Queueing Systems 10 (1992) 271–278

Short Communication

Counterexamples for comparisons of queues with finite waiting rooms

Ward Whitt

AT&T Bell Laboratories, Murray Hill, NJ 07974-0636, USA

Received 3 December 1990; revised 25 January 1991

We construct $G/G/1/k$ queueing models that fail to have anticipated monotonicity properties with respect to the capacity k. In one model the long-run average number of customers in the system is arbitrarily close to the capacity k, but it decreases to an arbitrarily small value when the capacity is increased. In another model the throughput is arbitrarily close to the arrival rate when the capacity is k, but the throughput decreases to an arbitrarily small value when the capacity is increased. These examples involving non i.i.d. service times, which are associated with external arrivals instead of being assigned when service begins, show that stochastic assumptions and arguments involving more than direct sample-path comparisons are essential for obtaining useful bounds and positive comparison results.

Keywords: Stochastic comparisons; finite waiting rooms; throughput; monotonicity.

1. Introduction

Consider a $G/G/s/k$ queue having s servers, $k-s$ extra waiting spaces, the first-come first-served (FCFS) service discipline, a stationary arrival process and a stationary sequence of service times independent of the arrival process. Let successive service times be associated with successive (external) arrivals. Let arrivals finding the waiting space full be blocked (lost) without affecting future arrivals. In such systems we usually expect the throughput and the long-run average queue length (number in system) to increase with the capacity k. This is illustrated by the explicit formulas available for the $M/M/s/k$ and $M/GI/1/k$ special cases; see Keilson [3] and p. 306 of Tijms [7] for $M/GI/1/k$ and Miyazawa [4] for extensions involving batches. Positive stochastic comparison results have also been established for cases in which explicit solutions are not available. In particular, theorem 1 of Sonderman [5] shows that number of departures in the interval $[0, t]$ is stochastically increasing in k for each t when the service times are i.i.d., so that the throughput (the long-run average

departure rate) is indeed increasing in k under this condition; see Tsoucas and Walrand [8] for extensions to networks. Having i.i.d. service times permits assigning successive service times to customers when they start service, instead of in order of arrival, without altering the distribution of the queueing processes. When service times are assigned in order of service initiation, positive sample-path comparisons can be made, as shown by the proof of theorem 1 of Sonderman [5] (and a forthcoming paper by the author and A.W. Berger). Moreover, theorem 1 of Sonderman [6] and theorems 7 and 10 of Whitt [9] establish stochastic comparisons for the queue lengths that imply the ordering for the long-run average queue length with i.i.d. exponential service times.

However, in general, the throughput and the long-run average queue length are not always increasing with k. The purpose of this paper is to give examples. These examples involving non-i.i.d. service times are interesting, because they show that simple sample-path arguments will not work in this case; i.e., the stochastic assumptions underlying positive results play a crucial role. The examples are also interesting because these performance measures are typically difficult to compute.

The first example in section 2 has the long-run average queue length with capacity k be k, while the long-run average queue length with capacity $m > k$ is arbitrarily small. The idea is relatively simple. Some customers have very long service times and others have very short service times. The k-capacity system is almost always filled with the customers having long service times, whereas the m-capacity system eventually has only customers with short service times, who produce just enough congestion to eventually block all the customers with long service times.

The second example in section 4 has arrival rate 1 and throughputs arbitrarily close to 1 in the k-capacity system but arbitrarily close to 0 in the m-capacity system. The example in section 2 has these throughputs reversed. It is significant that any throughput between 0 and the arrival rate is possible; i.e., without extra conditions, there are no crude bounds which restrict the range of possible throughput values. Moreover, the throughput at one capacity does not restrict the range of possible throughputs at another capacity.

The example in section 2 here corrects example 4.1 in Heyman and Whitt [2], which correctly shows that the queue lengths are not ordered for all t, but fails to correctly show that the long-run average queue lengths need not be ordered in the anticipated way. The comparison here also applies to arbitrary k and m and yields a more striking comparison. The queue length was not correctly calculated for the $D/A/1/2$ model there. (The queue length there should be $X_2(t) = 0$ for $6k + 1 + \epsilon \leqslant t < 6k + 2$ and $6k + 6 + 2\epsilon \leqslant t < 6k + 6 + 1$, $X_2(t) = 2$ for $6k + 3 \leqslant t < 6k + 4 + \epsilon$ and $6k + 5 \leqslant t < 6k + 6 + \epsilon$, for $k \geqslant 0$, with $X_2(t) = 1$ elsewhere. As given, the example shows that we do not have $X_1(t) \leqslant X_2(t)$ for all t (which was the main point in [2]). However, the long run averages there are $\bar{X}_1 = 1 < 1 + (5\epsilon/6) = \bar{X}_2$.)

The examples we construct in sections 2 and 4 are deterministic, but they are easily converted into models in the stationary stochastic process framework of Franken, König, Arndt and Schmidt [1], as we show in section 3. Moreover, the queue-length process can be made regenerative and aperiodic, so that the queue length converges in distribution as $t \to \infty$, while keeping the essential character of the deterministic examples.

2. A deterministic example for average queue lengths

We start with a deterministic example that shows that the long-run average queue lengths (number in system) are not increasing in the capacity k. We consider the case of $s = 1$ server. The number of waiting spaces is thus one less than the capacity. Let arrivals occur one at a time at each integer j for $j \geq 0$, so that the arrival process is a D process. At any j, if departures can occur, let them occur before the arrival, so that arrivals are always admitted at departure points.

We construct two systems with the same interarrival times and service times, one with capacity k and the other with capacity m, where $1 \leq k < m$. Arrivals finding the waiting room full are lost without affecting future arrivals (e.g., no retrials). Let both systems start out empty before the arrival at time 0.

Let $Q_i(t)$ be the queue length at time t in the system with capacity i. Let $\overline{Q}_i$ be the long-run average queue length in the system with capacity i, i.e.,

$$\overline{Q}_i \equiv \lim_{T\to\infty} \frac{1}{T}\int_0^T Q_i(t)\,\mathrm{d}t. \tag{1}$$

Recall that our object is to show that we can have $\overline{Q}_m < \overline{Q}_k$ when $k < m$. In particular, for any $\delta < 0$, we shall make a construction yielding

$$\overline{Q}_m < \delta \quad \text{and} \quad \overline{Q}_k = k. \tag{2}$$

Our example will depend on two parameters ϵ and n in addition to k and m. We choose a positive integer n and a small positive ϵ so that

$$2m\epsilon < 1 \quad \text{and} \quad 2m < n. \tag{3}$$

To obtain the extreme behavior in (2), we will make ϵ small and n large.

Our example is specified by giving the service times. Let the service time of customer j (the arrival at time j) be v_j. For nonnegative integers i, j, and n (as well as k and m), let

$$v_{ikn+j} = \begin{cases} n, & 0 \leq j \leq k-1, \\ m+k-1+\epsilon, & j = kn-m, \\ \epsilon, & \text{otherwise, } j \leq kn-1. \end{cases} \tag{4}$$

It is easy to see that

$$Q_k(t) = k \quad \text{for } t \geq k-1, \tag{5}$$

because customers $in+j$ with $0 \leq j \leq k-1$ are admitted, but no others; i.e., only the customers with the long (n) service times get in and they keep the k-capacity system full. From (5), it is obvious that $\overline{Q}_k = k$ as claimed in (2).

On the other hand, in the m-capacity system customers with the shorter service times do get in and they eventually block all the customers with longer (n) service times. The case of $k=3$, $m=4$, $n=10$ and $\epsilon = 0.1$ is displayed in fig. 1.

In particular, it is easy to see that

$$Q_m(t) = m \quad \text{for } m-1 \leq t \leq kn. \tag{6}$$

At time kn (after the departure and arrival then), the m-capacity system contains $m-1$ customers with service time ϵ and the arrival at kn with service time n. (Since $2m < n$, the customer arriving at $kn-m$ did not get in.)

Since the service discipline is FCFS, customer kn cannot start service at time kn. However, the $(m-1)$ customers with ϵ service time in the system at time kn are gone by time $kn+(m-1)\epsilon < kn+1$. Hence, customer kn starts service at time $kn+(m-1)\epsilon$. At time $kn+k-1$, the m-capacity system contains the k customers $kn+j$, $0 \leq j \leq k-1$, all with service time n. Consequently, we have

$$Q_m(t) = m \quad \text{or} \quad m-1 \quad \text{for } kn+m-1 \leq t \leq 2kn. \tag{7}$$

Moreover, $Q_m(2kn) = m$ and at time $2kn$ customer $kn+k-1$ is in service with $(m-1)\epsilon$ remaining service time. Hence, customer $2kn$ is not admitted. (We have now succeeded in blocking a customer with a long service time.)

At time $2kn$, customer $kn+k-1$ is in service with $(m-1)\epsilon$ remaining service time, while $m-1$ customers are in queue with ϵ service time. Since $2(m-1)\epsilon < 1$ by (3), at time $2kn+2(m-1)\epsilon$, and thus also until time $2kn+1$, the m-capacity system is empty.

From the above, we see that $Q(2kn+k-1) = k-1$ with these $k-1$ customers all having service time n, with the customer in service having remaining service time $n-(k-2)$. Consequently, $Q((3k-1)n+1) = m$, where all these m customers have ϵ service times. Moreover, $Q((3k-1)n+2) = 1$. Indeed,

$$Q(3kn-j) = \begin{cases} 1, & m < j \leq n-2, \\ m-j+1, & 0 < j \leq m, \\ m, & j = 0. \end{cases} \tag{8}$$

In fact, a periodic pattern of period kn begins at time $3kn-n+2$. By (4), customer $3kn-m$ departs at $3kn+k-1+\epsilon$. Hence, none of the k customers $3kn+j$, $0 \leq j \leq k-1$, are admitted. (We have finally succeeded in blocking all customers with long service times.) However, the m customers in the system at time $3kn+k-1$ are all gone by $3kn+k$. Hence

$$Q(3kn+j) = 1, \quad k \leq j \leq n-m,$$

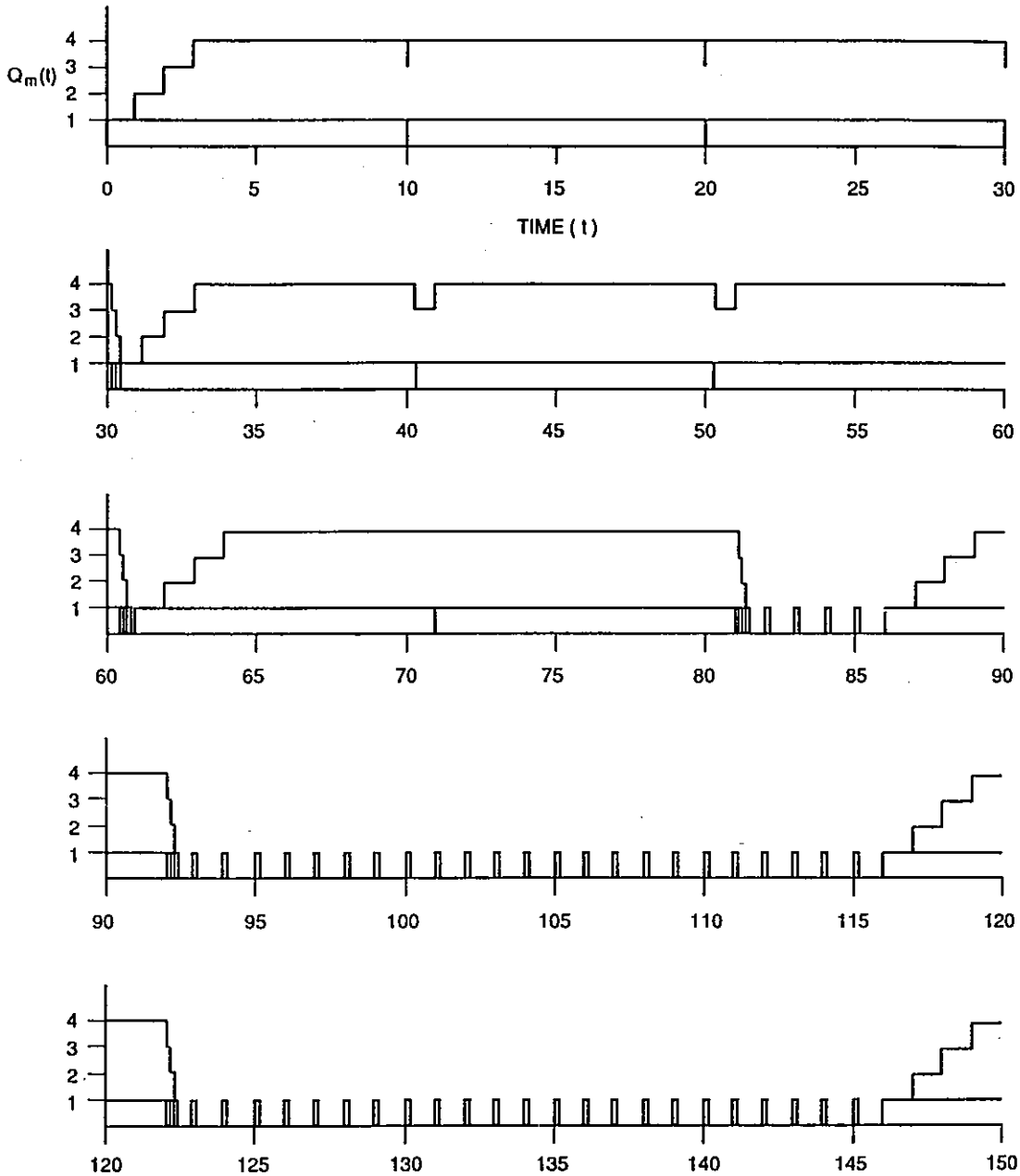

Fig. 1. A sample path of $Q_m(t)$ when $k=3$, $m=4$, $n=10$ and $\epsilon=0.1$. The service time of the customer in service is also shown. Customers with long ($n=10$) service times are admitted only at times 0, 1, 2, 30, 31, 32, 61 and 62.

and

$$Q(4kn-j)=m-j+1, \quad 0<j\leqslant m, \tag{9}$$

with customer $4kn-m$ being in service in the interval $[4kn-m, 4kn+k-1+\epsilon)$. Hence, none of the k customers $4kn+j$, $0\leqslant j\leqslant k-1$, are admitted.

As indicated above, periodic behavior of period kn begins at $3kn - n + 2$. For convenience, we calculate what happens in the intervals $[ikn - m, ikn + k]$ and $[ikn + k, (i+1)k - m]$ for any $i \geqslant 4$. First,

$$\begin{aligned}
&\int_{ikn-m}^{ikn+k} Q_m(t)\,\mathrm{d}t \\
&\quad = \int_{ikn-m}^{ikn} Q_m(t)\,\mathrm{d}t + \int_{ikn}^{ikn+k-1} Q_m(t)\,\mathrm{d}t + \int_{ikn+k-1}^{ikn+k} Q_m(t)\,\mathrm{d}t \\
&\quad = (1+2+\cdots+m) + (m(k-1)) + (m\epsilon + (m-1)\epsilon + \cdots + \epsilon) \\
&\quad = \frac{m(m+1)}{2} + m(k-1) + \frac{m(m+1)\epsilon}{2} \\
&\quad = \frac{m}{2}[(m+1)(1+\epsilon) + 2(k-1)] \qquad (10)
\end{aligned}$$

and, second,

$$\int_{ikn+k}^{(i+1)kn-m} Q_m(t)\,\mathrm{d}t = (kn - k - m)\epsilon \quad \text{for } i \geqslant 4. \qquad (11)$$

Thus, the long-run average queue length is

$$\begin{aligned}
\bar{Q}_m &\equiv \lim_{T\to\infty} \frac{1}{T}\int_0^T Q_m(t)\,\mathrm{d}t \\
&= \frac{1}{kn}\int_{ikn-m}^{(i+1)kn-m} Q_m(t)\,\mathrm{d}t \\
&= \frac{m}{2}\,\frac{[(m+1)(1+\epsilon) + 2(k-1)]}{kn} + \frac{(kn-k-m)\epsilon}{kn} \\
&< \frac{3m^2}{2kn} + \epsilon. \qquad (12)
\end{aligned}$$

Hence, for any given k and m, it is easy to choose ϵ and n to make $\bar{Q}_m$ arbitrarily small.

It is interesting that the throughputs in these two systems are also quite different, with a reversed ordering. Let $D_i(t)$ be the number of departures in the interval $[0, t]$ in the system with capacity i. Let the throughput be

$$\theta_i \equiv \lim_{t\to\infty} D_i(t)/t. \qquad (13)$$

For our example, it is easy to see that $\theta_k = n^{-1}$, while $\theta_m = 1 - n^{-1}$. (For the system with capacity m, eventually only the k customers with long service times are blocked in each cycle of kn arrivals.) Consequently, for any k and m, we may choose n to make θ_m arbitrarily close to the arrival rate with θ_k arbitrarily small.

3. The stationary-process framework

Since the arrival times in section 2 are deterministic and the service times are periodic of period kn, it is easy to construct associated stationary input. It simply suffices to consider the input defined in section 2 with the time origin placed uniformly in the interval $[0, kn]$. Then the arrival point process together with the service times is a stationary marked point process in the sense of Franken et al. [1].

With this stationary input having the origin at x, where $0 \leqslant x \leqslant kn$, we obtain the given average queue lengths if we initialize with an empty system at time $-x$ using the specified input. However, there is not a unique stationary distribution for the k-capacity system. If we initialize differently, then we may obtain a different stationary regime. To see this, note that a different periodic stationary regime prevails in the k-capacity system if we start the system in section 2 empty at times t, $0 < t \leqslant kn - m$. Then the customers with long (n) service times do not get in the system. The k-capacity system then behaves like the m-capacity system. Customers $ikn - m$ cause the customers with large service times to be blocked.

To obtain a unique stationary distribution for the k-capacity system as well as the m-capacity system, it suffices to slightly modify the service times in (4) as follows. We let

$$v_{ikn+j} = k + \epsilon, \quad j = -(m+k),\ j \geqslant 1, \tag{14}$$

and let v_{ikn+j} be defined as in (4) otherwise. The arrival at time $ikn - (m+k)$ enters service then and stays in service until $ikn - m + \epsilon$. Hence, assuming that $Q(ikn - (mk)) = 1$, $Q(ikn - m) = k$, so that customer $ikn - m$ is admitted in the m-capacity system, but not in the k-capacity system. Since $2m < n$ by (3), $m + k \leqslant n - 2$, so that (14) does not affect what happens before the periodic pattern begins at $3kn - n + 2$.

We can further make the system regenerative and have queue lengths converge in distribution as $t \to \infty$ by creating regenerative cycles as follows. We start each cycle with the input in section 2 for a time interval of length lkn, where l is a large positive integer. Then we include kn arrivals at integer time points with zero service times followed by an independent exponential interval with mean 1. By choosing l large, the mean of the stationary (and limiting) distribution of this regenerative process can be made arbitrarily close to the average in section 2. (Bounds on the difference are easily computed.)

The starting point of the regenerative cycle above can be taken as the epoch a customer with non-zero service time arrives after there have been no arrivals for a period strictly greater than 1. (The exponential idle intervals plus the adjoining interarrival time are the only interarrival times not equal to one.) This condition guarantees that the regenerative cycle starts with an empty system immediately

after the exponential interval. The mean cycle length is obviously $(l+1)kn+1$. The exponential intervals also remove the periodic behavior.

We remark that with this regenerative modification we do not need the change in (14).

4. A deterministic example for throughput

We now modify the example in section 2 to achieve throughputs θ_m small with θ_k near 1. For this purpose, suppose that (3) holds and let the service times be

$$v_{in+j} = \begin{cases} k+\epsilon, & j=0, \quad i \geqslant 0, \\ n-k-1, & j=k, \quad i \geqslant 0, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Then $Q_k(in+k) = Q_m(in+k) = k$ for $m > k$, so that customer $in+k$ is never admitted to the k-capacity system but is always admitted to the m-capacity system. All other customers are admitted to the k-capacity system, so that $\theta_k = 1 - n^{-1}$. In the system with capacity m, customers $in+j$ with $0 \leqslant j \leqslant m-1$ are admitted and served, while customers $in+j$ with $m \leqslant j \leqslant n-1$ are blocked. Hence, $\theta_m = m/n$. Hence, for any k and m, we can choose n so that θ_k is arbitrarily close to 1 while θ_m is arbitrarily close to 0.

This model is easily converted to the stationary-process framework, just as in section 3.

References

[1] P. Franken, D. König, U. Arndt and V. Schmidt, *Queues and Point Processes* (Akademie-Verlag, Berlin, 1981).

[2] D.P. Heyman and W. Whitt, Limits for queues as the waiting room grows, Queueing Systems 5 (1989) 381–392.

[3] J. Keilson, The ergodic queue length distribution for queueing systems with finite capacity, J. Roy. Statist. Soc. Ser B 28 (1966) 190–201.

[4] M. Miyazawa, Complementary generating functions for the $M^X/GI/1/k$ and $GI/M^Y/1/k$ queues and their application to the comparison of loss probabilities, J. Appl. Prob. 27 (1990) 684–692.

[5] D. Sonderman, Comparing multi-server queues with finite waiting rooms, I: same number of servers, Adv. Appl. Prob. 11 (1979) 439–447.

[6] D. Sonderman, Comparing multi-server queues with finite waiting rooms, II: different numbers of servers, Adv. Appl. Prob. 11 (1979) 448–455.

[7] H.C. Tijms, *Stochastic Modelling and Analysis: A Computational Approach* (Wiley, New York, 1986).

[8] P. Tsoucas and J. Walrand, Monotonicity of throughput in non-Markovian networks, J. Appl. Prob. 26 (1989) 134–141.

[9] W. Whitt, Comparing counting processes and queues, Adv. Appl. Prob. 13 (1981) 207–220.