Deciding Which Queue to Join: Some Counterexamples

Ward Whitt

Operations Research, Vol. 34, No. 1 (Jan. - Feb., 1986), 55-62.

Stable URL:
http://links.jstor.org/sici?sici=0030-364X%28198601%2F02%2934%3A1%3C55%3ADWQTJS%3E2.0.CO%3B2-O

Operations Research is currently published by INFORMS.

DECIDING WHICH QUEUE TO JOIN: SOME COUNTEREXAMPLES

WARD WHITT

AT&T Bell Laboratories, Holmdel, New Jersey

(Received January 1984; accepted January 1985)

Consider a queueing system with two or more servers, each with its own queue with infinite capacity. Customers arrive according to some stochastic process (e.g., a Poisson process) and immediately upon arrival must join one of the queues, thereafter to be served on a first-come first-served basis, with no jockeying or defections allowed. The service times are independent and identically distributed with a known distribution. Moreover, the service times are independent of the arrival process and the customer decisions. The only information about the history of the system available for deciding which queue to join is the number of customers currently waiting and being served at each server. Joining the shortest queue is known to minimize each customer's individual expected delay and the long-run average delay per customer when the service-time distribution is exponential or has nondecreasing failure rate. We show that there are service-time distributions for which it is not optimal to always join the shortest queue. We also show that if, in addition, the elapsed service times of customers in service are known, the long-run average delay is not always minimized by customers joining the queue that minimizes their individual expected delays.

We often must decide which queue to join, so it is natural to expect that the question has been thoroughly analyzed. Unfortunately, however, queueing theory provides less help than we might expect. It would be nice if we could always use a simple rule such as "Join the Shortest Queue," but the decision clearly depends on the available information. Suppose that we want to minimize our individual expected delay. If we can estimate the service time of each waiting customer, for example, by looking at the shopping baskets in a supermarket, then we do not necessarily want to join the shortest queue. We want to join the queue with the least work (total service time) in front of us, which need not be the shortest queue.

In some situations, such as waiting for a teller at a bank, we have little basis for estimating the service times of waiting customers. Moreover, the customers need not actually be making the decisions. For example, the customers might be jobs in a production facility or packets in a communication network. Then system managers want to determine the optimal dynamic routing, using an algorithm that may depend only on the number at each queue.

When we know only the number of customers at each queue, the shortest-line rule seems like the obvious candidate, but is it optimal? To be more specific: consider a queueing system with two identical servers, each with its own queue having unlimited waiting space. Customers arrive according to some stochastic process and immediately upon arrival must join one of the queues, thereafter to be served on a first-come first-served (FCFS) basis with no jockeying or defecting allowed. The service times are independent and identically distributed (i.i.d.) and independent of the arrival process and the customer assignments. The only information available for deciding which queue to join is the number of customers being served or waiting at each server.

In this paper, we prove that the shortest-line rule is *not* always optimal. The shortest-line rule may be a good rule to use, but we show that it need not minimize each customer's individual expected delay or the long-run average delay per customer (which is equivalent to the expected equilibrium delay). Moreover, the shortest-line rule need not maximize the expected number of departures by a given time. We do not identify an optimal rule, but we do identify another rule that is better than the shortest-line rule in a particular setting.

The difficulty with the model is not the arrival process, because we also assume a Poisson arrival process; it is the service-time distribution. Even though the decision maker does not know the actual service-times of the customers already in the system when the decision must be made, there is extra information in the general service-time distribution. In fact, for the purely Markovian model with Poisson arrival process and exponential service times, Winston (1977) proved that the shortest-line rule is indeed optimal

Subject classification: 681 assigning customers to servers, 693 server-selection rules, 696 dynamic routing

0030-364X/86/3401-0055 $01.25

with regard to all these criteria and stronger ones based on the notion of stochastic order. Weber (1978) subsequently extended this result to service-time distributions with nondecreasing failure rate and arbitrary arrival processes. The case of an arbitrary arrival process was also treated by Ephremides, Varaiya and Walrand (1980). Lehtonen (1981) carried out a sample-space construction for the purely Markovian model, which implies stochastic order comparisons for the entire departure process (the discrete-time analogue of ordering $\leqslant_4$ in Whitt 1981a). When trying to obtain positive results, one discovers that some conditions on the service-time distribution are evidently necessary, but one also discovers that counterexamples are hard to construct because it is difficult to describe the behavior of any policy, e.g., it is hard to calculate the expected equilibrium delay.

As in Whitt (1981b, 1984), we use light-traffic asymptotics to construct our counterexamples. In Section 1 we construct a counterexample with two queues using a service-time distribution having a U-shaped failure rate. In Section 2 we discuss the issue of breaking ties. In Section 3 we change the information conditions. We assume that the decision maker knows the length of time each customer in service has been in service in addition to the queue lengths. (The service times are realized only through service.) With this information, it is possible to calculate what each customer's expected delay would be at each queue, using the conditional distributions of the remaining service times. A natural rule is to join the queue giving the shortest expected delay. However, we show in Section 3 that the rule that minimizes each customer's individual expected delay need not minimize the long-run average delay per customer. This is another queueing situation in which individual and social optima do not coincide; see Bell and Stidham (1983) and references that they cite.

In Section 4 we consider multiserver teams, where each team has its own queue with unlimited waiting space. Immediately upon arrival, customers must join one of the queues, thereafter to be served on a FCFS basis by the first available server in the team. This problem arises when routing telephone calls to groups of operators and has been studied by Houck (1982). For the case of i.i.d. exponential service times, Houck found in numerical examples that the rule that assigns customers to the team that minimizes their individual expected delays yields a long-run average delay for all customers nearly as small as in the combined system having a single queue and the FCFS rule. Since the combined system with the FCFS rule provides a lower bound for the long-run average delay (see Wolff 1977, Gittins 1978, and Smith and Whitt 1981), this shortest-expected-delay rule must be at least nearly optimal given that each team has its own queue, at least for the cases considered. However, we prove that this shortest-expected-delay rule in fact does not minimize the long-run average delay for multiserver teams. Unlike the previous two examples for single-server teams, the service-time distribution in this example is exponential.

Houck also had conjectured that the expected equilibrium delay using the shortest-expected-delay rule for two teams of exponential servers could be bounded above by the expected equilibrium delay in a simple overflow system, in which all arrivals are first routed to one team which has no extra waiting room (i.e., no queue), and all overflows are routed to the second team which has a queue. However, the preferred alternative in our counterexample in Section 4 is such an overflow system. Hence, the overflow system does not provide an upper bound on the expected equilibrium delay in general.

All the results in this paper are negative (counterexamples). In addition to previously mentioned papers, positive control results for related models are contained in Davis (1977), Larsen (1981), Larsen and Agrawala (1983), Ramakrishnan (1983) and Hajek (1984). Stidham (1984) is a more recent survey with many references.

As in Whitt (1981b, 1984), our counterexamples are constructed using light-traffic analysis. For early use of light-traffic analysis, see p. 295 of Benes (1965) and Bloomfield and Cox (1972). There are many other recent uses of this method; e.g., Daley and Rolski (1984), Pinedo and Wolff (1982), and Wolff (1982). Light-traffic analysis is especially useful in conjunction with heavy-traffic analysis to generate approximations; see Burman and Smith (1983a, b) and Reiman and Simon (1984a, b).

1. The Shortest-Line Rule

To specify the model, let there be two servers, each with its own queue with unlimited capacity, and let the arrival process be a Poisson process. (The arrival process is not critical, but the Poisson assumption simplifies the proof.) Let the service-time distribution be a mixture of a point mass at 0 with probability $1 - \epsilon$ and a point mass at n with probability ϵ, where ϵ is small. In this section the decision maker knows only the queue lengths upon arrival; the rest of the history is not known. The dominating alternative decision rule matches the shortest-line rule when any queue is empty and when the differ-

ence between the numbers is zero or one; but the dominating rule has customers join the *longer* queue when neither is empty and the difference is two more more.

This service-time distribution captures the essence of a distribution with a U-shaped failure rate. The idea, which we will develop further, is that with such a distribution a queue shorter by more than a single customer may indicate that a service completion is more likely to have occurred more recently there, so that the next service completion is more likely to occur at the other server with the longer queue. Moreover, the probability mass near the origin may lead to many other service completions at the same server, so that the new arrival would wait less and depart sooner by joining this longer queue. Using this idea, we will show that the dominating rule is better than the shortest-line rule for all sufficiently small ϵ.

Note that for any fixed arrival rate, our service-time distribution puts the system in light traffic as $\epsilon \to 0$. It is not critical that there be any mass at zero in the service-time distribution. For example, the mass at zero could be moved to γ where γ is appropriately small compared to ϵ, e.g., $\gamma = \epsilon^{10}$. Moreover, the service-time distribution need not have atoms. For example, the mass at γ could be replaced by a density over the interval $[0, \gamma]$. Only minor modification of the arguments are needed, providing that γ is indeed appropriately small.

Since most of the mass is at 0, with high probability many customers will depart immediately upon arrival before anyone stays in service. Eventually, however, both servers will become busy for the first time and the queues will begin to grow. We assume that the shortest line rule is used when one queue is empty or the difference between the queue lengths is one, and an unspecified procedure when the queue lengths are equal, e.g., at random. (See Section 2.) Then one server will finish and, with high probability, several other customers will depart since they have zero service times. In fact, if ϵ is very small, it is very likely that all customers in that queue will instantaneously depart. Eventually, however, one of two kinds of events will occur following such an epoch. Either one of the customers entering service in the queue with the service completion will have nonzero service time or a new arrival with nonzero service time will occur before the long queue has a service completion (and with high probability empties out).

The next arrival after either of these events will be the first arrival to see a difference in the queue lengths of more than one. (With small probability, the difference could be zero or one, in which case the two rules are still identical.) Now, with high probability, it is better for this next arrival to join the longer queue because the server with the longer queue has been in service longer and will depart sooner. Moreover, with high probability, all the other customers in this queue will depart at the same instant. By this argument, we see that, if the longer queue is not too long, then the first customer to see a difference of at least two should join the longer queue. Moreover, subsequent arrivals should also join the longer queue as long as the difference is not too great. (It is obvious that, for any ϵ and n, there is a cutoff point, but we do not try to determine it.)

The issue is not entirely settled, however, because we have only considered the first customers to see a difference in queue lengths of at least two. Subsequent departures could lead to the server with the longer queue having the most recent service completion, in which case it is clearly better to join the shorter queue. Moreover, the history is not known when the decision is made.

We shall consider the expected equilibrium delay, which is equivalent to the long-run average delay per customer. As part of our analysis, we will show that these limiting concepts are well defined for the two rules under consideration. The key to obtaining concrete results is to consider light traffic, which we achieve by letting $\epsilon \to 0$. We analyze the equilibrium delay by using the regenerative structure and considering a typical busy cycle. We first sketch the argument, and then give more details.

A busy cycle is the interval between successive arrivals to an empty system. Since ϵ is small, most busy cycles are a single interarrival time. For ϵ small, the probability that k or more customers in a busy cycle have nonzero service times is of order ϵ^k. (Since this point is critical, we give a precise statement and following proof). Hence, the probability of ever having at least one (two) busy servers in a busy cycle is of order ϵ (ϵ^2). A customer can find a difference of at least two in the queue lengths only if there are at least three nonzero service times in that busy cycle, which has probability of order ϵ^3. The important point is that the first such difference of at least two, after which customers should usually join the longer queue, occurs with probability of order ϵ^3, but additional nonzero service times, after which the desirable behavior is unclear, occur with probability of order ϵ^4.

The upshot of this analysis is that, as $\epsilon \to 0$, the proportion of those arrivals seeing a difference of at least two in the queue lengths (who must arrive after the third nonzero service time) who arrive after a

fourth nonzero service time is asymptotically negligible. With very high probability (approaching one as $\epsilon \to 0$), the most recent service completion has occurred at the server with the shorter line when the difference in queue lengths is at least two.

Moreover, since the arrival rate and the nonzero service time are fixed, the probability of the queue growing very long given any fixed number of nonzero service times is asymptotically negligible as $\epsilon \to 0$. Hence, the alternative rule is better than the shortest-line rule for all sufficiently small ϵ. This result is obviously true not only for the expected equilibrium delay, but for the other criteria as well. (Detailed calculations for the expected equilibrium delay are given later.)

We have not identified an optimal rule, however, because our new rule can be improved too. For any given ϵ and n, arbitrarily long queue lengths are possible. If the difference between the queue lengths is sufficiently great, then it is better to join the shorter queue even if a service completion is about to occur in the longer queue. For small ϵ, this improvement is asymptotically negligible compared to the previous improvement, but it shows that we have not found an optimal policy.

Now we provide extra details. In particular, we first show that the probability that at least k customers in a busy cycle have nonzero service times is indeed of order ϵ^k for ϵ sufficiently small. We then indicate how to express the expected equilibrium waiting time in powers of ϵ, so that the desired comparison can be made.

Let $A(t)$ count the number of arrivals in the interval $[0, t]$ for $t \geq 0$. We have assumed that $A(t)$ is a Poisson process with arrival rate λ. Let $N(t)$ count the number of arrivals in the interval $[0, t]$ that have nonzero service times, i.e., service times of length n. (These customers need not begin service in the interval $[0, t]$, however.) Since $N(t)$ is obtained from $A(t)$ by Bernoulli splitting with probability ϵ, $N(t)$ is a Poisson process with rate $\lambda\epsilon$.

Let a busy cycle be the time interval between successive arrivals that find the system completely empty; let C denote the number of customers served in a busy cycle; and let B denote the number of customers with nonzero service served in a busy cycle.

Theorem 1. *For any positive integer k, there are positive constants M_1 and M_2 such that*

$$M_1\epsilon^k \leq P(B \geq k) \leq M_2\epsilon^k$$

for all ϵ, $0 < \epsilon < 1$, and all rules for assigning customers to the servers.

To establish Theorem 1, we use the following elementary lemma about a Poisson random variable X with mean λ.

Lemma. *For all $\lambda \geq 0$ and positive integers k,*

$$\lambda^k \geq P(X \geq k) \geq \lambda^k e^{-\lambda}/k!.$$

Proof. The lower bound is just $P(X = k)$. The upper bound calculation is

$$P(X \geq k) = \sum_{j=k}^{\infty} e^{-\lambda}\lambda^j/j! = \lambda^k \sum_{j=0}^{\infty} e^{-\lambda}\lambda^j/(j+k)!$$

$$\leq \lambda^k \sum_{j=0}^{\infty} e^{-\lambda}\lambda^j/j! = \lambda^k.$$

Proof of Theorem 1. Let D be the event that the first customer in the busy cycle has a nonzero service time. For $k \geq 1$,

$$P(B \geq k) = \epsilon P(B \geq k \mid D),$$

where

$$P(N(n) \geq k-1) \leq P(B \geq k \mid D) \leq P(N(n(k-1)) \geq k-1)$$

because, first, if we have $k-1$ more nonzero times in the interval $[0, n]$, then clearly $B \geq k$ given that the first customer has a nonzero service time. Secondly, if $B \geq k$, then clearly there must be at least $k-1$ more nonzero service times in the interval $[0, (k-1)n]$.

Next we can invoke the Lemma to get

$$P(N(n(k-1)) \geq k-1) \leq (\lambda\epsilon n(k-1))^{k-1}$$

and

$$P(N(n) \geq k-1) \geq (\lambda\epsilon n)^{k-1}e^{-\lambda\epsilon n}/(k-1)!,$$

so that the claimed bounds hold with the constants

$$M_1 = (\lambda n)^{k-1}e^{-\lambda n}/(k-1)! \quad \text{and} \quad M_2 = (\lambda n(k-1))^{k-1}.$$

We now provide extra details on calculating the expected equilibrium waiting time. By regenerative process theory, the expected equilibrium waiting time, say EW, is the expected waiting time per busy cycle divided by the expected number of customers served in a busy cycle, i.e.,

$$\text{EW} = E\sum_{k=1}^{C} W_k/\text{EC},$$

where $W_1 = 0$ is the waiting time of the initial customer who finds the system completely empty. Given Theorem 1, the following light-traffic asymptotics are not difficult.

Theorem 2. *For both decision rules,*

$$\mathrm{EC} = 1 + c_1\epsilon + c_2\epsilon^2 + o(\epsilon^2),$$

where c_1 and c_2 are independent of the rule. Moreover,

$$E\left(\sum_{k=1}^{c} W_k\right) = b_2\epsilon^2 + b_3\epsilon^3 + o(\epsilon^3),$$

where b_2 is independent of the rules while b_3 is not. Hence,

$$\mathrm{EW} = \frac{b_2\epsilon^2 + b_3\epsilon^3 + o(\epsilon^3)}{1 + c_1\epsilon + c_2\epsilon^2 + o(\epsilon^2)}$$

$$= b_2\epsilon^2 + (b_3 - c_1b_2)\epsilon^3 + o(\epsilon^4). \qquad (1)$$

As a consequence of Theorem 2, it suffices to consider the leading non-identical coefficient in the expansion of EW in powers of ϵ, i.e., $b_3 - c_1b_2$ in (1). The coefficients c_1 and b_2 are independent of the rule, but b_3 is smaller with the alternative decision rule.

2. Breaking Ties

In this section we discuss rules for breaking ties, i.e., rules for assigning customers to single-server queues when only the queue lengths are known and some of the queue lengths are equal. Since all such rules can be regarded as being consistent with the shortest-line rule, we did not consider this aspect in Section 1.

One possible rule is an *ordered-selection rule.* This rule assumes that the queues can be numbered. Then the customer joins the lowest numbered queue among those with equal queue lengths. Note that this rule does not use any information about past assignments, which we have assumed is not available.

For the example in Section 1, it is easy to see that the ordered-selection rule for breaking ties provides the basis for a significant improvement. Of course, the rule does not matter until both servers are busy, but with the ordered-selection rule, when both servers are busy, we know that with very high probability, as $\epsilon \to 0$, server 1 will have a service completion first. This is so because, with high probability as $\epsilon \to 0$, these are the only nonzero service times in the current busy cycle. This observation allows us to improve the coefficient b_2 in (1). Moreover, there is opportunity for further improvement. The new rule using the ordered-selection rule until both servers are busy followed by all customers joining queue 1 is better than the rule developed in Section 1 for that example for all sufficiently small ϵ. The rule in Section 1 still dominates the shortest-line rule and is valid if either the servers are indistinguishable or if ties must be broken by the random-assignment rule.

It also appears that the ordered-selection rule should be used to break ties with increasing failure rate (IFR) service-time distributions if the ages are not known and the servers are distinguishable, but we have not yet proved this result. Remark (i) on p. 412 of Weber might be construed as missing this point, but Weber is claiming optimality only under the condition that the random-assignment rule must be used to break ties. It is easy to see that the ordered-selection rule to break ties can improve the shortest-line rule for IFR distributions. We expect that this rule is usually nearly optimal, but we would be surprised if it were always optimal.

For decreasing failure rate (DFR) distributions, it is natural to use ordered-selection rule to break ties among idle servers and then reverse the order when the servers are busy.

3. The Shortest-Expected-Delay Rule

When the service-time distributions are not exponential, it is natural to consider the shortest-expected-delay (SED) rule, in which customers join the queue that will minimize their individual expected delays. However, if the residual service-time distributions are not known, then the individual expected delays depend on the rules used by other customers, which suggests a game-theoretic approach (that we will not investigate in this paper).

Suppose that the ages (times in service) are known, so that it is possible to compute each customer's individual expected delay at each queue. It is simply the expected remaining service time of the customer in service given the age plus the product of the number of customers waiting and the expected service time. Weber showed that if the service-time distribution is IFR then the SED rule minimizes the expected equilibrium delay. We now show that this is not true for all service-time distributions.

Here is our example: Let the service-time distribution be a two-point distribution assuming the values 1 and 5 each with probability ½. Let there be two servers and a Poisson arrival process with arrival rate ϵ, so that the traffic intensity is $\rho = 1.5\epsilon$. (Counterexamples are even easier to construct with other arrival processes.) Consider an arrival epoch when both servers are busy but nobody else is waiting. Let the ages be $1 - \delta$ and 3, so that the expected residual service times are $2 + \delta$ and 2, respectively. The SED rule dictates joining the first queue with age 3 for all δ and ϵ with $0 \leq \delta < 1$ and $0 \leq \epsilon < \frac{2}{3}$. (We need $\epsilon < \frac{2}{3}$ only for stability.) We will now show that the other action can be better for sufficiently small ϵ and δ.

First, note that if there is only one arrival in the next interval of length 10, then it is necessary to consider only the current customer and the following one; the current decision can have no further effect because then the system is completely empty either before the next arrival or before the subsequent arrival. (The length 10 is somewhat more than necessary.) The probability of having two or more arrivals in such an interval is of order ϵ^2, while the probability of having one is of order ϵ. Hence, for all sufficiently small ϵ, it suffices to assume that there is at most one arrival in this interval. Given any small ϵ, so that ϵ^2 is negligible compared to ϵ, we then choose δ so that δ is also negligible compared to ϵ.

It is now easy to calculate the expected delay for the next customer given the SED rule and its alternative. First suppose that the SED rule is used for both the current customer and the next customer. Let $1_A(u)$ be the indicator function of the set A, i.e., $1_A(u) = 1$ if $u \in A$ and 0 otherwise. Given exactly one arrival in the next time interval [0, 10] which occurs at u, the conditional expected delay for the next customer is

$$f_1(u) = (2 + \delta - u)1_{[0,\delta)}(u) + \frac{(4 + \delta - u)}{2} 1_{[\delta,3)}(u) + \frac{(4 + \delta - u)}{4} 1_{[3,4+\delta)}(u).$$

Next suppose that the alternative rule is used for the current customer, but the SED rule is used for the next customer. The conditional expected delay is then

$$f_2(u) = (2 - u)1_{[0,1+\delta)}(u) + 3\,\frac{(2 - u)}{4} 1_{[1+\delta,2)}(u)$$

and

$$\begin{aligned} f_1(u) - f_2(u) &= \delta 1_{[0,\delta)}(u) + \frac{(\delta + u)}{2} 1_{[\delta,1+\delta)}(u) \\ &\quad + \frac{(2 + 2\delta + u)}{4} 1_{[1+\delta,2)}(u) + \frac{(4 + \delta - u)}{2} 1_{[2,3)}(u) \\ &\quad + \frac{(4 + \delta - u)}{4} 1_{[3,4+\delta)}(u) \\ &\geqslant 0 \end{aligned}$$

for all u, $0 \leqslant u \leqslant 10$. Since the conditional distribution of the next arrival, given that there is exactly one arrival in [0, 10], is uniform over [0, 10], the sum of the expected delays always using the SED rule is greater by

$$\epsilon \int_0^{10} [f_1(u) - f_2(u)]\, du - O(\epsilon^2).$$

Hence, by choosing ϵ so that we can neglect $O(\epsilon^2)$, we see that the alternative is strictly better.

There are two remaining details. First, we have established that there is a situation in which it is better not to use the SED rule, but any specific pair of ages will be realized with probability zero. However, the example remains valid for ages in the intervals $[1 - \delta, 1)$ and $(3 - \delta, 3]$, respectively, for appropriate small positive δ. Hence, even though the situation we have analyzed is rare, it will occur with probability one, and affects the expected equilibrium delay.

Next, the second customer arriving in the interval [0, 10] could find ages in the interval $[1 - \delta, 1)]$ and $(3 - \delta, 3]$, which would mean that the SED rule should not be used with our alternative rule, contrary to our assumption for calculating $f_2(u)$, but the probability of these ages occurring is of order δ, so we can disregard it by choosing δ sufficiently small.

4. Teams of Exponential Servers

We now consider two teams of exponential servers, with each team having its own queue and operating according to the FCFS rule. We show that the SED rule does not minimize the expected equilibrium delay even with i.i.d. exponential service times. We consider two teams of different sizes with a Poisson arrival process. Then, even though the individual service-time distributions are identical, the delay distributions can be quite different at the two queues. The key is to consider teams of very different sizes in light traffic.

Our alternative to the SED rule is an overflow (OF) rule. With this rule, all arriving customers are routed to the first team and served there if there are any free servers. All customers finding no free servers at the first team are routed to the second team, where there is a queue. The SED rule evidently is typically better than the OF rule, but we show that this is not always the case.

Let the first team have a single server and let the other team have n servers, where n is large. Let the arrival rate and the individual service rate equal one, so that $\rho = 1/(n + 1)$ in the combined system. Note that the SED rule and the OF rule coincide until all servers are busy and there are n customers waiting in addition to n in service at the second team at an arrival epoch. Then the expected delay before beginning service for this arrival would be 1 at the first team

and $(n + 1)/n$ at the second team. Also note that the delay distribution at the first team is exponential with mean 1 and, by the law of large numbers, at the second team it is almost deterministic. The SED rule dictates sending this arrival to the first team. However, it is easy to see that the OF rule is better in this situation.

First, at the time of the next arrival, with all but negligible probability, there will be many departures from the second team, so that we need not consider the possibility of any future arrivals going to team 1 unless team 1 is empty. Moreover, the effect of an additional customer at the second team on the expected delay of any future customer is at most $1/n$. The issue is how the current assignment affects the probability that team 1 is empty. If the OF rule is used, then team 1 will be empty for the next arrival with probability ½. If the SED rule is used, then the probability is ¼. (Two service completions must occur at the first team before the next arrival.)

Suppose that the first customer uses the OF rule and joins the second team. Approximately, ignoring terms that are asymptotically negligible as $n \to \infty$, the second customer will have delay 0 with probability $(1 + e^{-2})/2$ and will join the second team and have delay $1 - t$ with density e^{-2t}, $0 \leq t \leq 1$. The expected delay for the second customer is thus $(1 + e^{-2})/4 = 0.284$. On the other hand, suppose that the first customer uses the SED rule and joins the first team. Then the second customer will have delay 0 with probability $(1 + 5e^{-2})/4$ and will join the second team and have delay $1 - t$ with density $e^{-2t} + te^{-2t}$, $0 \leq t \leq 1$. The expected delay for the second customer is thus $(1 + 4e^{-2})/4 = 0.385$. There is clearly a significant impact on the next arrival and could be on future arrivals as well (in the same direction). On the other hand, the cumulative impact on the expected delays of all future customers of an extra customer at the second team is of order $1/n$.

It is of course possible that the SED rule would begin to dominate the OF rule as more arrivals are sent to the second team when it has at least $2n$ customers, but with the OF rule, ever having $n + k + 1$ customers is asymptotically negligible compared to having $n + k$ for any k. (The arrival rate is 1 and the departure rate is n.) Hence, the comparison of the expected equilibrium delays has been established.

5. Concluding Remarks and Directions for Research

1. The example in Section 4 is easily modified to cover two single-server teams with different exponential service-time distributions. For the example, one server at rate n is essentially the same as n servers at rate one. Ephremides, Varaiya and Walrand comment that the SED rule is not optimal for two heterogeneous exponential servers, but they do not prove that SED does not minimize the long-run average delay. Variants of this model are also treated by Davis, Larsen, Larsen and Agrawala, Ramakrishnan, and Hajek.

2. We have shown that several natural selection rules are not optimal in various situations, but we have not identified any optimal rules. Identifying optimal rules in these situations would obviously be interesting, but appears to be difficult. Moreover, knowing an optimal rule might not be so useful because the optimal rule may be very complicated.

3. We conjecture that the positive results for single-server teams extend to multiserver teams all having the same number of servers.

4. Even though we have shown that certain natural simple rules are not optimal, we have not shown that they are typically bad. In fact, they may be nearly optimal; see Houck. It would be nice to quantify and perhaps even bound the degree of non-optimality under various modeling assumptions. Natural simple rules such as the threshold in Larsen and Agrawala are obviously worth additional study even if they are not optimal. It should be possible to make useful comparisons and establish heuristic design principles; see Whitt (1985).

5. A possible refinement of the shortest-expected-delay rule is to assign customers to servers so that the expected sum of the delays of that customer plus the next one or two customers is minimized or approximately minimized. This strategy requires calculating the delay distributions at the servers.

6. In contrast to our light-traffic examples, Kingman (1961), Foschini and Salz (1978) and Reiman (1983) have shown that in heavy traffic the shortest-line rule in the setting of Section 1 behaves as well as the combined system with the FCFS discipline, so that the shortest-line rule is asymptotically optimal in heavy traffic. (Reiman first treats general service-time distributions.)

7. In Section 3 we showed that the expected equilibrium waiting time is not minimized by having each customer minimize his expected waiting time upon arrival, given the model specification, the queue lengths and the ages. However, if we do not know the ages, then it is not possible to compute such expected waiting times without knowing the other customers' rules. It is thus natural to look for a noncooperative equilibrium rule, such that when all customers follow it, none is motivated to deviate. There are many

questions: When does such an equilibrium rule exist? When is it unique? What is its structure? When does it yield the minimum equilibrium waiting time? In Section 1 we have shown that the shortest-line rule is not such an equilibrium rule (for some service-time distributions).

8. As we have seen, much depends on the information available. It would be interesting to systematically investigate the impact of different kinds of information. Then we could compare the benefits of different kinds of information with the costs of getting it.

Acknowledgment

I am grateful to David Houck, Moshe Segal, Bob Smith, Y. T. Wang, Richard Weber and the referees for their assistance.

References

BELL, C. E., AND S. STIDHAM, JR. 1983. Individual versus Social Optimization in the Allocation of Customers to Alternative Servers. *Mgmt. Sci.* **29,** 831–839.

BENES, V. 1965. *Mathematical Theory of Connecting Networks and Telephone Traffic.* Academic Press, New York.

BLOOMFIELD, P., AND D. R. COX. 1972. A Low Traffic Approximation for Queues. *J. Appl. Prob.* **9,** 832–840.

BURMAN, D. Y., AND D. R. SMITH. 1983a. A Light-Traffic Theorem for Multi-server Queues. *Math. Opns. Res.* **8,** 15–25.

BURMAN, D. Y., AND D. R. SMITH. 1983b. Asymptotic Analysis of a Queueing Model with Bursty Traffic. *Bell Syst. Tech. J.* **62,** 1433–1453.

DALEY, D. J., AND T. ROLSKI. 1984. A Light Traffic Approximation for a Single-Server Queue. *Math. Opns. Res.* **9,** 624–628.

DAVIS, F. 1977. Optimal Control of Arrivals to a Two-Server Queueing System with Separate Queues. Ph.D. dissertation, Program in Operations Research, North Carolina State University at Raleigh.

EPHREMIDES, A., P. VARAIYA AND J. WALRAND. 1980. A Simple Dynamic Routing Problem. *IEEE Trans. Aut. Control* **25,** 690–693.

FOSCHINI, G., AND J. SALZ. 1978. A Basic Dynamic Routing Problem and Diffusion. *IEEE Trans. Comm.* **26,** 320–327.

GITTINS, J. C. 1978. A Comparison of Service Disciplines for *GI/G/m* Queues. *Math. Opnsforsch. Stati. Ser. Opt.* **9,** 255–260.

HAJEK, B. 1984. Optimal Control of Two Interacting Service Stations. *IEEE Trans. Aut. Control* **AC-29,** 491–499.

HOUCK, D. J. 1982. Algorithms for Routing Calls to Parallel Queueing Systems. Unpublished work.

KINGMAN, J. F. C. 1961. Two Similar Queues in Parallel. *Ann. Math. Stat.* **32,** 1314–1323.

LARSEN, R. L. 1981. Control of Multiple Exponential Servers with Applications to Computer Systems. Ph.D. dissertation, Department of Computer Science, University of Maryland.

LARSEN, R. L., AND A. K. AGRAWALA 1983. Control of a Heterogeneous Two-Server Exponential Queueing System. *IEEE Trans. Software Engr.* **SE9,** 522–526.

LEHTONEN, T. 1981. On the Optimality of the Shortest Line Discipline. Chapter IV in Stochastic Comparisons for Many Server Queues. Ph.D. dissertation, The Helsinki School of Economics.

PINEDO, M., AND R. W. WOLFF. 1982. A Comparison between Tandem Queues with Dependent and Independent Service Times. *Opns. Res.* **30,** 464–479.

RAMAKRISHNAN, K. K. 1983. The Design and Analysis of Resource Allocation Policies in Distributed Systems. Ph.D. dissertation, Department of Computer Science, University of Maryland.

REIMAN, M. I. 1983. Some Diffusion Approximations with State Space Collapse. In *Proceedings of the International Seminar on Modeling and Performance Evaluation Methodology.* Springer-Verlag, New York.

REIMAN, M. I., AND B. SIMON. 1984a. An Interpolation Approximation for Queueing Systems with Poisson Input. AT&T Bell Laboratories, Murray Hill, N.J.

REIMAN, M. I., AND B. SIMON. 1984b. Open Queueing Systems in Light Traffic. AT&T Bell Laboratories, Murray Hill, N.J.

SMITH, D. R., AND W. WHITT. 1981. Resource Sharing for Efficiency in Traffic Systems. *Bell Syst. Tech. J.* **60,** 39–55.

STIDHAM, S. JR. 1984. Optimal Control of Admission to a Queueing System. Department of Industrial Engineering and Operations Research, North Carolina State University at Raleigh.

WEBER, R. W. 1978. On the Optimal Assignment of Customers to Parallel Servers. J. Appl. Prob. **15,** 406–413.

WHITT, W. 1981a. Comparing Counting Processes and Queues. *Adv. Appl. Prob.* **13,** 207–220.

WHITT, W. 1981b. On Stochastic Bounds for the Delay Distribution in the *GI/G/s* Queue. *Opns. Res.* **29,** 604–608.

WHITT, W. 1984. Minimizing Delays in the *GI/G/1* Queue. *Opns. Res.* **32,** 41–51.

WHITT, W. 1985. The Best Order for Queues in Series. *Mgmt. Sci.* **31,** 475–487.

WINSTON, W. 1977. Optimality of the Shortest Line Discipline. *J. Appl. Prob.* **14,** 181–189.

WOLFF, R. W. 1977. An Upper Bound for Multi-channel Queues. *J. Appl. Prob.* **14,** 884–889.

WOLFF, R. W. 1982. Tandem Queues with Dependent Service Times in Light Traffic. *Opns. Res.* **30,** 619–635.