Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters*

Tania Jiménez‡ Ger Koole§†

‡ CESIMO, Universidad de los Andes, Mérida, Venezuela

§ Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands

OR Spectrum **26**: 413–422 (Special issue on Call Center Management), 2004

Abstract

Temporary overload situations in queues can be approximated by fluid queues. We strengthen earlier results on the comparison of multi-server tandem systems with their fluid limits. At the same time we give conditions under which economies of scale hold. We apply the results to call centers.

Keywords: call centers, fluid limits, economies of scale, inhomogeneous Poisson processes.

1 Introduction

Queueing systems with inhomogeneous Poisson arrivals are extremely hard to analyze exactly. However, in practice, constant rate Poisson input is the exception. For example, in call centers traffic is usually modeled as arriving according to an inhomogeneous Poisson process with a piecewise constant arrival rate. As long as the system is stable under all possible arrival rates, then a simple approximation based on stationarity often suffices, see Green & Kolesar [4]. The idea of this *pointwise stationary approximation* (PSA) is that the performance at each moment is approximated by the stationary performance of the system with the parameters as they are at that point in time. This method of computing the performance is indeed common practice in call centers, where usually the Erlang delay formula is chosen to approximate stationary performance. See Gans et al. [2] for an overview of the modeling of call centers.

The situation changes when an overload situation occurs during some of the time. Stationarity cannot be used during the overload period, as this would result in an infinite queue

†Communicating author. Full address: Department of Mathematics, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands. Email koole@few.vu.nl, fax +31 20 4447653

length, while in reality the queue length remains finite, it is just building up. And worse, stationarity fails also in subsequent underload intervals, because of the large quantities of jobs that are transferred to later intervals.

There are several potential ways to deal with temporary overload situations. In this paper we focus on the use of fluid limits, but let us first discuss some other possibilities.

In Jennings et al. [6] a non-homogeneous multi-server queue is approximated by an infinite-server 'queue'. This is relevant to time-inhomogeneous systems because servers still being busy due to a high load earlier in time are modeled. On the other hand, queueing phenomena including a queue building up in overload are not modeled. For this reason this method is expected to work better than the PSA in situations when the rate changes quickly, because then queueing hardly ever occurs. This intuition is confirmed by the results in Jennings et al. [6] and Massey & Whitt [10]. In this paper we are interested in situations where delayed customers due to longer overload periods play a crucial role. Therefore a method based on an infinite-server queue is less suitable.

An exact, but numerically demanding method is *uniformization* (see, e.g., Section 6.7 of Ross [13]). In Ingolfsson et al. [5] this method is successfully used in the context of call centers. (This method is only exact for piecewise homogeneous processes.) In this paper we use simulation to verify the results based on fluid models.

The idea of using fluid limits to model overload situations dates back to Newell [11]. He sees the fluid limit as the average over many samples of the input. This results, for exponential service times, in multiplying the arrival rate and the service rate with the same number k. The limit as $k \to \infty$ is the fluid limit. In Altman, Jiménez, & Koole [1] it was shown for an important class of systems and objectives that this fluid limit has a lower value than the original system. A second way to define a fluid limit is by scaling the arrival rate and the number of servers, but keeping the service rate constant. This converges to a different fluid limit (see Mandelbaum, Massey & Reiman [8]), that has in general a higher value than the fluid limit first introduced. We start with showing that this second fluid limit is again a lower bound on the performance of the real system, by considering first, using coupling arguments, the effects of scaling. This gives results on scaling that are interesting on their own.

Next we move to call centers and approximations using fluid limits. In Mandelbaum, Massey, Reiman & Rider [9] it is shown that the second fluid limit approximates the real performance exceptionally well in the situation of call centers. However, there numerical examples deal almost exclusively with overload situations. Indeed, the examples in [1] show that the first fluid limit gives a poor approximation in situations with a low load. This is not different for the second fluid limit. This leads us to the following simple performance approximation, to be used in all situations: the performance at some point in time t is approximated by the maximum of the performance of the fluid limit at t and of the performance of the stationary $M/M/s$ queue with as input rate the minimal rate up to t. For obvious reasons this is again a lower bound of the performance. Intuitively this bound performs well in clear underload and overload situations. The approximation is worst in cases where the system is very close to the critical load. This is also numerically shown.

We illustrate these ideas in the context of call centers, where we base our analysis on

data from an existing call center.

2 Economies of scale

Fluid limits for queueing systems can be taken in several ways. In this paper we look at a method that consists of increasing the number of servers and the arrival rate with the same factor. This means that the number of customers in the system increases as well, thus the performance measure (e.g., the number of customers) has to be scaled as well. This is actually equivalent to increasing the *scale* of the system.

In this section we answer the question of when increasing scale leads to improved results, i.e., when economies of scale hold. That the answer to the last question is non-trivial is shown by the following numerical example.

Consider first the stationary $M(\lambda)/M(\mu)/s$ system, and let $W_q(\lambda, s)$ be its waiting time distribution. Then $W_q(\lambda, s)$ is exponentially distributed with parameter $s\mu - \lambda$ with probability $C(s, \lambda/\mu)$ (the so-called delay probability), and $W_q(\lambda, s) = 0$ with probability $1 - C(s, \lambda/\mu)$. It is well-known that $C(ks, k\lambda/\mu) \leq C(s, \lambda/\mu)$ for $k \in \mathbb{N}$. From this it follows that $\mathbb{P}(W_q(\lambda, s) > t) \geq \mathbb{P}(W_q(k\lambda, ks) > t)$. This result does, surprisingly enough, not hold for the transient system. To see this, consider an initially empty $M(\lambda_t)/M/1$ queue where $\lambda_t \gg \mu$ for small t: we only consider arrivals until u with u such that $\int_0^u \lambda_t dt = 2$. Denote with W_{qt} the waiting time of a customer that arrives at t. Note that $W_{qu}(\lambda_t, 1) > 0$ if there is one customer in the system at u. Because $\lambda_t \gg \mu$ we find that $\mathbb{P}(W_{qu}(\lambda_t, 1) > 0) \approx \mathbb{P}(1$ arrival in Poisson process with parameter $2)$. Similarly, $W_{qu}(2\lambda_t, 2) > 0$ if there are two customers in the system at u. Thus $\mathbb{P}(W_{qu}(\lambda_t, 1) > 0) \approx 1 - e^{-2} \approx 0.86$, but $\mathbb{P}(W_{qu}(2\lambda_t, 2) > 0) \approx 1 - e^{-4} - 4e^{-4} \approx 0.91$. Thus the probability of being served immediately is bigger in the single-server system, therefore increasing scale does not give advantages.

The question we will answer next is for which types of performance measures an increase of scale leads to performance improvements. Let Q_t (Q_t^k) be the queue length process of an $M(\lambda_t)/M/s$ ($M(k\lambda_t)/M/ks$) queue. We are interested in comparing (functions of) Q_t and Q_t^k for all or some $t > 0$. To be able to do so we first couple both systems, leading to the following theorem.

Theorem 2.1 *Let $Q_t^{(1)}, \ldots, Q_t^{(k)}$ represent k independent copies of Q_t. Then $Q_t^{(1)}, \ldots, Q_t^{(k)}$ and Q_t^k can be coupled such that for all realizations $Q_t^{(1)}(\omega), \ldots, Q_t^{(k)}(\omega)$ and $Q_t^k(\omega)$ we have $Q_t^{(1)}(\omega) + \cdots + Q_t^{(k)}(\omega) \geq Q_t^k(\omega)$, assuming that it holds initially.*

Proof We will consider the k parallel $M(\lambda_t)/M/s$ queues as a single system. Both systems can be coupled as follows: when an arrival occurs at the $M(k\lambda_t)/M/ks$ queue there is also an arrival at one of the parallel queues, each with probability $1/k$. The departures are coupled in the following way: there are ks independent Poisson processes, each governing the potential departures of a server in each system. The systems are coupled in such a way that an active server in the $M(k\lambda_t)/M/ks$ queue corresponds to an

active server in one of the $M(\lambda_t)/M/s$ queues. With this coupling it is easily seen that $Q_t^{(1)}(\omega) + \cdots + Q_t^{(k)}(\omega) \geq Q_t^k(\omega)$ is preserved. □

Remark Up to now we have assumed that only the arrival rate changes over time. It is also possible to make the service rate and the number of servers change, without any additional difficulties.

Next we introduce performance measures C and C^k, which are functions of the state of the original and the scaled system. In this paper we will use increasing and convex in the non-strict sense. With $\leq_s$ we indicate the usual stochastic order: $X \leq_s Y$ means that X and Y can be coupled such that $X(\omega) \leq Y(\omega)$ for all ω.

Theorem 2.2 *Let C^k be increasing and convex and $C^k(kx) \leq C(x)$. Then $\mathbb{E}C^k(Q_t^k) \leq \mathbb{E}C(Q_t)$ for all t, assuming that initially $Q_0^k \leq_s kQ_0$.*

Proof According to Theorem 2.1 and the fact that the current initial condition is stronger than in Theorem 2.1, Q_t^k and Q_t can be coupled such that for $Q_t^{(i)}(\omega) = x_i$ and $Q_t^k(\omega) = x$, $x_1 + \cdots + x_k \geq x$. By $C^k(kx) \leq C(x)$ we have

$$C(x_1) + \cdots + C(x_k) \geq C^k(kx_1) + \cdots + C^k(kx_k).$$

Next, by convexity of C^k we find

$$\frac{C^k(kx_1) + \cdots + C^k(kx_k)}{k} \geq C^k((kx_1 + \cdots + kx_k)/k) = C^k(x_1 + \cdots + x_k).$$

Finally, as C^k is increasing, we have

$$C^k(x_1 + \cdots + x_k) \geq C^k(x).$$

Combining all and taking expectations gives $\mathbb{E}C(Q_t) \geq \mathbb{E}C^k(Q_t^k)$. □

Let us interpret Theorem 2.2, by considering some relevant cost functions and see if they satisfy the conditions. Consider the expected number of calls in the system. If we increase the scale by a factor k we expect, due to economies of scale, that the number of calls increases by less than k, i.e., we expect $\mathbb{E}Q_t^k \leq \mathbb{E}kQ_t$. This follows directly from Theorem 2.2 by taking $C(x) = x$ and $C^k(x) = x/k$, and for which choice $C^k(kx) = C(x)$. We see that, in correspondence with the fluid limit, we have to scale the queue lengths as well. Other cost functions can be of interest, such as the expected number of customers in the queue $C(x) = \max\{x-s, 0\} = (x-s)^+$ and its scaled counterpart $C^k(x) = (x-ks)^+/k$. The function C^k is convex and again $C^k(kx) = (x-s)^+ = C(x)$.

Next we consider waiting times. The expected waiting time $\mathbb{E}W_{qt}$ is given by $\mathbb{E}W_{qt} = (Q_t + 1 - s)^+/(s\mu)$, for the scaled system by $\mathbb{E}W_{qt}^k = (Q_t^k + 1 - ks)^+/(ks\mu)$. Thus $C(x) = (x + 1 - s)^+/(s\mu)$ and $C^k(x) = (x + 1 - ks)^+/(ks\mu)$. It is easily seen that $C^k(kx) \leq C(x)$; note that $C^k(kx) < C(x)$ if $x > s$.

Through a counterexample we saw earlier in this section that $\mathbb{P}(W_{qt}^k > v) \leq \mathbb{P}(W_{qt} > v)$ does not hold in general. This is because of the fact that $\mathbb{P}(W_{qt}^k > v)$ is not convex in the queue length. We work this out for $v = 0$, then $\mathbb{P}(W_{qt}^k > v) = \mathbb{I}\{Q_t^k > s\}$, with $\mathbb{I}$ the indicator function: $\mathbb{I}\{A\} = 1$ if A holds, 0 otherwise. This function is increasing, but not convex. A performance indicator that is convex and relevant to call centers is $\mathbb{E}(W_{qt} - v)^+$, the expected time that waiting exceeds v (see Koole [7]). This can be seen as follows. Let $E_1, E_2, \ldots$ be independent exponential random variables (r.v.s) with parameter $s\mu$, and define the gamma distributions $G_n = E_1 + \cdots + E_n$. Define also $G_0 = 0$. Then

$$\mathbb{E}(W_{qt} - v)^+ = \mathbb{E}\sum_{n=0}^{Q_t - s} \mathbb{P}(G_n \leq v < G_{n+1}) \frac{Q_t + 1 - s - n}{s\mu} = \mathbb{E}\sum_{n=0}^{Q_t - s} \mathbb{P}(v < G_{n+1}) \frac{1}{s\mu}.$$

From this we deduce that $C(x) = \sum_{n=0}^{x-s} \mathbb{P}(v < G_{n+1}) \frac{1}{s\mu}$. Thus $C(x) = 0$ as long as $x < s$, if $x \geq s$ then $C(x) - C(x-1) = \mathbb{P}(v < G_{x-s+1}) \frac{1}{s\mu}$. This number is increasing in x, and therefore C is convex.

Now consider the scaled system, with performance $\mathbb{E}(W_{qt}^k - v)^+$. Similarly, let $E_1^k, E_2^k, \ldots$ be independent exponential r.v.s with parameter $ks\mu$, and take $G_n^k = E_1^k + \cdots E_n^k$, $G_0^k = 0$. Then $C^k(x) = \sum_{n=0}^{x-ks} \mathbb{P}(v < G_{n+1}^k) \frac{1}{ks\mu}$. This is again convex, the question that remains to be answered is whether $C^k(kx) \leq C(x)$, which is equivalent to:

$$\mathbb{E}(G_{kn+1}^k - v)^+ \leq \mathbb{E}(G_{n+1} - v)^+,$$

which follows from the fact that G_{kn}^k is stochastically less variable than G_n (see Ross [12], Section 8.5).

3 Fluid limits and performance bounds

In general, there are different ways to take fluid limits. In our context it is crucial that the time is not scaled, because we are interested in modeling a non-stationary arrival process. Then there are two ways of taking fluid limits. Let us briefly discuss the classical one before we continue with the one we focus on in this paper.

The classical method (already proposed by Newell [11] in 1971) consists of averaging the arrival process over multiple realizations. This is, for the $M(\lambda_t)/M(\mu)/s$ queue, equivalent to scaling λ_t and μ. In the limit the amount of work that arrives during a period $[t_0, t_1]$ is equal to $\int_{t_0}^{t-1} \lambda_t dt/\mu$ a.s. It disappears at a fixed rate $s\mu$, as long as there is fluid. The limit is thus a continuous non-negative function $\tilde{z}_t$, with $\tilde{z}_0$ given, and derivative $\tilde{z}_t' = \lambda_t - s\mu$ if $\tilde{z}_t > 0$, and $\tilde{z}_t' = (\lambda_t - s\mu)^+$ if $\tilde{z}_t = 0$.

This type of fluid limit can also be obtained for general service times. The method is less suitable for multi-server queues, as it gives the same limit as for single-server queues (with the server speed appropriately scaled). This way of scaling is therefore most appropriate for the $M_t/G/1$ queue. See Altman et al. [1] for results concerning comparison with fluid limits that parallel the results obtained here with for the fluid limit discussed next.

The second way of taking fluid limits is by increasing s and λ with the same factor. This is actually equivalent to increasing the *scale* of the system. It also means that the number of customers in the system increases as well, and therefore the performance measure (e.g., the number of customers) has to be scaled. The resulting fluid limit $z_t = \lim_{k\to\infty} Q_t^k/k$ is defined by its derivative $z_t' = \lambda_t - \min(z_t, s)\mu$. As a result single and multi-server queues give different limits.

We are interested in $\lim_{k\to\infty} \mathbb{E}C^k(Q_t^k)$. We have the following result, showing that for well-chosen C^k and C this fluid limit gives again a lower bound to the performance.

Theorem 3.1 *For C^k increasing and convex for each $k \in \mathbb{N}$ and $C^k(kx) \le C(x)$ for all k, x then $\tilde{C}(x) = \limsup_{k\to\infty} C^k(kx)$ is such that $\tilde{C}(z_t) \le C(Q_t)$.*

Proof The function $\tilde{C}$ is well-defined because $\tilde{C}(x) \le \sup_k C^k(kx) \le C(x)$. By dominated convergence, using Theorem 2.2, we find $\tilde{C}(z_t) = \limsup_{k\to\infty} \mathbb{E}C^k(Q_t^k)$. Combining this with $\mathbb{E}C^k(Q_t^k) \le \mathbb{E}C(Q_t)$ gives the desired result. □

A special case of Theorem 3.1 is when $C^k(kx) = C(x)$. Then $\tilde{C} = C$. This was the case when considering the number of calls in the system. This shows that the fluid limit is a lower bound for the number of calls in the system. In the case of expected waiting times $\lim_{k\to\infty} C^k(kx) = (x-s)^+/(s\mu) = \tilde{C}(x)$. In the case of the expected time that waiting exceeds v we find $\lim_{k\to\infty} C^k(kx) = (x-v)^+$.

Theorem 3.1, together with the following lemma, motivates us to concentrate on the fluid limit z_t.

Lemma 3.2 *If $z_0 = \tilde{z}_0$, then $z_t \ge \tilde{z}_t$ for all $t \ge 0$.*

Proof If $z_t = \tilde{z}_t$, then $z_t' = \lambda_t - \min(z_t, s)\mu \ge \lambda_t - \mathbb{I}\{z_t\}s\mu = \tilde{z}_t'$. From this the result follows directly. □

In Mandelbaum et al. [9] z is used for a call center, but mainly in overload. For a stable system with a constant arrival rate (i.e., $\lambda < s\mu$) the fluid limit tends to λ/μ: if $z_t > \lambda/\mu$, then $z_t' < 0$, and vice versa. This explains why the fluid approximation is so bad for these systems: for the $M/M/1$ queue for example, the average queue length is given by $\sum_{n=1}^{\infty} n(1-\lambda/\mu)(\lambda/\mu)^n = \lambda/(\lambda-\mu)$.

In both fluid limits waiting does not occur due to random fluctuations. In the first type of limit that is because the randomness is averaged out of the arrival process, in the second type this is due to the scaling of the system. Waiting can still occur because of fluctuations in the arrival rate. It is exactly for these reasons that fluid limits are useful for transient analysis, especially because there are no exact results for models with a non-homogeneous arrival process.

As an example, take the stationary $M(\lambda)/M/1$ queue. The average number of customers in the system is given by $\lambda/(\mu-\lambda)$. The limit for the first fluid limit is 0, and for the second it is λ/μ: this is the fraction of servers that is busy. Customers spend

respectively on average $1/(\mu-\lambda)$, 0, and $1/\mu$ time units in the system. Note that for the first fluid limit customers become infinitesimally small; this explains the sojourn time of 0.

We consider a queueing system in $[0,T]$. We assume that $[0,T]$ is divided in n intervals $[0,t_1],[t_1,t_2],\ldots,[t_{n-1},T]$ (with $0=t_0<t_1<\cdots<t_{n-1}<t_n=T$). During interval i (from t_{i-1} to t_i) customers arrive according to a Poisson process with rate λ_i. As before we keep μ and s constant. We consider thus an $M/M/s$ system with inhomogeneous piecewise constant arrival rate. Furthermore, we assume that the system is in a stationary situation for some arrival rate λ_0 (with $\lambda_0<s\mu$) at time 0. This is the usual way to model call centers, see Gans et al. [2], although it is advisable in many situations to model abandonments as well. Also the choice to keep s constant is not very realistic, although we will show an application in which it is the case. Making s time-dependent is possible, but to keep the exposition simple we choose not to do it. Both abandonments and time-varying numbers of servers are subject of ongoing research.

Now consider some performance measure C that is increasing. Define

$$\lambda_t^* = \min\{\lambda_0, \min_{i|t_{i-1}<t} \lambda_i\},$$

the minimal rate up to t. Denote with $\hat{Q}_\lambda$ the stationary distribution of the $M(\lambda)/M(\mu)/s$ queue.

Lemma 3.3 $\mathbb{E}C(Q_t) \geq \mathbb{E}C(\hat{Q}_{\lambda_t^*})$ *for all* $t \geq 0$.

Proof Using a coupling argument it is easily seen that $Q_t \geq_s \hat{Q}_{\lambda_t^*}$ for all t. The result follows because C is increasing. □

Now assume we also have some $\tilde{C}(x) = \limsup_k C^k(kx)$ with all C^k convex and increasing. We propose the following approximation for the performance at t, denoted by L_t:

$$L_t = \max\{\tilde{C}(z_t), \mathbb{E}C(\hat{Q}_{\lambda_t^*})\},$$

which is the maximum of the performance of the fluid limit at t and of the performance of the stationary $M/M/s$ queue with as input rate the minimal rate up to t. Thanks to Lemma 3.3 and Theorem 3.1 we know that this is a lower bound to the system performance, i.e., $\mathbb{E}C(Q_t) \geq L_t$.

In the next section we show how this approximation works in the context of a call center.

4 A call center application

In this section we study a simple time-varying model for a call center. After introducing the model we apply the results of the previous section to obtain a lower bound on the performance.

We simulated a call center with 32 servers having exponential service times with mean 240 seconds using the Glider simulation language [3]. We simulated two situations, both with first overload and then underload. In the first situation we used a Poisson arrival process with an average of 270 calls the first 30 minutes and an average of 225 calls per 30 minutes for the rest of the simulation. Thus there is an overload situation during the first 30 minutes ($\rho = 1.125$), and after that the system is stable ($\rho = 0.9375$). In the second situation we had an average of 331 calls during the first 30 minutes, then 223 calls in the following 30 minutes and 162 calls per 30 minutes during the rest of the simulation. Thus there is again an overload situation during the first 30 minutes ($\rho = 1.378$), and after that the system is stable. The numbers of the second situation are inspired by an actual call center where a peak in traffic occured as soon as it opened. Management decided not to schedule more agents, but to base the number of agents on the stable situation after 30 or 60 minutes.

We performed 100 simulations of the systems. The simulated queue lengths are shown in Figure 1, together with the fluid limits and the performance under the minimal arrival rate. (For practical reasons we started the simulation empty, in contrast with the lower bound of the previous section.) To obtain the fluid limit we had to solve $z_t' = \lambda_i - \min(z_t, s)\mu$. Its solution is as follows: $z_t = z_{u_0} + (\lambda_i - s\mu)(t - u_0)$ on $t \in [u_0, u_1]$ as long as $z_t \geq s$ and $[u_0, u_1] \subset [t_{i-1}, t_i]$, i.e., λ_t is constant and equal to λ_i on $[u_0, u_1]$; $z_t = \lambda_i/\mu + (z_{u_0} - \lambda_i/\mu)\exp(-\mu(t - u_0))$ on $t \in [u_0, u_1]$ as long as $z_t \leq s$ and $[u_0, u_1] \subset [t_{i-1}, t_i]$.

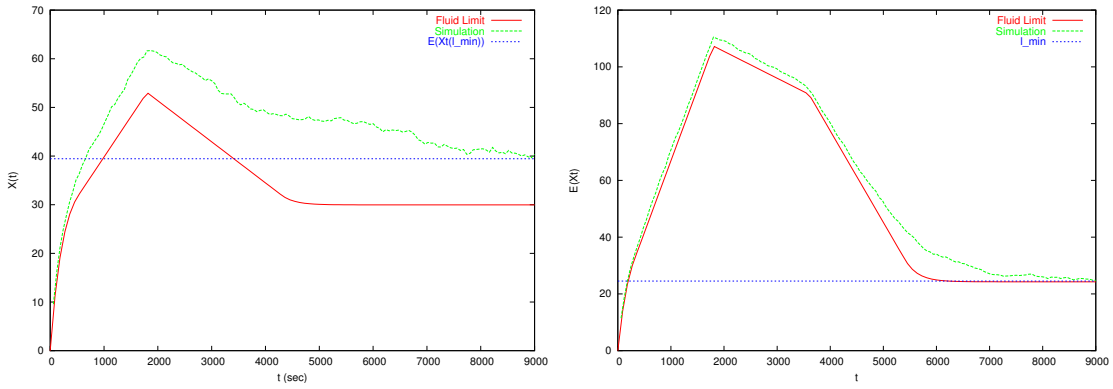

Figure 1: Number of calls in the systems for both situations

It is striking to see in Figure 1 how the queue builds up during the first 30 minutes, and how it decreases slowly thereafter, followed nicely by the fluid limit, especially in situation 2. The fluid limit is less useful later on (unless the load is very low, as it is in the second situation), but here the stationary approximation gives reasonable results (for the second situation they almost coincide).

Similar observations apply to Figure 2, in which $\mathbb{E}(W_t - 60)^+$ is plotted together with $(z_t - 60)^+$ and $\mathbb{E}(Y - 60)^+$, with Y the stationary performance under the minimal arrival rate.

It is clear from the plots that more experimental work has to be done: the two situations

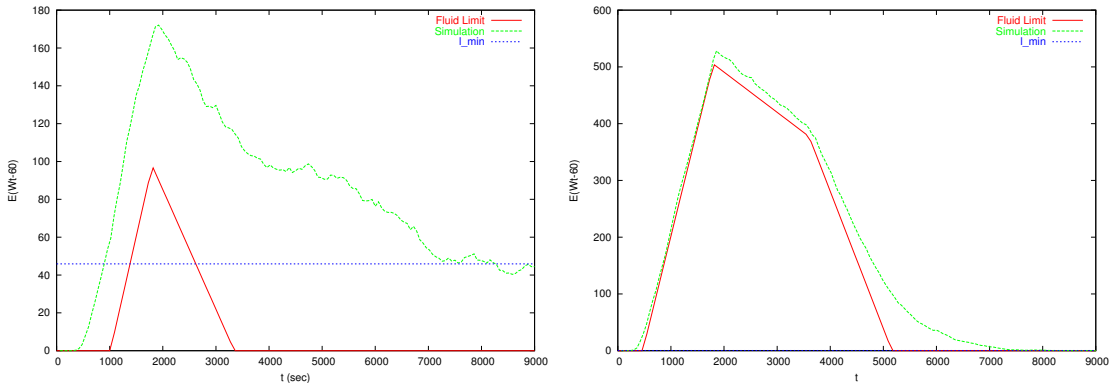

Figure 2: Average of waiting times that exceed one minute for both situations

that we analyzed give different results. We conclude that the new lower bound, based on the fluid limit and the stationary regime for the minimum of the arrival rates, is considerably better than using the fluid limit alone. It is also remarkable that the fluid limit follows the simulation better in the second situation. The reason might be that the slope in the second situation is steeper. This corresponds to our intuition that the approximation is worst close to the critical load. Further research is needed to investigate the practical usefulness of the lower bound presented in this paper. An interesting extension would be to include abandonments or even retrials in the process.

Acknowlegdment We are grateful to the anonymous referees for their useful remarks that helped us improve the paper.

References

[1] E. Altman, T. Jiménez, and G.M. Koole. On the comparison of queueing systems with their fluid limits. *Probability in the Engineering and Informational Sciences*, 15:165–178, 2001.

[2] N. Gans, G.M. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5:79–141, 2003.

[3] The Glider simulation language. http://afrodita.faces.ula.ve/Glider.

[4] L. Green and P. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 1991.

[5] A. Ingolfsson, E. Cabral, and X. Wu. Combining integer programming and the randomization method to schedule employees. Submitted, 2003.

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

[7] G.M. Koole. Redefining the service level in call centers. Working paper, 2003.

[8] A. Mandelbaum, W.A. Massey, and M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998.

[9] A. Mandelbaum, W.A. Massey, M.I. Reiman, and R. Rider. Time varying multiserver queues with abandonments and retrials. In P. Key and D. Smith, editors, *Proceedings of the 16th International Teletraffic Conference*, 1999.

[10] W.A. Massey and W. Whitt. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems*, 25:157–172, 1997.

[11] G.F. Newell. *Applications of Queueing Theory*. Chapman and Hall, 1971.

[12] S.M. Ross. *Stochastic Processes*. Wiley, 1983.

[13] S.M. Ross. *Introduction to Probability Models*. Academic Press, 7th edition, 1997.