A Fluid Approximation for Large-Scale Service Systems

Yunan Liu, Ward Whitt
Department of Industrial Engineering and Operations Research
Columbia University
500 West 120th Street
New York, NY 10027-6699
{yl2342,ww2040}@columbia.edu

ABSTRACT

We introduce and analyze a deterministic fluid model that serves as an approximation for the $G_t/GI/s_t + GI$ many-server queueing model, which has a general time-varying arrival process (the G_t), a general service-time distribution (the first GI), a time-dependent number of servers (the s_t) and allows abandonment from queue according to a general abandonment-time distribution (the $+GI$). This fluid model approximates the associated queueing system when the arrival rate and number of servers are both large. We also show that the system dynamics greatly simplifies in two special cases: (i) when the service time distribution is exponential (M) and (ii) when the service time distribution is deterministic (D) and the model is stationary. We develop an efficient algorithm to compute all standard performance functions in both cases.

In case (i), we establish an *asymptotic loss of memory* (ALOM) property, i.e., asymptotic independence from the initial conditions as time evolves. We show that the difference in the performance functions with different initial conditions dissipates over time exponentially fast, under regularity conditions. In contrast, in case (ii) we show that ALOM fails dramatically. Instead, although all model parameters are constants, we show that the performance rapidly approaches a *periodic steady state* (PSS) with a period equal to the service time, whenever the system does not start with the unique stationary distribution. Moreover, the form of the PSS depends on the initial condition. Simulation and a heavy-traffic limit confirm that this anomalous behavior also occurs in the large-scale queueing model.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Queueing Theory

General Terms

Performance, Theory

Keywords

large-scale service systems, queues with time-varying arrivals; nonstationary queues; many-server queues; customer abandonment; deterministic fluid model; non-Markovian queues, asymptotic loss of memory, periodic steady state

1. INTRODUCTION

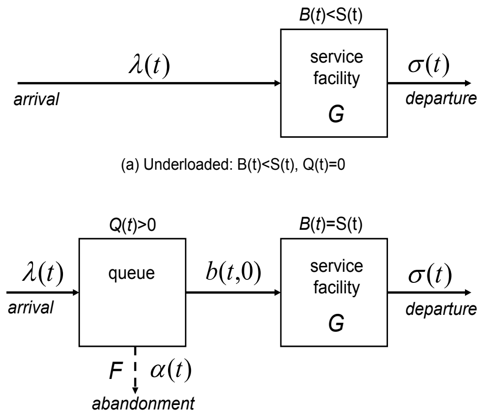
Motivated by the need for tools to improve the performance of large-scale service systems, such as customer contact centers (telephone call centers) and hospitals, we introduce and analyze a deterministic fluid model that serves as an approximation for a many-server non-Markovian queueing model with time-varying parameters; see [1] and [3] for background on contact centers, see [15] for hospitals. Large-scale service systems tend to be quite complicated, with multiple classes of customers and multiple pools of servers. Here we restrict attention to the basic model with a single class of customers handled by a single group of homogeneous servers, working in parallel.

The motivating queueing model is $G_t/GI/s_t+GI$, which has a general time-varying arrival process (the G_t) partially characterized by its arrival rate function, independent and identically distributed (i.i.d.) service times with a general service-time distribution (the first GI), a time-varying (deterministic) staffing function (number of servers, the s_t), and abandonment allowed from queue according to i.i.d. patience times with a general distribution (the $+GI$). The stochastic model has four complicating features that make it difficult to analyze directly: (i) time varying arrival rate and staffing function, (ii) customer abandonment, (iii) large scale (many servers) and (iv) non-Markovian probability structure. We aim to provide tools that will increase understanding of these complicating features and thereby help improve system performance. For setting appropriate staffing policies to cope with time-varying demands in meeting performance goals subject to performance constraints, see [4, 6]. See [2, 5, 12] for recent development.

The fluid model is intended to serve as an approximation for the queueing model when both the number of servers and the arrival rate are large, and the system experiences occasional periods of significant overloading. Such approximations are usually justified by many-server heavy-traffic limits as the arrival rate and number of servers increase, e.g., see [13], but this work is concerned with analyzing the performance of the approximating fluid model. Specifically, here we provide an overview of our recent work in [7]-[11]. In §2, we briefly describe the transient performance of the $G_t/GI/s_t+GI$ fluid queue. In §3, we discuss the *asymptotic loss of memory* (ALOM) property of the $G_t/M/s_t+GI$ fluid queue. In §4, we discuss the convergence to a *periodic steady state* (PSS) in the $GI/D/s+GI$ fluid queue with constant model parameters.

2. THE MODEL AND ITS PERFORMANCE

In the $G_t/GI/s_t + GI$ fluid model there is a deterministic arrival process with arrival-rate function $\lambda \equiv \{\lambda(t) : t \geq 0\}$; i.e., the total input of fluid over the interval $[0, t]$



(a) Underloaded: B(t)<S(t), Q(t)=0

(b) Overloaded: B(t)=S(t), Q(t)>0

Figure 1: Flows in overloaded and underloaded intervals.

is $\Lambda(t) \equiv \int_0^t \lambda(u)\,du$, $t \geq 0$. There is a staffing (service capacity) function $s \equiv \{s(t) : t \geq 0\}$. There are service-time and abandon-time cdf's G and F, respectively, with pdf's g and f. These cdf's apply deterministically yielding proportions, i.e., a proportion $G(x)$ ($F(x)$) of the fluid entering service (the queue) completes service (abandons) and departs by time x after it has entered. There is unlimited waiting space and the service discipline is FCFS.

The transient dynamics can be characterized by two two-parameter performance functions: $Q(t,y)$ and $B(t,y)$, where $Q(t,y)$ ($B(t,y)$) is the quantity of fluid in queue (in service) at time t that has been in queue for time less than or equal to y. Under regularity conditions, these functions are absolutely continuous in the second argument, i.e.,

$$Q(t,y) = \int_0^y q(t,x)dx \quad \text{and} \quad B(t,y) = \int_0^y b(t,x)dx, \quad y \geq 0,$$

where the fluid density in queue b and the density in service q are non-negative integrable functions that satisfy the following fundamental evolution equations:

$$\begin{aligned} q(t+u, x+u) &= q(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)}, \quad 0 \leq x < w(t), \\ b(t+u, x+u) &= b(t,x)\frac{\bar{G}(x+u)}{\bar{G}(x)}. \end{aligned}$$

Under regularity conditions, we show that the head-of-line waiting time w is a non-negative continuous function that solves the following ODE:

$$w'(t) = 1 - \frac{b(t,0)}{q(t,w(t))}.$$

We provide an algorithm to compute the transient dynamics of $b(t,y)$ and $q(t,y)$ for $0 \leq t \leq T$, assuming that the system alternates between underloaded intervals (with $B(t) \leq s(t)$) and overloaded intervals (with $Q(t) > 0$ and $B(t) = s(t)$). Theorem 3.1 in [8] shows that the two-parameter density function $b(t,x)$ solves a fixed-point equation in each overloaded interval. The two density functions b and q can be used to determine all performance measures, such as the total fluid in queue $Q(t)$ and in service $B(t)$, the potential waiting time $v(t)$ (the virtual waiting time of an infinitely patient arrival), the abandonment rate $\alpha(t)$ and the departure rate $\sigma(t)$. Figure 1 illustrates the flow dynamics in different regimes.

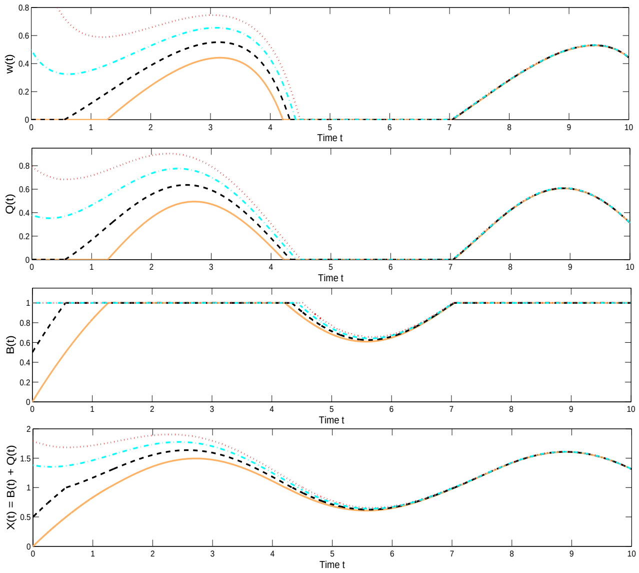


Figure 2: The performance measures of a $G_t/M/s{+}M$ example with different initial conditions.

3. EXPONENTIAL SERVICE TIMES

The algorithm in [8] greatly simplifies when the service time distribution is exponential (M). Then the fixed-point equation for $b(t,x)$ has an explicit solution.

We illustrate the algorithm by considering an example with a sinusoidal arrival rate $\lambda(t) = 1 + 0.6\sin(t)$, constant staffing $s(t) = 1$, exponential abandonment time with rate $\theta = 0.5$ and service rate $\mu = 1$. (This would apply to a queueing system with n servers after multiplying $\lambda(t)$ and $s(t)$ by n; we would then scale up the performance as well.) We consider four different initial conditions: (i) empty, (ii) underloaded with $B(0) = 0.5$ and $Q(0) = 0$, (iii) overloaded with $B(0) = 1$ and $Q(0) = 0.4$ and (iv) overloaded with $B(0) = 1$ and $Q(0) = 0.8$. (Simulations confirm that these plots describe the queue performance well for large n.)

Figure 2 shows that the differences of fluid contents in these four cases converge to zero extremely fast. Thus Figure 2 shows the asymptotic loss of memory (ALOM) property; i.e., the performance is asymptotically independent of the initial conditions as time evolves. Indeed, in [10] we show that the differences dissipate exponentially fast, under regularity conditions.

In [10], we apply ALOM to establish convergence to steady state in the $G/M/s + GI$ fluid queue, characterized by [14], and convergence to a PSS in the $G_t/M/s + GI$ fluid queue when the arrival rate function is periodic. In [11], we also treat the open network generalization of the $G_t/M_t/s_t{+}GI_t$ fluid queue where the service and patience distributions at each queue are allowed to be time-varying.

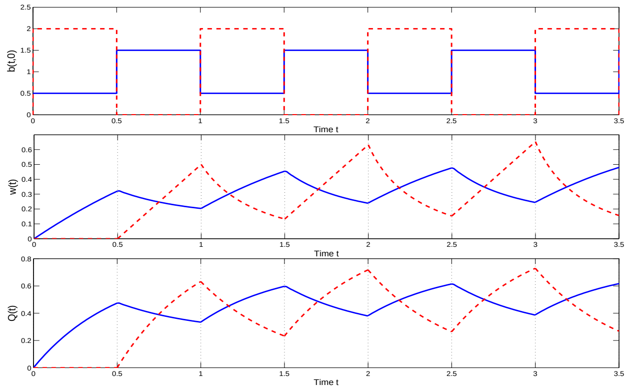


Figure 3: The PSS performance of the $GI/D/s + M$ fluid queue with different initial conditions.

4. STATIONARY MODEL WITH D SERVICE

In §§2 and 3 our analysis is based on the assumption that the service time distribution G has a density g. Now we consider the $GI/D/s + GI$ fluid queue that has constant parameters and deterministic service times $1/\mu$. Deterministic service times are of applied interest, because computer-generated service times, such as automated messages, may well be deterministic. On the one hand, we know the queue-length process $X(t)$ of a $M/D/n + M$ stochastic queue has a well-defined steady state as $t \to \infty$ for all arrival rates, because abandonment ensures stability. We demonstrate by showing that $\{X(t) : t \geq 0\}$ is a regenerative process with finite regeneration cycle τ, under regularity conditions. However, the times between regenerations grow exponentially fast as $n \to \infty$ where n is the number of servers. This suggests that in a $M/D/n + M$ queue with large n and λ, the queue length converges to that steady state very slowly. (That is confirmed by our analysis.)

On the other hand, as $n \to \infty$, the sequence of appropriately scaled (under the fluid scaling) stochastic queues converges to the $M/D/s + M$ fluid model. We consider the case $\rho = \lambda/s\mu > 1$, under which the system is overloaded, but still stable because of the abandonment. We show the performance in the fluid model converges to a PSS as time evolves. Moreover, there are infinitely many possible PSS's, depending on the initial conditions. The system is stationary if and only if the system starts with the unique stationary fluid content. We establish the many-server heavy-traffic limit in [9], under regularity conditions.

In Figure 3, we illustrate this periodic behavior by plotting the performance of the $M/D/s+M$ fluid model with $\lambda = 2$, $\mu = s = 1$ and exponential abandonment with rate $\theta = 2$. We consider two initial conditions: (i) critically loaded with $b(0, x) = 1.5 \cdot 1_{\{0 \leq x \leq 1/2\}} + 0.5 \cdot 1_{\{1/2 < x \leq 1\}}$, $Q(0) = 0$ and (ii) empty. As shown in Figure 3, both cases yield a PSS with period $1/\mu = 1$, but the performance in these two cases differs dramatically. Simulations show that the associated large-scale queueing systems behaves the same way for long intervals of time.

5. ACKNOWLEDGMENTS

This research was supported by NSF grant CMMI 0948190.

6. REFERENCES

[1] Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16:665–688, 2007.

[2] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54:324–338, 2008.

[3] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: tutorial, reviews and research prospects. *Manufacturing Service Operations Management*, 5:79–141, 2003.

[4] L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16:13–39, 2007.

[5] I. Gurvich and W. Whitt. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Operations Management*, 11:237–253, 2009.

[6] O. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt. Server staffing to meet time-vary demand. *Management Science*, 42:1383–1394, 1996.

[7] Y. Liu and W. Whitt. A fluid model for a large-scale service system experienceing periods of overloading. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[8] Y. Liu and W. Whitt. A fluid model for a many-server $G_t/GI/s_t + GI$ queue. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[9] Y. Liu and W. Whitt. The heavily loaded many-server queue with abandonment and deterministic service times. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[10] Y. Liu and W. Whitt. Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with customer abandonment. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[11] Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[12] Y. Liu and W. Whitt. Stabilizing customer abandonments in many-server queues with time-varying arrivals. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[13] G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probability Surveys*, 4:193–267, 2007.

[14] W. Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54:37–54, 2006.

[15] G. Yom-Tov and A. Mandelbaum. Time-varying QED queues with reentrant customers in support of healthcare staffing. *The Technion, Haifa, Israel*, 2010.