43rd IEEE Conference on Decision and Control
December 14-17, 2004
Atlantis, Paradise Island, Bahamas

Fluid Approximation and Perturbation Analysis of a Dynamic Priority Call Center

Min Chen, Jian-Qiang Hu, Michael C. Fu

Abstract— **We analyze a call center with multiclass calls and dynamic priority service discipline, in which a lower priority customer becomes high priority when its waiting time exceeds a given service level threshold. For each priority queue, the service discipline is first come, first served. Based on a fluid approximation of the system, we apply infinitesimal perturbation analysis (IPA) to derive estimators for the derivative of the queue lengths with respect to the threshold parameter. We establish unbiasedness of the estimators, and report numerical results via simulation.**

I. Introduction

A call center is a network of customer service representatives and the physical infrastructure needed to provide services to customers from remote sources. In traditional call centers, customers access the call center by talking to customer service representatives directly over the phone. However, modern call centers have become more complicated, both in size and in operational complexity, allowing additional communication channels, e.g., automatic answering systems, e-mails, faxes. Along with the added complexity, more demanding service requirements have created new challenges in modeling, analysis, and design of call centers. As an alternative to the traditional modeling approach using discrete queueing models, fluid models provide a simpler and more efficient approach for performance evaluation. In order to capture the stochastic characteristic of queueing networks, the idea of stochastic fluid models (SFMs) was introduced by Anick, D.Mitra, and Sondhi [4], which has been followed by many developments in queueing network analysis, such as the stability of queuing networks (see [10], [11], [14] and [15]) and fluid approximations of multiclass queueing networks (see [6], [7], [8] and [9]). The fluid model can provide a first-order approximation of the corresponding queueing system, leading to heuristics for controlling and scheduling communication networks or manufacturing systems.

In call center systems, quality of service (QoS) is a critical measure of performance. For example, in [20], the objective of the call center management is that 80% of the calls are answered within 20 seconds, and no more than 3% of the calls are abandoned before reaching a representative. However, modern call centers have multiple classes of customers, e.g., live calls, e-mail, and fax, along with "preferred" or "elite" customers, for which the QoS requirements are more demanding. In order to guarantee differentiated QoS requirements for multiple classes of customers, different queueing disciplines may be required (see [18], [12] and [13] for priority fluid models, [9] for a process-sharing model, and [1], [17] for first come, first served).

Strict priority is often used to classify customer classes in call centers, in order to meet the more stringent requirements of high priority customers. However, during periods of heavy traffic, this may lead to difficulties in meeting the QoS requirements for the lower priority customers, i.e., the waiting times for lower priority customers may become unacceptably large. With this in mind, many call centers have adopted dynamic priorities, in order to avoid excessive waiting for the low priority customers. One of the simplest versions is to upgrade a low priority customer to a higher priority if its waiting time exceeds some service level threshold θ. This would decrease the average waiting time of the low priority customers at the cost of some increase for high priority customers, but by choosing the parameter θ appropriately, the overall QoS may be improved. A fluid model for this type of call center is first introduced in [2], where the threshold is an exponentially distributed random variable, and not a fixed constant. As a result, the model is Markovian, and results from [6] and [7] can be applied. In this paper, we develop a fluid model of the dynamic priority call center for the case where the threshold is a fixed constant. In addition to developing estimators for the average queue lengths (or average waiting time), we also

Min Chen and Jian-Qiang Hu are with Department of Manufacturing Engineering & Center for Information and Systems Engineering (CISE), Boston University, Brookline, MA 02446, USA. They were supported in part by the National Science Foundation under Grant EEC-0088073. anthem16@bu.edu, hqiang@bu.edu

Michael C. Fu is with The Robert H. Smith School of Business & Institute for Systems Research, University of Maryland, College Park, MD 20742, USA. He was supported in part by the National Science Foundation under Grants DMI-9988867 and DMI-0323220, and by the Air Force Office of Scientific Research under Grant F496200110161. mfu@rhsmith.umd.edu

derive estimators for the sensitivity of the performance measure with respect to the threshold parameter θ using infinitesimal perturbation analysis (IPA). The dynamic priority makes this fluid model quite different from previous fluid models in the literature. This uniqueness makes this problem very interesting from the perspective of sample path analysis, which yields some insightful properties. Moreover, by simulation, we see that this fluid model is a good approximation of the call center of interest when the arrival rate and service rate are scaled up proportionally.

The remainder of this paper is organized as follows. In section 2, we introduce the discrete model of the dynamic priority call center and develop the corresponding fluid model. In section 3, we derive the IPA estimators of the queue lengths with respect to the waiting time threshold. In section 4, we prove the unbiasedness of the estimators. Numerical results and computational results are reported in section 5, followed by conclusions and future work in section 6.

II. Fluid Model for a Dynamic Priority Call Center

For the purpose of performance analysis, call centers have been traditionally modeled as the discrete queuing networks. However, for more complicated system, the analysis of the queuing networks is not only inefficient but also intractable. A reasonable direction is to provide tractable "relaxations" of the problem via higher abstraction level models that are simple, e.g., the fluid model approach we take here.

The dynamic priority call center under consideration, introduced in [2], consists of two classes of arriving customers: low priority customers and high priority customers, who have preemptive-resume priority over the low priority customers. A low priority call that has not completed service within a given amount of time (namely, the service level threshold θ) is upgraded to the higher priority. Thus, the low priority customers have dynamic priority.

In [2], under the assumption that the threshold θ is exponentially distributed, a fluid approximation of this call center is given, and the performance of the fluid model is compared with the actual performance, estimated by discrete event simulation. This assumption simplifies the analysis because it leads to a Markovian model, whose limit is the fluid model developed in [2]. However, this assumption allows later arriving customers of the same low priority class to overtake earlier customers. In most call centers, customers with the same priority are served in the sequence they enter the system, so we use first come, first served queueing discipline for customers within each priority class. Most importantly, we take the service level threshold θ to be a constant. One key problem in call center design is to determine an appropriate value for the threshold parameter θ.

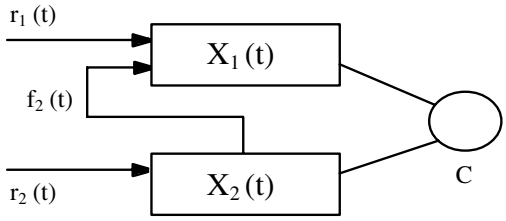

Fig. 1. Fluid model of call center with dynamic priority

Using a deterministic threshold leads to a non-Markovian system, which makes the fluid model more complicated. A diagram of the call center fluid model is shown in Fig.1. We assume infinite capacity for both buffers. Denote the class 1 (high priority) arrival rate at time t by $r_1(t)$, the class 2 (low priority) arrival rate at time t by $r_2(t)$, let C be the maximum outflow rate (service rate), and assume $r_1(t) \neq C.w.p.1$, so that when the 1st buffer is empty and $r_2(t) \neq 0$, there must be some service that is allocated to the 2nd buffer. We can normalize the maximum outflow rate of fluid server C as 1. The flow in the 2nd buffer will enter the end of the 1st buffer after its waiting time exceeds the threshold θ, which is a constant. Let $f_2(t)$ be the flow rate from the 2nd buffer to the 1st one. The service will be allocated only to the 1st buffer if it is not empty, and will allocate the remaining capacity to the 2nd buffer otherwise. Let $X_1(t)$ and $X_2(t)$ be the total queue lengths of the 1st and 2nd buffer at time t, respectively, with $c_1(t)$ and $c_2(t)$ the corresponding service rates. Let $X_1^1(t)$ and $X_1^2(t)$ be the queue lengths of the respective class 1 and upgraded class 2 flows in the 1st buffer at time t, with $c_1^1(t)$ and $c_1^2(t)$ the corresponding service rates. Our goal is to estimate the average flow lengths of class 1 and 2 customers in each buffer.

A. Dynamic Equations of the Fluid Model

Before we formulate the dynamic equations, we note that in the 1st buffer, typically there are two classes of customers: high priority customers and upgraded low priority customers. Since the queue discipline is first-come, first-served, it is necessary to capture the characteristics of a first-in, first-out (FIFO) fluid model.

For the FIFO fluid model (e.g., see [1]), we have

$$c_1^1(t+W_1(t)) = \frac{r_1(t)}{r_1(t)+f_2(t)}c_1(t+W_1(t)), \quad (1)$$

$$c_1^2(t+W_1(t)) = \frac{f_2(t)}{r_1(t)+f_2(t)}c_1(t+W_1(t)), \quad (2)$$

where $W_1(t) = X_1(t)/C$.

The basic dynamics of the fluid model are described by the following equations:

$$X_1(t) = X_1(0) + \int_0^t (r_1(s) + f_2(s) - c_1(s))ds,$$

$$X_2(t) = X_2(0) + \int_0^t (r_2(s) - f_2(s) - c_2(s))ds,$$

$$X_1^1(t) = X_1^1(0) + \int_0^t (r_1(s) - c_1^1(s))ds,$$

$$X_1^2(t) = X_1^2(0) + \int_0^t (f_2(s) - c_1^2(s))ds,$$

$$c_1(t) = \begin{cases} C & \text{if } X_1(t) \neq 0, \\ r_1(t) & \text{otherwise,} \end{cases}$$

(Note that if $X_1(t) = 0, f_2(t) = 0$.)

$$c_2(t) = \begin{cases} 0 & \text{if } X_1(t) \neq 0, \\ C - r_1(t) & \text{if } X_1(t) = 0 \quad X_2(t) \neq 0, \\ r_2(t) & \text{otherwise.} \end{cases} \quad (3)$$

Next we try to express $f_2(t)$ in terms of the other known functions. Consider the class 2 flow entering the 2nd buffer at time period $(t, t+dt)$, where dt is small enough to guarantee that this small amount of flow $r_2(t)dt$ will be served in the same buffer w.p.1. Let $\tau(t)$ be the system time (waiting time plus service time) of this small amount of flow $r_2(t)dt$. This flow has two ways to be served: it receives service from buffer 2 at time $t+\tau(t)$, so that $\tau(t) < \theta$, or it receives service in buffer 1 at time $t+\tau(t)$, and $\tau(t) > \theta$:

$$\tau(t) = \begin{cases} \tau & \tau \leq \theta, \\ \theta + \dfrac{X_1(t+\theta)}{C} & \tau > \theta, \end{cases} \quad (4)$$

where $\tau = \inf_{a \geq 0}\{a : a = \dfrac{X_1(t) + \int_t^{t+a} r_1(s)ds + X_2(t)}{C}\}$. The condition in (4) indicates whether or not a class 2 customer arriving at t would be upgraded at $t+\theta$.

We denote the function $g(t,\theta)$ to be the $g(t)$ in the sample path under the parameter θ. Thus, we have

$$f_2(t,\theta) = r_2(t-\theta)1\{\tau(t-\theta,\theta) > \theta\}. \quad (5)$$

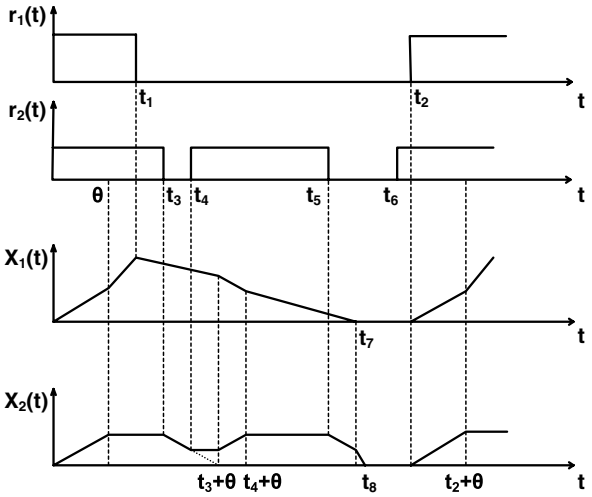

Fig. 2. A typical sample path of this fluid model

This fluid model is described mathematically by (1)-(5), from which we can generate sample paths of the system. In the next section, we will use infinitesimal perturbation analysis (IPA) to derive sample path derivative estimators of the queue lengths with respected to the service level threshold θ.

III. Infinitesimal Perturbation Analysis

Infinitesimal perturbation analysis (IPA) (see [5]) was applied to stochastic fluid models in [3] and [16], in which a simple threshold-based buffer control policy was proposed. In the fluid model of the dynamic priority call center, we can control system performance metrics like queue length by adjusting the service level threshold θ. The goal of this section is to estimate the derivative of the queue lengths with respect to the parameter θ.

For simplicity, the following assumptions are made: $r_1(t)$ is piecewise continuous, $r_2(t)$ is piecewise constant, and both are bounded from above.

For the function $g(t,\theta)$, we define

$$\Delta g(t,\theta) = g(t,\theta+\Delta\theta) - g(t,\theta) = \bar{g}(t) - g(t),$$

where $g(t,\theta)$ is the parametric function in the original sample path, and $g(t,\theta+\Delta\theta)$(or $\bar{g}(t)$) denotes the one in the perturbed sample path. Then $\Delta g(t,\theta)$ is the difference between the two.

For illustrative purpose, we present in Fig.2 a sample path of the fluid system in which both $r_1(t)$ and $r_2(t)$ are ON-OFF processes. Let r_1 (resp. r_2) be the inflow rate of $r_1(t)$ (resp. $r_2(t)$) when it is in ON state. In OFF state, both $r_1(t)$ and $r_2(t)$ are equal to zero. For simplicity,

we assume $r_1 > C > r_2$(recall C is the constant service rate). In the sample path shown in Fig.2, at time θ, the flow that enters the $2nd$ queue at time 0 still remains in the $2nd$ queue; therefore, it is upgraded and is moved to the end of the 1st queue. This implies that the upgrading rate at time θ is equal to $r_2(0)$, i.e., the inflow rate of the $2nd$ queue at time 0. This delay effect causes the drift rate (inflow rate - outflow rate) changes for both queues at time θ. For the same reason we conclude that the flow entering the $2nd$ queue during period $(0, t_3)$ will be upgraded during period $(\theta, t_3 + \theta)$, while the flow entering at time t_4 will be upgraded at time $t_4 + \theta$. And during period $(t_3 + \theta, t_4 + \theta)$, no flow will be upgraded. After time t_7, as the $1st$ queue becomes empty and its source is in OFF state, the server can serve the $2nd$ queue at its full capacity, therefore there is no upgrading during (t_7, t_8).

At time t_6, the source of the $2nd$ queue switches to ON state, but the queue remains empty until t_2, when the source of the $1st$ queue switches to ON state. Obviously, the flow in the $2nd$ queue that is not being served from time t_2 will be upgraded at $t_2 + \theta$.

We define the Busy Period (BP) as an interval in which the queue length is not 0, and Empty Period (EP) otherwise. From now on, unless otherwise noted, the BP(or EP) is defined for the system, i.e., BP is the period in which at lease one queue length is not 0. Define the Full Period (FP) as an interval of BP, in which the $2nd$ buffer is forwarding the flow into the $1st$ buffer, and None Full Period (NFP) otherwise. For instance, in Fig.2, $(0, t_8)$ is a BP, (t_8, t_2) is a EP. $(\theta, t_3 + \theta)$ and $(t_4 + \theta, t_7)$ are 2 FPs, while the other periods are $NFPs$. Consider a typical sample path, which is composed of alterntating BPs and EPs. As $X(t) = \int_0^t r_1(s) + r_2(s) - c(s)ds$, which is independent of the threshold θ, so the BPs and EPs are independent of θ, e.g., in Fig.2, t_8 is independent of θ. Therefore, no perturbation of θ could be propagated from one BP to another. So typically, we only need to consider one BP, starting with the end of the previous EP. This BP may consist of several FPs and $NFPs$.

Next consider the time when an FP ends. There are two possibilities that could cause the ending of an FP:

1. $X_1(t)$ becomes 0 at time t, e.g., t_7 in Fig.2. It is obvious that FP ends at time t. After time t, the system is back to NFP, so that θ will have no effect on the buffer contents until the end of this BP or the beginning of the next FP in the same BP, whichever starts first.

2. At time t, $r_2(t^- - \theta) \neq 0$ and $r_2(t^+ - \theta) = 0$. Then at time t, even if $X_1(t)$ is not 0, the $2nd$ class flow that enters the system later than $t - \theta$ has not reached the service level threshold θ, so no flow at the $2nd$ buffer is upgraded to the $1st$ buffer, then FP ends. For example in Fig.2, at time $t_3 + \theta$, FP ends and the flow entering the $2nd$ queue at time t_4 can only be upgraded at time $t_4 + \theta$. This effect causes the NFP $(t_3 + \theta, t_4 + \theta)$.

For the first case, during the following NFP period, θ has no effect on the queue lengths, until the next FP starts. For the second case, although $r_2(t)$ becomes 0 at $t - \theta$, we treat the system as if the FP does not end, i.e., the flow rate at $t^+ - \theta$ is $r_2(t^+ - \theta) = 0$, and this zero rate flow enters the low priority buffer, waits for the service, and is then upgraded to the high priority buffer at time t. E.g, in Fig.2, during (t_3, t_4), the flow (with rate $r_2(t) = 0$) enters the $2nd$ queue, and is upgraded to the $1st$ queue during $(t_3 + \theta, t_4 + \theta)$. Then (θ, t_7) becomes an FP in the current BP. This eliminates the second case that causes the end of FP, without affecting the analytical and numerical results of the model.

Based on the above discussions, each FP starts with the flow from the $2nd$ queue being upgraded and ends when the $1st$ queue becomes empty. The perturbation of θ won't be propagated from one FP to another. Hence, for the purpose of our perturbation analysis, when considering the time t in a typical BP, we can focus on the most recent FP in the same BP, since each FP is independent of the others in the sample path. Without loss of generality, we assume $(0, t)$ does not contain the end of the FP. Additionally, we denote the $\overline{FP}$ as the FP of the perturbed sample path. We now state the following results (their proofs can be found in [21])

Lemma 1 $X(t)$ is a continuous function on t, and independent of θ.

Lemma 2 For small enough $\Delta\theta$, w.p.1, $\Delta f_2(t)$ is 0 for all the other t except for the period $(FP \cap \overline{NFP}) \cup (NFP \cap \overline{FP})$.

Theorem 1 The sample derivatives of $X_1(t, \theta)$ and $X_2(t, \theta)$ with respect to θ satisfy:

$$\frac{\partial X_1(t,\theta)}{\partial \theta} = \begin{cases} -r_2(z-\theta) & \text{if } t \in FP, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

$$\frac{\partial X_2(t,\theta)}{\partial \theta} = \begin{cases} r_2(z-\theta) & \text{if } t \in FP, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where z is the starting time point of the FP that time t belongs to.

Theorem 2 The sample derivatives of $X_1^1(t)$ and $X_1^2(t)$ satisfy

$$\frac{\partial X_1^1(t,\theta)}{\partial \theta} = \begin{cases} -r_2(z-\theta)c_1^1(t) \\ \qquad \text{if } t > z + X_1(z) \text{ and } t \in FP \\ 0 \qquad \text{otherwise,} \end{cases} \tag{8}$$

$$\frac{\partial X_1^2(t,\theta)}{\partial \theta} = \begin{cases} -r_2(z-\theta)c_1^2(t) \\ \qquad \text{if } t > z + X_1(z) \text{ and } t \in FP \\ -r_2(z-\theta) \\ \qquad \text{if } t < z + X_1(z) \text{ and } t \in FP, \\ 0 \qquad \text{otherwise.} \end{cases} \tag{9}$$

where z is the starting time point of the FP that time t belongs to.

IV. Unbiasedness of the estimator

In this section, the unbiasedness of the estimator derived from above section will be proven. In general, the unbiasedness of an IPA derivative $\frac{d}{d\theta}J(\theta)$ has been shown to be ensured by the following two conditions (See [19]):

Condition 1 For every $\theta \in \Theta$, the sample derivative $\frac{d}{d\theta}J(\theta)$ exists w.p.1.

Condition 2 W.p.1, the random function $\frac{d}{d\theta}J(\theta)$ is Lipschitz continuous throughout Θ, and the Lipschitz constant has finite first moment.

Theorem 3 The sample derivatives of $X_1(t)$, $X_2(t)$, $X_1^1(t)$, and $X_1^2(t)$ with respect to θ are unbiased estimators of $dE(X_1(t,\theta))/d\theta$, $dE(X_2(t,\theta))/d\theta$, $dE(X_1^1(t,\theta))/d\theta$, and $dE(X_1^2(t,\theta))/d\theta$, respectively.

V. Numerical Example

In the previous sections, $r_1(t)$ and $r_2(t)$ are allowed to be any piecewise continuous and piecewise constant functions, respectively. Here, we consider an example in which the 1st class flow is an ON-OFF piecewise constant flow with flow rates r_1 and 0, and the ON (OFF) period exponentially distributed with rate $\mu(\lambda)$, and the 2nd class flow is a constant with flow rate r_2. From the equations (3) ,(1), (2), (4) and (5), we see that $f_2(t)$ can take on two values r_2 or 0, whereas $c_1^1(t)$ can take on three possible values: $1, 0, \frac{r_1}{r_1+r_2}$. The following equations are straightforward from the previous section:

$$\frac{\partial X_1(t,\theta)}{\partial \theta} = \begin{cases} -r_2 & \text{if } t \in FP, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

$$\frac{\partial X_2(t,\theta)}{\partial \theta} = \begin{cases} r_2 & \text{if } t \in FP, \\ 0 & \text{otherwise,} \end{cases}$$

$$\frac{\partial X_1^1(t,\theta)}{\partial \theta} = \begin{cases} -r_2 c_1^1(t) \\ \qquad \text{if } t > z + X_1(z) \text{ and } t \in FP \\ 0 \qquad \text{otherwise,} \end{cases}$$

$$\frac{\partial X_1^2(t,\theta)}{\partial \theta} = \begin{cases} -r_2 c_1^2(t) \\ \qquad \text{if } t > z + X_1(z) \text{ and } t \in FP \\ -\ r_2 \\ \qquad \text{if } t < z + X_1(z) \text{ and } t \in FP, \\ 0 \qquad \text{otherwise} \end{cases} \tag{11}$$

where z is the starting point of the current FP.

Table I gives the results comparing the IPA estimator with the finite difference estimator.

θ	class	IPA Result	Finite difference	Error
0.5	1st	-0.698($\pm$0.012)	-0.696($\pm$0.009)	0.2%
	upgrade 2nd	-0.175($\pm$0.008)	-0.177($\pm$0.012)	0.2%
	2nd	0.873($\pm$0.002)	0.873($\pm$0.002)	0%
1	1st	-0.692($\pm$0.017)	-0.689($\pm$0.17)	0.4%
	upgrade 2nd	-0.173($\pm$0.010)	-0.176($\pm$0.009)	0.4%
	2nd	0.866($\pm$0.002)	0.866($\pm$0.001)	0%
1.5	1st	-0.664($\pm$0.018)	-0.674($\pm$0.012)	1.5%
	upgrade 2nd	-0.166($\pm$0.004)	-0.160($\pm$0.007)	3.6%
	2nd	0.830($\pm$0.003)	0.834($\pm$0.002)	0.5%
2	1st	-0.680($\pm$0.010)	-0.674($\pm$0.005)	0.9%
	upgrade 2nd	-0.170($\pm$0.006)	-0.179($\pm$0.004)	5.3%
	2nd	0.850($\pm$0.003)	0.853($\pm$0.002)	0.35%

TABLE I
IPA RESULTS FOR NUMERICAL EXAMPLE
(STANDARD ERRORS IN PARENTHESES)

VI. Conclusions and Future Work

We developed a fluid model of a dynamic priority call center scheduler, and applied IPA to obtain derivative estimators of queue lengths with respect to the service level threshold of low priority calls. Still in progress is a rigorous proof establishing that the scaled limit of the discrete queueing model is in fact the fluid model we have proposed; numerical simulations seem to support this claim. Other future research directions include extending the results of the single fluid queue to fluid networks, and applying the approach to communication networks.

References

[1] M. Bramson. "Convergence to equilibria for fluid models of certain FIFO and processor sharing queueing networks," *Queueing Systems: Theory and Applications,* Vol. 22, pp. 5-45 1996.

[2] A.D. Ridley, M.C. Fu, and W.A. Massey, "Fluid approximations for a priority call center with time-varying arrivals," in *Proceedings of the 2003 Winter Simulation Conference,* pp. 1817-23, 2003.
[3] C.G. Cassandras, Y. Wardi, B. Melamed, G. Sun, and C.G. Panayiotou, "Perturbation Analysis for On-Line Control and Optimization of Stochastic Fluid Models," *IEEE Trans. on Automatic Control,*, vol. 47, no. 8, pp. 1234-1248, 2002.
[4] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Probability,* Vol. 20, pp. 646-676, 1988.
[5] Y.C. Ho and X.R. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems,* Kluwer Academic Publishers, 1991.
[6] A. Mandelbaum, W.A. Massey, and M.I. Reiman, "Strong Approximation for markovian networks," *Queueing Systems: Theory and Applications,* Vol. 30 pp. 149-201, 1998.
[7] A. Mandelbaum and W.A. Massey, "Strong approximation for time dependent queues," *Mathematics of Operations Research,* Vol. 20, No. 1, pp. 33-64, 1995.
[8] H. Chen, "Fluid approximations and stability of multiclass queueing networks: work-conserving disciplines," *Annals of Applied Probability*, No. 5, pp. 636-665, 1995.
[9] H. Chen, "Fluid approximations for a process-sharing queue," *Queueing Systems: Theory and Applications,* No. 27, pp. 99-125, 1997.
[10] J.G. Dai, "On positive Harris Recurrence of multiclass queueing networks: A unified approach via fluid limit models," *Annals of Applied Probability*, Vol. 5, pp. 49-77 1995.
[11] J.G. Dai, "Stability of open multiclass queueing networks via fluid models," *Stochastic Networks,* F. Kelly and R.J. Williams (Editors), pp. 71-90, Springer-Verlag, 1995.
[12] Y. Liu and W.B.Gong, "On Fluid Queueing System with Strict Priority," *IEEE Trans. on Automatic Control,* December, 2003.
[13] Y.Liu and W.B.Gong, "Perturbation Analysis for Stochastic Fluid Queueing Systems," *Discrete Event Dynamic Systems: Theory and Applications,* 12, 391-416, 2002.
[14] D. Gamarnik, "Using fluid models to prove stability of adversarial queueing networks," *IEEE Trans. Automat. Control,* vol. 45, pp. 741-746, 2000.
[15] D. Bertsimas, D. Gamarnik, and J.N. Taitaiklia, "Stability conditions for multiclass fluid queueing networks," *39th Symposium on Foundations of Computer Science,* pp. 1618-1631, 1998.
[16] G. Sun, *Perturbation Analysis and On-Line Control for Discrete Event Systems via Stochastic Fluid Models,* Ph.D Thesis, Department of Manufacturing Engineering, Boston University, 2003.
[17] H. Chen and H.Q. Zhang, "Stability of multiclass fluid networks under FIFO service discipline," *Mathematics of Operations Research,* 22, pp 691-725, 1997.
[18] J. Zhang, "Performance Study of Markov Modulated Fluid Flow Models with Priority Traffic," *INFOCOM 1993,* pp. 10-17, 1993.
[19] P. Glasserman, *Gradient Estimation via Perturbation Analysis,* Kluwer Academic Publishers, 1991.
[20] G. Koole, *Call center mathematics A scientific method for understanding and improving contact centers*, Oct 2003.
[21] M. Chen, J.Q. Hu, and M.C. Fu, *Fluid Approximation and Perturbation Analysis of a Dynamic Priority Call Center,* Boston University Technical Report, Department of Manufacturing Engineering, Boston University, 2004. (url: http://lagrange.bu.edu/papers/CDCfluid.pdf).