i(t) \\ &\quad +\sum_{j\leq K}\left[D_{ji}^{+}-D_{ji}^{-}(t)\right]-U_i(t), \end{aligned} \tag{1}$$

where $A_i(t)$ is the cumulative number of exogenous clients that arrived at queue i by time t, $D_i(t)$ is the number of service completions at queue i by time t, and $S_i(t)$ is the number of removed customers due to an exogenous signal by time t. The process $D_{ji}^{+}(t)$ denotes the total number of customers that left queue j and joined queue i as a regular customer by time t, and $D_{ji}^{-}(t)$ is the total number of customers removed from queue i due to a negative arrival originated from queue j.

Let N_i^{α} be standard Poisson processes with càdlàg sample paths, for $\alpha\in\{a,s,d\}$ and $i\in\{1,...,K\}$. Also, define the processes $\tilde{N}_i^{\alpha}(t)\triangleq N_i^{\alpha}(\int_0^t\Lambda_i^{\alpha}(X(s))ds)$, where $\Lambda_i^{\alpha}:\mathbb{R}_+^K\to\mathbb{R}_+$ are measurable functions. Then we define $A_i(t)\triangleq\tilde{N}_i^{a}(t)$, $S_i(t)\triangleq\tilde{N}_i^{s}(t)$, $D_i(t)\triangleq\int_0^t\mathbb{I}_{\{X_i(s-)>0\}}\tilde{N}_i^{d}(t)$, $D_{ij}^{+}(t)\triangleq\int_0^t\mathbb{I}_{ij}^{+}(s)dD_i(s)$, and $D_{ij}^{-}(t)\triangleq\int_0^t\mathbb{I}_{ij}^{-}(s)d\tilde{D}_{ij}(s)$, where $\tilde{D}_{ij}(t)\triangleq\int_0^t\mathbb{I}_{\{X_i(s-)>0,X_j(s-)>0\}}d\tilde{N}_i^{d}(s)$.

The processes $\mathbb{I}_{ji}^{+}(t)$ and $\mathbb{I}_{ji}^{-}(t)$ are defined as the indicator functions of the events that a customer leaving queue j at time t is routed to queue i as a positive or negative customer, respectively. The process $U_i(t)$ denotes the cumulative number of customers not allowed to enter the queue due to the buffer being full by time t. If the buffer size is infinite for queue i, the process $U_i(t)$ can be considered as the "zero" process.

All stochastic processes given above are defined on the same probability space $(\Omega,\mathscr{F},P)$. References to it are not necessary and will be omitted henceforth. Let $\mathscr{F}_t$ be the minimal σ-algebra that measures all driving processes defined above up to time t (i.e., $\{\mathscr{F}_t,\ t\geq 0\}$ is a filtration). In addition, the following assumption will be used. Amongst other things, it guarantees that the counting processes defined above are nonexplosive and have a martingale representation, which will be given below. The condition on the continuity and boundedness of the rates can be relaxed and that is discussed in [44].

Assumption 2.1: (a) The random quantities $X_i(0)$, N_i^{α}, $i\in\{1,...,K\}$, $\alpha\in\{a,s,d\}$, are mutually independent.

(b) The functions $\Lambda_i^{\alpha}(\cdot)$, $i\in\{1,...,K\}$, $\alpha\in\{a,s,d\}$, are continuous and bounded.

(c) $\mathbb{E}\left[\mathbb{I}_{ij}^{\alpha}(t)\middle|\mathscr{F}_t^r\right]=Q_{ij}^{\alpha}(X(t-))$, for $i,j\in\{1,...,K\}$, $\alpha\in\{+,-\}$, where $Q_{ij}^{\alpha}:\mathbb{R}_+^K\to[0,1]$ is a measurable function and $\mathscr{F}_t^r$ is the minimal σ-algebra that measures all driving processes up to time t, not including the current routing decision.

Owning to the assumption above, the jump processes A_i, D_i, S_i and $\tilde{D}_{ij}$ have the following martingale decompositions (see [46] pg. 625 and [11], T8 pg.27)

$$\begin{aligned} A_i(t) &= M_i^{a}(t)+\int_0^t\Lambda_i^{a}(X(s))ds \\ D_i(t) &= M_i^{d}(t)+\int_0^t\mathbb{I}_{\{X_i(s)>0\}}\Lambda_i^{d}(X(s))ds \\ S_i(t) &= M_i^{s}(t)+\int_0^t\mathbb{I}_{\{X_i(s)>0\}}\Lambda_i^{s}(X(s))ds \\ \tilde{D}_{ij}(t) &= \tilde{M}_{ij}^{d}(t)+\int_0^t\mathbb{I}_{\{X_i(s)>0,X_j(s)>0\}}\Lambda_i^{d}(X(s))ds, \end{aligned}$$

where M_i^a, M_i^s, M_i^d, and $\tilde{M}_{ij}^d$ are $\mathscr{F}_t$-martingales. In order to have a martingale decomposition for D_{ij}^+ and D_{ij}^- define

$$M_{ij}^+(t) \triangleq \int_0^t \left(\mathbb{I}_{ij}^+(s) - Q_{ij}^+(X(s-))\right) dD_i(s)$$
$$M_{ij}^-(t) \triangleq \int_0^t \left(\mathbb{I}_{ij}^-(s) - Q_{ij}^-(X(s-))\right) d\tilde{D}_{ij}(s)$$

The same argument used in ([46], pg.626) can be used to show that M_{ij}^+ and M_{ij}^- are $\mathscr{F}_t$-martingales. Now it is possible to write

$$\begin{aligned} D_{ij}^+(t) &= \int_0^t \left(\mathbb{I}_{ij}^+(s) - Q_{ij}^+(X(s-))\right) dD_i(s) \\ &\quad + \int_0^t Q_{ij}^+(X(s-)) dD_i(s) \\ &= M_{ij}^+(t) + \int_0^t Q_{ij}^+(X(s-)) dM_i^d(s) \\ &\quad + \int_0^t Q_{ij}^+(X(s)) \mathbb{I}_{\{X_i(s)>0\}} \Lambda_i^d(X(s)) ds \\ D_{ij}^-(t) &= M_{ij}^-(t) + \int_0^t Q_{ij}^-(X(s-)) d\tilde{M}_{ij}^d(s) \\ &\quad + \int_0^t Q_{ij}^-(X(s)) \mathbb{I}_{\{X_i(s)>0, X_j(s)>0\}} \Lambda_i^d(X(s)) ds. \end{aligned}$$

Hence, the process $X = (X_i, i = 1...K)'$ accepts the following representation

$$X(t) = X(0) + \int_0^t B(X(s)) ds + M(t) - U(t)$$

where

$$\begin{aligned} B_i(x) &\triangleq \Lambda_i^a(x) - \Lambda_i^d(x)\mathbb{I}_{\{x_i>0\}} - \Lambda_i^s(x)\mathbb{I}_{\{x_i>0\}} \\ &\quad + \sum_{j\le K} \left[Q_{ji}^+(x)\mathbb{I}_{\{x_j>0\}} - Q_{ji}^-(x)\mathbb{I}_{\{x_j>0, x_i>0\}}\right] \Lambda_j^d(x) \\ M_i(t) &\triangleq M_i^a(t) - M_i^d(t) - M_i^s(t) + \sum_{j\le K} \Big[M_{ji}^+(t) - M_{ji}^-(t) \\ &\quad + \int_0^t Q_{ji}^+(X(s-)) dM_j^d(s) - \int_0^t Q_{ji}^-(X(s-)) d\tilde{M}_{ji}^d(s)\Big], \end{aligned}$$

and M_i is an $\mathscr{F}_t$-martingale.

III. Heavy Traffic Limit

As it is usual in heavy traffic analysis, we consider a sequence of queueing networks $(X^n, n > 0)$ indexed by the parameter n. As n increases, the system approaches heavy traffic in the sense that the rate of customers entering the system approaches that of customers leaving the system. The following scale is usually employed:

$$x^n(t) \triangleq X^n(nt)/\sqrt{n}.$$

Let any mathematical object defined in the previous section with respect to X^n be now indexed with an upper script n (e.g., $\mathscr{F}_t^n$, $\Lambda_i^{a,n}$, $Q_{ij}^{+,n}$, etc.). Similarly, any counting process defined in the previous section (e.g., A_i, S_i, D_i, D_{ij}^+, D_{ij}^-, U_i) is now replaced by its scaled equivalent (e.g., A_i^n, S_i^n, D_i^n, $D_{ij}^{+,n}$, $D_{ij}^{-,n}$, U_i^n). For example, $A_i^n(t)$ now denotes $1/\sqrt{n}$ times the number of exogenous customers that arrived at queue i by time nt, that is,

$$A_i^n(t) \triangleq \frac{1}{\sqrt{n}} N_i^a \left(\int_0^{nt} \lambda_i^{a,n}(x^n(s/n)) ds\right)$$

with $\lambda_i^{a,n}(\xi) \triangleq \Lambda_i^{a,n}(\sqrt{n}\xi)$, $\xi \in \mathbb{R}^K$, for each $n > 0$, and the martingale decomposition becomes

$$A_i^n(t) = M_i^{a,n}(t) + \sqrt{n}\int_0^t \lambda_i^{a,n}(x^n(s)) ds,$$

after a change of variable. Likewise, let $q_{ij}^{\alpha,n}(\xi) \triangleq Q_{ij}^{\alpha,n}(\sqrt{n}\xi)$, for $\alpha \in \{+,-\}$, $i,j \in \{1,...,K\}$, $\xi \in \mathbb{R}^K$. Also, let the size of the buffer for the scaled process be B_i for the ith queue (B_i may be infinite).

In the first assumption below, it will be defined how state dependence is introduced. Even though the dependence is very small for large n, it has significant effect in the limit. As it will be seen, the functions $f_i^\alpha(x)$ and $f_{ij}^\beta(x)$ will appear in the drift term of the limit equation.

Assumption 3.1: For $o(\cdot)$ uniformly in x, we assume that:

(a) There exists non-negative constants r_i^α and r_{ij}^β, and bounded and continuous functions $f_i^\alpha(x)$ and $f_{ij}^\beta(x)$, $j,i \in \{1,...,K\}$, $\alpha \in \{a,d,s\}$ and $\beta \in \{+,-\}$, such that $\lambda_i^{\alpha,n}(x) = r_i^\alpha + f_i^\alpha(x)/\sqrt{n} + o(1/\sqrt{n})$, and $q_{ij}^{\beta,n}(x) = r_{ij}^\beta + f_{ij}^\beta(x)/\sqrt{n} + o(1/\sqrt{n})$.

(b) For any $i \in \{1,..,K\}$ $r_i^a + \sum_{j\le K} r_j^d r_{ji}^+ = r_i^d + r_i^s + \sum_{j\le K} r_j^d r_{ji}^-$, which is usually called the heavy traffic condition. This condition tells us that for large n, the rate of customers joining a queue in the network is very close to the rate of customer leaving this queue.

The next assumption will be on the *reflection directions* and they will become more clear throughout the development of the proof of Theorem 3.3. Define the following matrices in $\mathbb{R}^{K\times K}$:

$$\begin{array}{ll} I_r \triangleq \mathrm{diag}(r_i^d + r_i^s)_{i=1,...,K} & \Theta_{ij} \triangleq r_i^d(r_{ij}^+ - r_{ij}^-) \\ R^S \triangleq \left(I_r - \Theta' + \Omega^S\right) & R \triangleq R^\emptyset \\ \Omega^S \triangleq \mathrm{diag}\left(\sum_{j\in Z\setminus S\cup\{i\}} r_{ji}^- r_j^d\right)_{i=1,...,K} & \end{array} \quad (2)$$

where $S \subseteq Z \triangleq \{1,...,K\}$, and $\mathrm{diag}(a_i)_{i=1,...,m} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with entries a_i.

Assumption 3.2: (a) For any $S \subseteq \{1,...,K\}$, with $|S| \ge 2$, there exists an $\alpha = (\alpha_1,...,\alpha_{|S|})' \ne 0$, where $|S|$ denotes the number of elements of S, such that $\alpha_i \ge 0$ and

$$R_{\{i\in S\}}\alpha = R^S_{\{i\in S\}} e, \quad \text{with } e = (1,...,1)'.$$

The subscript $\{i \in S\}$ on any matrix A indicates that $A_{\{i\in S\}} \in \mathbb{R}^{K\times|S|}$ is formed by the columns of A with indices in S.

(b) The matrix R satisfies the completely-$\mathscr{S}$ condition (see [40], pg. 121).

We are now able to present the heavy traffic limit in the theorem below.

Theorem 3.3: Let $x^n(0)$ converge weakly to $x(0)$. With assumptions 2.1, 3.1 and 3.2, $\{x^n(\cdot)\}$ is tight and any weakly

convergent subsequence satisfies

$$\begin{aligned} x(t) &= x(0)+\int_0^t b(x(s))ds+M(t)+z(t), \qquad (3)\\ z(t) &= Ry(t)-u(t), \end{aligned}$$

where $0 \le x_i(t) \le B_i$, $i \in \{1,...,K\}$, and

$$\begin{aligned} b_i(x) &= f_i^a(x)-f_i^d(x)-f_i^s(x)\\ &\quad+\sum_{j=1}^{K}\left[f_j^d(x)(r_{ji}^+-r_{ji}^-)+r_j^d(f_{ji}^+(x)-f_{ji}^-(x))\right], \qquad (4)\\ M_i(t) &= M_i^a(t)-M_i^d(t)-M_i^s(t)\\ &\quad+\sum_{j=1}^{K}\left[(r_{ji}^+-r_{ji}^-)M_j^d(t)+M_{ji}^+(t)-M_{ji}^-(t)\right]. \end{aligned}$$

The M_i^α, $\alpha \in \{a,d,s,r\}$, $i \in \{1,...,K\}$, are mutually independent Wiener processes, where $M_i^r(t) \triangleq (M_{i1}^+(t),...,M_{iK}^+(t),M_{i1}^-(t),...,M_{iK}^-(t))'$. M_i^α has variances r_i^α, for $\alpha \in \{a,d,s\}$, and M_i^r has covariance matrix

$$(\Sigma_i)_{jk} = r_i^d \cdot \begin{pmatrix} (\Sigma_i^+) & (\Sigma_i^{+-}) \\ (\Sigma_i^{+-})' & (\Sigma_i^-) \end{pmatrix}$$

with
$$(\Sigma_i^\alpha)_{jk} = \begin{cases} (1-r_{ij}^\alpha)r_{ij}^\alpha & \text{if } j=k \\ -r_{ij}^\alpha r_{ik}^\alpha & \text{otherwise,} \end{cases}$$

for $\alpha \in \{+,-\}$, and $(\Sigma_i^{+-})_{jk} = -r_{ij}^+ r_{ik}^-$, where Σ_i^{+-}, $\Sigma_i^\alpha \in \mathbb{R}^{K\times K}$.

The process $z(\cdot)$ is the reflection term ($y_i(0)=0$, $y_i(\cdot)$ are continuous, nondecreasing, and can increase only at t where $x_i(t)=0$. Similarly, if $B_i<\infty$, $u_i(0)=0$, $u_i(\cdot)$ are continuous, nondecreasing and increase only when $x_i(t)=B_i$).

Proof: See [44] for details. ■

IV. Networks Satisfying Condition 3.2a

Conditions 2.1, 3.1 and 3.2b are usual assumptions in heavy traffic approximations for state-dependent queueing systems [40]. Loosely speaking, that is also true for condition 3.2a, that essentially requires that any reflection direction appearing on the edge or corner of the state-space be a positive linear combinations of the reflections at the adjacent boundaries. In fact, if we consider a network where queues can receive signals from outside but not from within the network, the condition is automatically satisfied. However, it is not valid for any G-network. In this section, it is shown two types of network topologies that satisfy assumption 3.2a.

A. Two queues in tandem

Consider two queues in tandem where each queue may send regular or negative customers to each other, as seen in figure 1(a). Both queues receive customers and signals from exogenous sources, and there is no feedback, in the sense that a customer that has just left queue i may not be routed immediately to queue i.

For this case, the matrices I_r, Θ, Ω^S and R are defined as follows:

$$\begin{aligned} \Omega^\emptyset &= \begin{pmatrix} r_{21}^- r_2^d & 0 \\ 0 & r_{12}^- r_1^d \end{pmatrix}\\ \Omega^{\{1,2\}} &= 0\\ I_r &= \begin{pmatrix} r_1^d+r_1^s & 0 \\ 0 & r_2^d+r_2^s \end{pmatrix}\\ \Theta &= \begin{pmatrix} 0 & r_1^d(r_{12}^+-r_{12}^-) \\ r_2^d(r_{21}^+-r_{21}^-) & 0 \end{pmatrix}\\ R &= \begin{pmatrix} r_1^d+r_1^s+r_{21}^- r_2^d & -r_2^d(r_{21}^+-r_{21}^-) \\ -r_1^d(r_{12}^+-r_{12}^-) & r_2^d+r_2^s+r_{12}^- r_1^d \end{pmatrix}\\ R^{\{1,2\}} &= \begin{pmatrix} r_1^d+r_1^s & -r_2^d(r_{21}^+-r_{21}^-) \\ -r_1^d(r_{12}^+-r_{12}^-) & r_2^d+r_2^s \end{pmatrix}. \end{aligned}$$

Hence, the condition is verified if there is an $\alpha=(\alpha_1,\alpha_2)'$ with positive components such that $R\alpha = R^{\{1,2\}}e$, which is true as long as $r_1^d, r_2^d > 0$.

B. Two layer feedforward network

Let us now consider a feedforward network with two layers, in the sense that the queues on the first layer may send customers to queues in the second layer, but not the converse, see figure 1(b) for reference. Each queue can also receive regular and negative exogenous arrivals and there is no feedback. Suppose that there are K_1 queues on the first layer and K_2 on the second. Define $K=K_1+K_2$. We index the queues starting on the first layer in such a way that if $K_1 < i \le K$, the ith queue is in the second layer.

Define $Z_1=\{1,...,K_1\}$. For this scenario, the matrix R^S is given by

$$R^S = \begin{pmatrix} \mathrm{diag}(r_i^d+r_i^s)_{i=1,...,K_1} & 0 \\ -\tilde{\Theta}' & \Psi \end{pmatrix}$$

where $\tilde{\Theta} \in \mathbb{R}^{K_1\times K_2}$ and $\Psi \in \mathbb{R}^{K_2\times K_2}$ are defined as

$$\tilde{\Theta} = \begin{pmatrix} r_1(r_{1(K_1+1)}^+ - r_{1(K_1+1)}^-) & \cdots & r_1(r_{1K}^+ - r_{1K}^-) \\ \vdots & & \vdots \\ r_{K_1}(r_{K_1(K_1+1)}^+ - r_{K_1(K_1+1)}^-) & \cdots & r_{K_1}(r_{K_1K}^+ - r_{K_1K}^-) \end{pmatrix},$$

$$\Psi = \mathrm{diag}\left(r_i^d+r_i^s+\sum_{j\in Z_1\setminus S} r_{ji}^- r_j^d\right)_{i=K_1+1,...,K}.$$

In order to verify the assumption 3.2a, let $S \subseteq \{1,...,K\}$ be an ordered set, with $|S| \ge 2$. Define $S_1 \subseteq \{1,...,K_1\}$ and $S_2 \subseteq \{K_1+1,...,K\}$ as ordered sets such that $S_1 \cup S_2 = S$. Then choose $\alpha = (\alpha_1,...,\alpha_{|S_1|},\beta_1,...,\beta_{|S_2|})'$ such that $\alpha_i = 1$, for $i=1,...,|S_1|$, and

$$\beta_j = \frac{r_k^d+r_k^s+\sum_{l\in Z_1\setminus S_1} r_{lk}^- r_l^d}{r_k^d+r_k^s+\sum_{l\in Z_1} r_{lk}^- r_l^d}, \qquad \text{for } j=1,...,|S_2|,$$

where k is the jth element of S_2. Now we can verify that $R_{\{i\in S\}}\alpha = R^S_{\{i\in S\}}e$. Since this works for any choice of S, assumption 3.2a is satisfied.

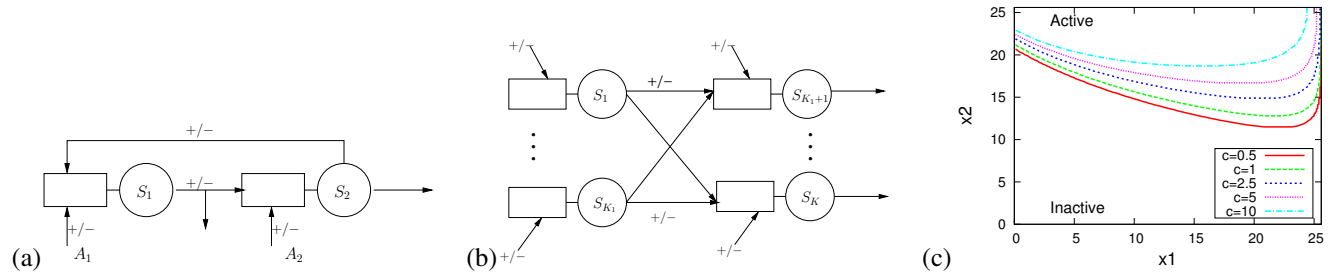

Fig. 1. (a) Two queues in tandem. The symbols + and − indicates the arrival of regular or negative customers, respectively. (b) Two layer feedforward network. (c) Switching curves for the optimal controls with varying values for c .

V. Numerical Experiments

In order to illustrate an application of the theorem, let us suppose that we have the system of subsection IV-A, given by figure 1(a). It will be assumed that every customer leaving queue 1 joins queue 2 as a regular customer, and queue 2 does not receive exogenous clients. Also, both queues have finite buffers. Suppose that queue 2 needs to reduce customer loss due to buffer overflow and it does that by sending signals to queue 1. Hence, every time queue 2 has its buffer almost full, it will start sending signals to the first queue. In this example, it will be shown how one can use the result derived here to choose the *optimal routing* strategy for a *system operating under heavy traffic*.

As it is common in application (e.g., [43], [42], [40]), we do not have a sequence of queues indexed by the parameter n. Rather, we have one queueing system that we want to approximate. Hence, we need to choose a large n such that the rates for our problem satisfy

$$\begin{aligned} \Lambda_i^{\alpha}(\sqrt{n}x) &= \lambda_i^{\alpha,n}(x) \approx r_i^{\alpha} + f_i^{\alpha}(x)/\sqrt{n}, \quad \alpha = a,d,s \\ Q_{ij}^{\beta}(\sqrt{n}x) &= q_{ij}^{\beta,n}(x) \approx r_{ij}^{\beta} + f_{ij}^{\beta}(x)/\sqrt{n}, \quad \beta = +,- \end{aligned}$$

and the heavy traffic condition holds (i.e., assumption 3.1b). Now, we can approximate the number of customers in each queue at time t, $X(t)$, by $X(nt) \sim \sqrt{n}x(t)$, where $x(\cdot)$ is the limit process given by (3).

For our example, let us suppose that

$$\lambda_1^{a,n}(x) = \lambda \quad \lambda_2^{a,n}(x) = 0 \quad \lambda_1^{s,n}(x) = 0 \quad \lambda_2^{s,n}(x) = 0$$

and that $\lambda_1^{d,n}(x) = \mu_1$ and $\lambda_2^{d,n}(x) = \mu_2$, where μ_1, μ_2 and λ are positive constants. By the heavy traffic assumption, there exist ("small") constants b_1 and b_2 such that $b_1 = \sqrt{n}(\mu_1 - \lambda)$ and $b_2 = \sqrt{n}(\mu_2 - \mu_1)$. That is, the rate of customers entering each queue is close to the rate of departing customers. Hence, $\lambda_1^{d,n}(x) = \lambda + b_1/\sqrt{n}$ and $\lambda_2^{d,n}(x) = \lambda + (b_1+b_2)/\sqrt{n}$. Also, we suppose that

$$\begin{aligned} q_{12}^{+,n}(x) &= 1 \qquad q_{21}^{+,n}(x) = 0 \\ q_{12}^{-,n}(x) &= 0 \qquad q_{21}^{-,n}(x) = g(x)/\sqrt{n}, \end{aligned}$$

where $g : \mathbb{R}^2 \to [0,1]$, and that the size of the (unscaled) buffers are $\sqrt{n}B_1$ for the first queue and $\sqrt{n}B_2$ for the second. The heavy traffic limit is given by

$$\begin{aligned} dx(t) = \begin{pmatrix} -b_1 - \lambda g(x(t)) \\ -b_2 \end{pmatrix} dt + \begin{pmatrix} 2\lambda & -\lambda \\ -\lambda & 2\lambda \end{pmatrix}^{1/2} dW(t) \\ + \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} dy(t) \; - du(t), \end{aligned}$$

where $A^{1/2}(A^{1/2})' = A$. It is perhaps noteworthy to mention that the function $g(\cdot)$ only acts upon the first component of $x(\cdot)$, even though we are interested in controlling the second. However, the second queue will be affected indirectly by the control through the reflection term.

We can now find the optimal choice of $g(\cdot)$ with respect to a cost function. For this example, we will use the following discounted cost:

$$W(x,g) = \mathbb{E}_x^g \left[\int_0^{\infty} e^{-\beta t} \left[cg(x(t))dt + vdu_2(t) \right] \right],$$

where c and v are constants associated with the cost of routing negative customers (or losing customers at the first queue) and the cost of losing customers due to buffer overflow at queue 2, respectively.

We use the Markov chain approximation method [41], [38] to find the optimal control numerically. We set $\beta = 0.01$, and the discretization parameter is set to $h = 0.1$. Figure 1(c) plots the control for different choices of c and v set to 200. Notice that the optimal control is of the switching type (i.e., after a given threshold, the control is used at maximum rate). This type of optimal control has also been found in different situations for the control of queueing systems [43].

It is interesting to see the shape of these switching curves. Notice that the curves move upwards at the right side of the state space. This can be explained by the delay of the control action since the control at queue 2 is done indirectly. When queue 1 and queue 2 are almost full, there most likely will be buffer overflow loss at queue 2 even if it sends signals to queue 1.

VI. Conclusions

We have presented heavy traffic limits for a class of state-dependent G-networks which satisfy assumption 3.2a. Two examples are given which satisfy this condition. Our current work concentrates in extending the results to any class of G-networks, and for networks with different kinds of signals.

VII. 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 by 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] E. Altman and H.J. Kushner. Control of polling in presence of vacations in heavy traffic with applications to satellite and mobile radio systems. *SIAM J. Control Optim.*, 41(1):217–252, 2002.

[3] A. Arazi, E. Ben-Jacob, and U. Yechiali. Bridging genetic networks and queueing theory. *Physica A: Statistical Mechanics and its Applications*, 332:585–616, 2004.

[4] A. Arazi, E. Ben-Jacob, and U. Yechiali. Controlling an oscillating jackson-type network having state-dependent service rates. *Mathematical Methods of Operations Research*, 62(3):453–466, 2005.

[5] J.R Artalejo. G-networks: A versatile approach for work removal in queueing networks. *European Journal of Operational Research*, 126:233–249, 2000.

[6] V. Atalay and E. Gelenbe. Parallel algorithm for color texture generation using the random neural network model. *International Journal of Pattern Recognition and Artificial Intelligence*, 6(2-3):437–446, 1992.

[7] V. Atalay, E. Gelenbe, and N. Yalabik. The random neural network model for texture generation. *International Journal of Pattern Recognition and Artificial Intelligence*, 6(1):131 – 141, 1992.

[8] P.P. Bocharov, E.V. Gavrilov, and A.V. Pechinkin. Exponential queuing network with dependent servicing, negative customers, and modification of the customer type. *Automation and Remote Control*, 65(7):35–59, 2004.

[9] A. Borovkov. Some limit theorems in the theory of mass service i. *Theor. Probability Appl.*, 9:550–565, 1964.

[10] A. Borovkov. Some limit theorems in the theory of mass service ii. *Theor. Probability Appl.*, 10:375–400, 1965.

[11] P. Brémaud. *Point Processes and Queues, Martingale Dynamics*. Springer-Verlag, New York, 1981.

[12] R. Buche and H.J. Kushner. Control of mobile communications with time-varying channels in heavy traffic. *IEEE Transactions on Automatic Control*, 47(6):992–1003, 2002.

[13] 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.

[14] X. Chao. A queueing network model with catastrophes and product form solution. *Operations Research Letters*, 18:75–79, 1995.

[15] J. Fourneau, E. Gelenbe, and R. Suros. G-networks with multiple classes of negative and positive customers. *Theoretical Computer Science*, 155(1):141–156, 1996.

[16] J. Fourneau and D. Verchère. G-networks with triggered batch state-dependent movement. In *MASCOTS '95: Proceedings of the 3rd International Workshop on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 33–37, Washington, DC, USA, 1995. IEEE Computer Society.

[17] D.P. Gaver. Diffusion approximations and models for certain congestion problems. *Journal of Applied Probability*, 5:607–623, 1968.

[18] E. Gelenbe. On approximate computer system models. *Journal of the Association for Computing Machinery*, 22(2):261–269, 1975.

[19] E. Gelenbe. Random neural networks with negative and positive signals and product form solution. *Neural Computation*, 1(4):502–511, 1989.

[20] E. Gelenbe. Product-form queueing networks with negative and positive customers. *Journal of Applied Probability*, 28:656–663, 1991.

[21] E. Gelenbe. G-networks with triggered customer movement. *Journal of Applied Probability*, 30:742–748, 1993.

[22] E. Gelenbe. G-networks: a unifying model for neural and queueing networks. *Annals of Operations Research*, 48(5), 1994.

[23] E. Gelenbe. Steady-state solution of probabilistic gene regulatory networks. *Physical Review E*, 76:031903–1–031903–8, 2007.

[24] E. Gelenbe and J. Fourneau. G-networks with resets. *Performance Evaluation*, 49(1-4):179–191, 2002.

[25] E. Gelenbe and K.F. Hussain. Learning in the multiple class random neural network. *IEEE Transactions on Neural Networks*, 13(6):1257–1267, 2002.

[26] E. Gelenbe, X. Mang, and R. Onvural. Diffusion based statistical call admission control in atm. *Performance Evaluation*, 27-28:411–436, 1996.

[27] E. Gelenbe, X. Mang, and R. Onvural. Bandwidth allocation and call admission control in high-speed networks. *IEEE Communications Magazine*, 35(5):122–129, 1997.

[28] E. Gelenbe and G. Pujolle. The behaviour of a single queue in a general queueing network. *Acta Informatica*, 7:123–136, 1976.

[29] E. Gelenbe and R. Schassberger. Stability of product form g-networks. *Probability in the Engineering and Informational Sciences*, 6:271–276, 1992.

[30] E. Gelenbe and A. Stafylopatis. Global behavior of homogeneous random neural systems. *Applied Mathematical Modelling*, 15:534–541, 1991.

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

[32] A. Gómez-Corral and M.E. Martos. Performance of two-stage tandem queues with blocking: the impact of several flows of signals. *Performance Evaluation*, 63(9):910–938, 2006.

[33] V. Guffens, E. Gelenbe, and G. Bastin. Qualitative dynamical analysis of queueing networks with inhibition. In *interperf '06: Proceedings from the 2006 workshop on Interdisciplinary systems approach in performance evaluation and design of computer & communications sytems*, 2006.

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

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

[36] W. Henderson, B.S. Nothcote, and P.G. Taylor. State dependent signalling in queueing networks. *Advances in Applied Probability*, 26:436–455, 1994.

[37] D.L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, i. *Adv. Appl. Prob.*, 2:150–177, 1970.

[38] D. Jarvis and H.J. Kushner. Codes for optimal stochastic control: documentation and users guide. Technical report, Brown University, Lefschetz Center for Dynamical Systems Report 96-3, 1996.

[39] J.F.C. Kingman. The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.*, 57:902–904, 1961.

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

[41] H.J. Kushner and P.G. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer–Verlag, New York, 1992.

[42] H.J. Kushner and L.F. Martins. Heavy traffic analysis of a data transmission system with many independent sources. *SIAM Journal on Applied Mathematics*, 53(4):1095–1122, 1993.

[43] H.J. Kushner, J. Yang, and D. Jarvis. Controlled and optimally controlled multiplexing systems: A numerical exploration. *Queueing Systems*, 20:255–291, 1995.

[44] S.C. Leite and M.D. Fragoso. Diffusion approximation of state dependent g-networks under heavy traffic. LNCC Internal Report No. 21/2008, 2008.

[45] Q. Li and Y.Q. Zhao. A map/g/1 queue with negative customers. *Queueing Systems*, 47:5–43, 2004.

[46] A. Mandelbaum and G. Pats. State-dependent stochastic network. part i: Approximations and applications with continuous diffusion limits. *The annals of Applied Probability*, 8:569–646, 1998.

[47] Yu. Prohorov. Transient phenomena in process of mass service. *Litovsk. Mat. Sb.*, 3:199–205, 1963. (In Russian).

[48] M.I. Reiman. Open queueing networks in heavy traffic. *Mathematics of Operations Research*, 9(3):441–458, 1984.

[49] Y.W. Shin. Multi-server retrial queue with negative customers and disasters. *Queueing Systems*, 55:223–237, 2007.

[50] W. Whitt. Complements to heavy traffic limit theorems for the $gi/g/1$ queue. *J. Appl. Prob.*, 9:185–191, 1972.

[51] W. Whitt. Heavy traffic limit theorems for queues: A survey. In M. Beckmann and H.P. Kunzi, editors, *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, New York, 1974.