THE HEAVY-TRAFFIC BOTTLENECK PHENOMENON IN OPEN QUEUEING NETWORKS

by

S. Suresh and *W. Whitt*

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

This note describes a simulation experiment involving nine exponential queues in series with a non-Poisson arrival process, which demonstrates that the heavy-traffic bottleneck phenomenon can occur in practice (at reasonable traffic intensities) as well as in theory (in the limit). The results reveal limitations in customary two-moment approximations for open queueing networks.

Key Words: queues, open queueing networks, queues in series, limit theorems, heavy traffic, two-moment approximations

February 7, 1989

Revision: January 4, 1990

1. Introduction

The purpose of this note is to describe a simulation experiment that provides insight into the steady-state performance of non-product-form open queueing networks. In particular, we show that the heavy-traffic bottleneck phenomenon in an open queueing network can occur approximately at reasonable traffic intensities.

By the heavy-traffic bottleneck phenomenon, we mean the state-space collapse that occurs if the traffic intensity of one queue approaches 1, while the traffic intensities at all other queues remain below $1-\varepsilon$ for some $\varepsilon > 0$. Heavy-traffic limit theorems by Iglehart and Whitt [5], Reiman [7, 8] and Chen and Mandelbaum [4] indicate that if the traffic intensity at one queue is sufficiently high, while the traffic intensities of all the other queues are substantially lower, then the standard steady-state random variables such as the waiting time at each queue and the number of customers in the network are distributed nearly the same (relatively to the level of congestion at the bottleneck queue) as if all the service times in the non-bottleneck queues were set equal to 0.

Since the number of customers in the bottleneck queue should go to infinity as its traffic intensity approaches 1, while the number of customers at other queues should stay finite, it is intuitively obvious that the proportion of customers in the network that are at the bottleneck queue should approach 1 in this limit. However, it is less obvious that the normalized steady-state waiting time at the bottleneck queue should be nearly the same as if the service times at all the other queues were set equal to 0, i.e., as if the other queues acted as instantaneous switches. This is the feature that we wish to identify in typical networks.

To exhibit the heavy-traffic bottleneck phenomenon in this form, we choose a relatively simple network. (It will be evident that the phenomenon will hold more generally.) In particular, we consider several single-server queue in series. Customers arrive at the first queue according to

a renewal process with interarrival times having a general distribution with mean 1 and squared coefficient of variation (variance divided by the square of the mean) c_{a1}^2. Each queue has unlimited waiting space, the first-in first-out discipline, and IID (independent and identically distributed) service times that are independent of the arrival process and the other service times. The service-time distribution at queue i has a general distribution with mean ρ_i, where $\rho_i < 1$, and squared coefficient of variation c_{si}^2. In this context, the heavy-traffic bottleneck phenomenon occurs if the traffic intensity of one queue is allowed to approach 1; then, by [5], the waiting-time distribution at this bottleneck queue is asymptotically the same as if the immediate arrival process (i.e., the departure process from the previous queue) were replaced by the external arrival process to the first queue with squared coefficient of variation c_{a1}^2. Our purpose is to show that this can be approximately true at reasonable traffic intensities.

Unfortunately, due to the non-exponential distributions, this model is very difficult to analyze exactly. A useful practical approach to this model and more general open queueing networks is the *parametric-decomposition approximation method*, as in Whitt [14], Segal and Whitt [10], Bitran and Tirupati [3] and references cited there. For our model of queues in series, the standard implementation of this approach is to approximate the arrival process to queue i by a renewal process with arrival rate 1 and squared coefficient of variation c_{ai}^2, where c_{ai}^2 is defined recursively by

$$c_{a,i+1}^2 = \rho_i^2 \, c_{si}^2 + (1 - \rho_i^2) \, c_{ai}^2 , \quad i \geq 1 , \tag{1}$$

see (38) of [14] and (23) of [15]. We then can approximate the mean steady-state waiting time (before beginning service) at queue i by

$$E[W_i] \approx \frac{\rho_i^2 (c_{ai}^2 + c_{si}^2)}{2(1-\rho_i)} \tag{2}$$

or some refinement such as provided by Kraemer and Langenbach-Belz [6]; see (2) and (44) of

[14].

As indicated in [15], approximation (1) can be viewed as the result of the pure *stationary-interval method*, i.e., an attempt to match c_{ai}^2 for $i > 1$ to the actual squared coefficient of variation of a stationary interval in the i^{th} arrival process (but ignoring the dependence among successive interarrival times). It is significant that (2) does *not* reflect the heavy-traffic phenomenon, because the approximating arrival variability parameter c_{ai}^2 at queue i is totally independent of ρ_i.

It may seem appropriate that c_{ai}^2 not depend on ρ_i, because the arrival process to queue i is exogenous to queue i. However, experience has shown that it may be desirable to let c_{ai}^2 depend on ρ_i, because the way the variability in the arrival process affects the queue depends on the traffic intensity in the queue.

An alternate approach described in [13, 15] is the *asymptotic method*, which attempts to choose a variability parameter c_{ai}^2 to match the central limit theorem behavior of the i^{th} arrival process. For queues in series, this leads to the approximation

$$c_{ai}^2 = c_{a1}^2 \quad \text{for all} \quad i \geq 1 \; . \tag{3}$$

Intuitively, (3) may not look too promising, but it is just what is predicted by the heavy-traffic theory when $\rho_i \to 1$. (This was the original motivation for the asymptotic method.) We thus regard actual system performance consistent with (3) and (2), instead of (1) and (2), when ρ_i is relatively high as strong evidence of the heavy-traffic bottleneck phenomenon.

Based on success approximating queues with superposition arrival processes in Albin [1] and Whitt [13, 14], Whitt [15] sought a *hybrid approximation* for the arrival variability parameters for queues in series, which appropriately combines the stationary-interval method and the asymptotic method. However, in the simulations considered in [15], (3) did not help. Until the present experiment, we have had no clear evidence indicating that (3) is relevant at typical traffic

intensities. However, benefits from modifying (1) in open queueing networks were noted by Albin and Kai [2]. Moreover, a modification of (1) that reflects (3) for two queues in series is presented in Suresh and Whitt [11]. However, the modification in [11] does not help significantly with the examples here.

In Section 2 we describe a specific experiment showing that (3) can be relevant at a bottleneck queue at typical traffic intensities. In Section 3 we consider a modification of that experiment to see the effect of inserting a low-variability (high-variability) queue in front of the first queue when the external arrival process has high (low) variability. Finally, we make a few concluding remarks in Section 4.

2. Nine Exponential Queues in Series

We now specify the model to demonstrate the relevance of (3). The traffic intensities were chosen to reflect the heavy-traffic bottleneck phenomenon, but not to be too extreme. For this purpose, the network was given 9 queues with $\rho_9 = 0.9$ and $\rho_i = 0.6$ for $1 \le i \le 8$. Similarly, the service-time and external interarrival-time distributions were chosen to be relatively standard. In particular, all the service-time distributions are exponential (so that $c_{si}^2 = 1$ for all i). Two cases were considered for the interarrival times: high variability and low variability. (Nothing would be learned from a Poisson arrival process, for which the exact solution is known and consistent with both (1) and (3).) The distribution for high variability is the hyperexponential (H_2) distribution with balanced means, as in (3.7) of [13], with $c_{a1}^2 = 8$. The distribution for low variability is deterministic (D) with $c_{a1}^2 = 0$.

The simulation estimates of the expected waiting times at each queue were obtained from ten replications of 30,000 arrivals, discarding the first 2,000 in each case to allow the system to approach steady state. These run lengths are not long enough to obtain high accuracy at the bottleneck queue (see [18]), but they are adequate to clearly demonstrate the heavy-traffic

bottleneck phenomenon. (Other experiments have subsequently been conducted with millions of arrivals that also support the results here.) The estimated mean steady-state waiting times at the last two queues in both cases are displayed in Table 1, together with estimates of 90% confidence intervals, which are based on the *t*-statistic applied to the ten independent replications. (As usual, since the estimates are not actually normally distributed, the *t*-statistic is an approximation.) Also shown in Table 1 are the values of three approximations.

The idea behind this experiment is that, if we did not have the heavy-traffic phenomenon, we would expect that the arrival process to each successive queue would become more like a Poisson process, so that the last queues would behave like M/M/1 queues. (See Remark 4.2 for further discussion.) Consistent with (1), we might expect that the non-Poisson variability in the external arrival process has been dissipated by the time we reach queue 9. However, from Table 1 it is clear that the observed mean waiting time at the bottleneck queue (queue 9) is much higher (lower) than in the M/M/1 model with the same traffic intensity when $c_{a1}^2 = 8.0$ ($c_{a1}^2 = 0.0$). The standard approximation (1) yields $c_{a9}^2 = 1.20$ and 0.97 in these cases, so that from (1) and (2) we would expect the mean waiting time to be about 10% higher and 2% lower than for the M/M/1 models in these two cases. In fact, the actual estimates are 272% higher and 38% lower, respectively.

In contrast, the pure asymptotic-method approximation combining (2) and (3) is much better at the bottleneck queue, providing very strong evidence of the heavy-traffic bottleneck phenomenon. However, ρ_9 could be even higher, so that we should not expect to see the full heavy-traffic effect. Indeed, the approximation combining (2) and (3) does not perform exceptionally well, yielding about a 20% error in each case.

As should be expected, the asymptotic method performs very poorly at the preceding non-bottleneck queue. Note that queue 8 would have the highest traffic intensity among the first 8

queues, and thus be the bottleneck queue in some sense, if we just increased its traffic intensity by a very small amount, say by 0.01. However, for practical purposes, for a queue to be a bottleneck, it is not enough for it to have the highest traffic intensity; its traffic intensity should be substantially greater than the traffic intensities at the other queues.

These examples show *limitations* in the parametric-decomposition approximations *as currently developed*. We still believe that improved parametric-decomposition approximations can be developed to cover these examples. These results suggest that, just as in [1, 13, 14], it should be appropriate to consider hybrid approximations of the stationary-interval and asymptotic methods. In general, it appears that an appropriate approximating arrival process variability parameter at queue i, say c_{ai}^2, should be a function of c_{a1}^2, c_{s1}^2, ..., $c_{s,i-1}^2$ and ρ_1, ..., ρ_i. We are fairly confident that c_{ai}^2 should satisfy the requirement that

$$\min \{c_{a1}^2, c_{s1}^2, \ldots, c_{s,i-1}^2\} \le c_{ai}^2 \le \max \{c_{a1}^2, c_{s1}^2, \ldots, c_{s,i-1}^2\}, \tag{4}$$

but we have just shown that neither (1) nor (3) is always good. However, we expect (1) to work reasonably well when the bounds in (4) are not too far apart.

Reiman [9] recently has proposed two parametric-decomposition approximations for open queueing networks that are strongly based on the heavy-traffic bottleneck phenomenon. The object is to determine c_{ai}^2 to use with (2). The first method is the *individual bottleneck decomposition* (IBD), which treats each queue as if it were the unique bottleneck queue. It is not difficult to see that IBD in fact coincides with the asymptotic method in [13, 15]: this is justified by the heavy-traffic limit theorems in [4, 5, 7, 8]. Reiman's second method, which seems more promising, is the *sequential bottleneck decomposition* (SBD), which starts by identifying the queue, say queue i, with the highest traffic intensity (assuming no ties) and applying the bottleneck approximation to it to determine c_{ai}^2. In a series network this amounts to using (3) at the queue with highest traffic intensity. The procedure continues by removing the bottleneck

queue from the network and replacing it by an external source (with consistent routing) having its service times as interarrival times. Then the procedure is repeated by identifying the queue with the next highest traffic intensity, and so forth. For a series network, this means that the original procedure is repeated for the queues before the first bottleneck queue, and separately for the queues after the first bottleneck queue, with the bottleneck queue being replaced by an external source with arrival variability parameter c_{si}^2. For the example in Section 2, this means using (3) and (2) at queue 9, i.e., $c_{a9}^2 = 8.0$ for the case in which $c_{a1}^2 = 8.0$. At queue 8 it also means using (3) and (2) i.e., $c_{a8}^2 = 8.0$, if ρ_8 is raised to 0.601. However, it means using $c_{a8}^2 = 1$ if instead ρ_7 is raised to 0.601. This example shows that SBD could benefit from refinement, but Reiman shows that it performs quite well in some cases. We regard SBD as another basic method along with the stationary-interval and asymptotic methods that can serve as a basis for refined hybrid methods.

While we do not intend to investigate specific new approximations for c_{ai}^2 here, we suggest some properties that we think c_{ai}^2 should satisfy. First, c_{ai}^2 could reasonably be a convex combination of c_{a1}^2, c_{s1}^2, ..., $c_{s,i-1}^2$ with weights that are continuous functions of $(\rho_1, ..., \rho_i)$. Moreover, the weight on c_{sj}^2 should be increasing in ρ_j and decreasing in ρ_k for $k \neq j$. Similarly, the weight on c_{a1}^2 should be increasing in ρ_i but decreasing in ρ_j for $j \neq i$. Moreover, any approximation should be consistent with SBD for a single bottleneck queue, i.e., queue j as $\rho_j \to 1$. As $\rho_j \to 1$, the approximation of c_{aj}^2 should approach the asymptotic method value and the approximation of c_{ak}^2 for $k \neq j$ should be consistent with replacing queue j by an external arrival process with arrival variability parameter c_{sj}^2. It is not obvious what should happen when two traffic intensities get large; then we would want consistency with the more complicated two-dimensional diffusion limit resulting from [4, 7].

3. Filtering Through a Queue

If there is high variability in an external arrival process, as in the first case above with $c_{a1}^2 = 8.0$, then we might consider controlling the variability by filtering the arrival process through a low-variability queue, i.e., we could insert a low variability queue in front of the other queues in series to absorb some of the fluctuations. Hence, in this section we consider a modification of the experiment above in which an extra queue with deterministic service times is inserted before the same nine exponential queues.

Before discussing our experiment in more detail, we note that a fairly obvious result holds, namely, that adding a queue can only increase the number of customers in the entire system at each time *t* and the time each customer spends in the system.

Proposition. If a new queue is added to a series of queues, then the number of customers in the system at each time and the time each customer spends in the system are greater than or equal to what they were before.

Proof. Note that the performance measures of interest are the same as if the inserted queue were always there but with zero service times. Then observe that the departure times from the inserted queue and all subsequent queues are nondecreasing in the service times; see Theorem 12 of [12]. Of course, the external arrival times are unchanged as are the arrival times at the queue where the service times are being changed. Finally, note that the time in system is the departure time minus the exogenous arrival time and the number in system at time *t* is the number of arrivals by *t* minus the number of departures by *t*. ■

Of course, this comparison result does not imply that it is never desirable to insert an additional queue, because we might prefer to have customers waiting at the inserted queue than at later queues. (In manufacturing, it is often desirable to delay starts to avoid having excessive partially completed work in process.)

Our new experiment consists of a new first queue with $c_{s1}^2 = 0$. The remaining 9 queues do not change; they get relabeled, so that now $\rho_{10} = 0.9$ and $\rho_i = 0.6$ for $2 \le i \le 9$. As before, $c_{si}^2 = 1$ for $2 \le i \le 10$. We consider three different traffic intensities for the first queue $\rho_1 = 0.4$, 0.6 and 0.9.

The simulation experiment was conducted in the same way as the previous one. The results are given in Table 2, along with the case $\rho_1 = 0.0$, which reduces to the previous case. The estimated halfwidth of the 90% confidence interval is given below each simulation estimate. The four cases based on the four values of ρ_1 were generated from the same random variables, so that comparisons between the cases are relatively reliable (but not independent).

From Table 2, we see that the smoothing effect increases as we increase ρ_1. However, for $\rho_1 = 0.6 = \rho_i$, $2 \le i \le 9$, the smoothing effect helps very little beyond the very next queue. In contrast, the smoothing effect for $\rho_1 = 0.9$ is great, but at the expense of substantial delay at the filter queue. Also given in Table 2 are the approximations using (1) and (2). Again, the approximations do not perform very well, especially in predicting the large delay at queue 10 when $\rho_1 \le 0.6$. Even with $\rho_1 = 0.9$, there remains a long-range variability effect on the final bottleneck queue not anticipated from (1); i.e., the estimated mean steady-state waiting time is 14.0, whereas the approximation based on (1) and (2) is nearly the same as the M/M/1 value of 8.1.

From Table 2, we see that the approximation does not perform well at queue 1 when $\rho_1 = 0.4$ and 0.6 and at queue 2 when $\rho_1 = 0$. In part, this is because $c_{a1}^2 = 8.0$ is relatively high variability, for which good approximations are hard to achieve. However, these cases are also ones for which the Kraemer and Langenbach-Belz refinements help significantly. Since these refinements always decrease the approximate value, they do not move the values at queue 10 in the correct direction.

In Table 3 we also report results for the dual example in which the external arrival process is

deterministic ($c_{s1}^2 = 0$) and the first queue has H_2 service times with $c_{s1}^2 = 8.0$. From Table 3 we see that the approximations based on (1) and (2) perform significantly better in this case. From the case $\rho_1 = 0.9$, we see that high variability in the service times can also cause a much greater waiting time in a subsequent bottleneck queue.

4. Concluding Remarks

4.1 Long-Range Variability Effects

The examples here illustrate how high or low variability in an external arrival process or the service times (the case $\rho_1 = 0.9$ in Table 3) can have only limited impact on immediately following queues, and yet have a dramatic effect on a later queue with a much higher traffic intensity. This phenomenon would have been more apparent if we considered deterministic service times at all queues (the pipelining effect in [16]), but perhaps less convincing. A different long-range variability effect for multi-class queueing networks is described in [17]. Upon reflection, it appears that the two phenomena actually are rather similar. Due to the relevant time scales, it is possible for an arrival process to pass through a subnetwork where it has little effect and reappear later largely unchanged. Here the low variability queues before the bottleneck queue do not significantly reduce the high variability in the larger time scale relevant for the bottleneck queue.

4.2 The Reiman-Simon Conjecture

In a certain sense, these examples also test a long-standing conjecture, communicated by M. I. Reiman and B. Simon among others, that the stationary departure process from *n* IID exponential single-server queues in series fed by an independent stationary arrival process converges to a Poisson process as $n \to \infty$. Of course, we only test the quality of the approximation of the alleged limit for finite *n*. Assuming that the conjecture is true (which we strongly believe), we might expect that the arrival process to the last queue in our example would be sufficiently close

to a Poisson process so that the mean steady-state waiting time is close to what it would be in an M/M/1 queue with $\rho = 0.9$. However, we have seen that this is not the case. Evidently n has to be much larger for the departure process to be close to the Poisson process, at least from the perspective of a following bottleneck queue. Evidently the required n for the mean steady-state queue length at a subsequent bottleneck queue with traffic intensity ρ_n to behave as if the arrival process were Poisson (after passing through $n-1$ queues with $\rho_i = 0.6$) goes to infinity as $\rho_n \to 1$. However, this does not contradict the conjecture, if the conjecture is understood to mean convergence of the finite-dimensional distributions, because as ρ_n increases the relevant time scale for the n^{th} arrival process increases with regard to its impact on the steady-state behavior of queue n. In other words, the conjectured convergence as $n \to \infty$ is evidently not uniform in the length of the time intervals considered.

4.3 Simulation Technique for Many Queues in Series

An effective way to simulate many queues in series if the joint distribution of characteristics at several queues is not required is to simulate the individual queues separately and recursively (the opposite of parallel processing). For any queue, given a sequence of arrival times $\{T_n\}$ and a sequence of service times $\{S_n\}$, we generate sequences of departure times $\{D_n\}$ and waiting times $\{W_n\}$ by

$$D_n = \max\{T_n, D_{n-1}\} + S_n \tag{5}$$

and

$$W_n = \max\{T_n, D_{n-1}\} - T_n \tag{6}$$

for $n \geq 1$. Of course, the departure times serve as the arrival times at the next queues.

If we want to reduce memory, we can work with a file containing only the arrival sequence $\{T_n\}$. We generate S_n as needed by a random number generator and collect cumulative statistics on D_n and W_n as we go along. To eliminate extra storage, we can replace T_n by D_n after we

have calculated D_n and W_n, so that $\{T_n\}$ becomes the arrival process to the next queue when we are finished applying (5) and (6) to the given sequence. In fact, we have used this approach to study variations of the model in Section 2 with up to 100 queues in series. (The heavy-traffic bottleneck phenomenon is still present.)

4.4 Improving System Performance

The heavy-traffic bottleneck phenomenon has important implications for improving performance of queues in series (and more general open queueing networks). If there is a bottleneck queue, then obviously we should try to reduce its traffic intensity. Next we should try to reduce the variability of the bottleneck service times and the external arrival process.

REFERENCES

[1] ALBIN, S. L., ‘‘Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues,’’ *Oper. Res.* 32 (1984), 1133-1162.

[2] ALBIN, S. L. and S. KAI, ‘‘Approximation for the Departure Process of a Queue in a Network,’’ *Naval. Res. Logist. Quart.* 33 (1986), 129-143.

[3] BITRAN, G. R. and D. TIRUPATI, ‘‘Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference,’’ *Management Sci.* 34 (1988), 75-100.

[4] CHEN, H. and A. MANDELBAUM, ‘‘Stochastic discrete flow networks: diffusion approximations and bottlenecks,’’ Graduate School of Business, Stanford University, 1988.

[5] IGLEHART, D. L. and W. WHITT, ‘‘Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks, and Batches,’’ *Adv. Appl. Prob.* 2 (1970), 355-369.

[6] KRAEMER, W. and M. LANGENBACH-BELZ, ‘‘Approximate Formulae for the Delay in the Queueing System GI/G/1,’’ *Eighth Int. Teletraffic Congress,* Melbourne, 1976, 235-1-8.

[7] REIMAN, M. I., ‘‘Some Diffusion Approximations with State-Space Collapse,’’ *Proc. Int. Seminar on Modeling and Perf. Eval. Methodology,* eds. F. Baccelli and G. Fayolle, Springer-Verlag, Berlin, 209-240, 1983.

[8] REIMAN, M. I., ‘‘Open Queueing Networks in Heavy Traffic,’’ *Math. Oper. Res.* 9 (1984), 441-458.

[9] REIMAN, M. I., ‘‘Asymptotically Exact Decomposition Approximations for Open

Queueing Networks,'' *Oper. Res. Letters,* to appear.

[10] SEGAL, M. and W. WHITT, ''A Queueing Network Analyzer for Manufacturing,'' in *Teletraffic Science for New Cost-Effective Systems, ITC12* (ed. M. Bonatti) North-Holland, Amsterdam, 1989, 1146-1152.

[11] SURESH, S. and W. WHITT, ''Arranging queues in series: a simulation experiment,'' *Management Sci.* 36 (1990), to appear.

[12] WHITT, W., ''Comparing counting processes and queues,'' *Adv. Appl. Prob.* 13 (1981), 207-220.

[13] WHITT, W., ''Approximating a Point Process by a Renewal Process, I: Two Basic Methods,'' *Oper. Res.* 39 (1982), 125-147.

[14] WHITT, W., ''The Queueing Network Analyzer,'' *Bell System Tech. J.* 62 (1983), 2779-2815.

[15] WHITT, W., ''Approximations for Departure Processes and Queues in Series,'' *Naval Res. Logist. Quart.* 31 (1984), 499-521.

[16] WHITT, W., ''The best order for queues in series,'' *Management Sci.* 31 (1985), 475-487.

[17] WHITT, W., ''A Light-Traffic Approximation for Single-Class Departure Processes from Multi-Class Queues,'' *Management Sci.* 34 (1988), 1333-1346.

[18] WHITT, W., ''Planning Queueing Simulations,'' *Management Sci.* 35 (1989), 1341-1366.

		high variability $c_{a1}^2 = 8.0$	low variability $c_{a1}^2 = 0.0$
queue 9 $\rho_9 = 0.9$	simulation estimate	30.1 ± 5.1	5.03 ± 0.22
	approximation (1) and (2)	8.9	8.0
	M/M/1 approximation	8.1	8.1
	approximation (3) and (2)	36.5	4.05
queue 8 $\rho_8 = 0.6$	simulation estimate	1.41 ± 0.07	0.775 ± 0.013
	approximation (1) and (2)	1.04	0.88
	M/M/1 approximation	0.90	0.90
	approximation (3) and (2)	4.05	0.45

Table 1. Simulation estimates of the mean steady-state waiting times at queues 9 and 8 in the network of nine queues in series in Section 2, plus associated approximations.