A Retrospective: Which Queue to Join?

Ward Whitt

Department of Industrial Engineering and Operations Research,
Columbia University, New York, NY, 10027 {ww2040@columbia.edu}

January 19, 2014

1 The Problem: Deciding Which Queue to Join

You need service, but you must wait your turn. You must join one of two queues. Which queue should you join? In particular, suppose that one queue has 3 waiting customers while the other has 27. Would you ever join the longer queue?

Yes, of course you might join the longer queue, because the service offered could be different. For example, at a movie theater showing several different films, there may be a different ticket line for each film.

But consider the case in which the same service is provided by each server, and our sole objective is to minimize our expected delay (waiting time in queue before starting service). Even then, we might join the longer queue, because the best decision clearly depends on the information available. For example, consider a supermarket. Some queues may be designated as express lines, where customers can have at most ten items. They clearly will move faster than the ordinary lines. Even among non-express lines, we may have experience about the processing ability of the checker or we can directly estimate the required service time of each waiting customer by looking at their shopping baskets. But suppose that the only information available to arriving customers for deciding which queue to join is the number of customers waiting or being served at each server. In this "low-information" setting, is it not obvious that it is better to join the shorter queue?

To further clarify the problem, suppose that each server has its own queue and that customers must decide immediately upon arrival which queue to join, thereafter to be served one at a time by their chosen server in a first-come first-served basis, with no jockeying (moving from one queue to another) or defections (leaving the system before receiving service) allowed. Suppose that the service times of all customers come from a sequence of independent and identically distributed ran-

dom variables, that are independent of the arrival process and the customer assignments. Suppose that the only information available to arriving customers for deciding which queue to join is the number of customers waiting or being served at each server. In this setting, would any sane person join the queue with 27 customers instead of the queue with 3 customers?

2 Is There Anything to Think About?

Consistent with our intuition about what must be best, Winston [**?**] proved that indeed joining the shorter queue minimizes the expected steady-state delay for the Markovian queueing model with Poisson arrival process and exponential service times. But could there possibly be service-time distributions for which joining the shorter queue is actually dominated by another policy? Is further generalization not just a harder proof?

However, in [**?**] we identified service-time distributions for which another decision rule is better than joining the shorter queue. For the special service-time distribution (introduced below), the alternative rule has arrivals join the shorter queue whenever the two queues are equal, breaking ties at random in both cases, or differ by only a single customer. However, the alternative rule has the arriving customer join the *longer* queue whenever the queue lengths differ by two or more. We do not join the shorter queue precisely when it seems most beneficial to do so, as in the initial example with queue lengths of 3 and 27.

3 All Service-Time Distributions Are Not the Same

For some service-distributions, finding one queue longer than the other by at least two customers can signal that it is actually usually better to join the longer queue. That state can provide important information about the remaining service times at the two servers. And, when most service times are very short, those remaining service times may produce most of the delay experienced by this new arrival.

It is helpful to consider a special service-time distribution that is easy to analyze (which will illustrate what must be true more generally). Let the service-time distribution take only two values. Let each service time be a fixed positive value m with probability p and 0 with probability $1 - p$, where p is very small. Since p is very small, most of the service times are 0. Usually there are no queues at all, but eventually a customer with service time m will arrive and enter service. Both joining the shorter queue and the alternative decision rule rule will have subsequent arrivals go to

the other idle server, where they usually will receive service instantly because the service time is 0, but eventually both servers will become busy serving customers with positive service times. Then the two queues will build up with waiting customers. Both policies dictate that the queues will differ by at most 1 until there is a service completion.

The key is to observe what happens at that first service completion epoch after both servers become busy. When there is a service completion at one of the queues, many of the waiting customers in that queue will depart together with the customer that was in service, because they have 0-length service times. Indeed, if p is very small, usually the entire queue will empty out at that service completion epoch. However, there will occur times when one of the waiting customers has service time m, causing some customers to remain in that queue. But then the queue lengths usually will differ by at least 2.

Now we claim that subsequent customers, arriving right after that service completion epoch, should join the longer queue. We can deduce from the queue difference of at least 2 that a service completion must have occurred more recently by the server with the shorter queue. Since all positive service times have the constant value m, the next service completion should occur at the server with the longer queue. Moreover, after that service time is complete, all waiting customers are likely to depart together, making the expected waiting time less for this new arrival if the new arrival joins the longer queue.

4 Can It Matter?

If indeed the probability p in this special service-time distribution is small, then most arriving customers can start service immediately, without any delay. From the long-run average point of view, the policy matters little. And, when arrivals first have to wait, both decision rules make the same decisions. The only difference occurs at those rare instants when the two queues are both positive and differ by two or more.

But, to the arrivals finding positive queues that differ by at least two, it can indeed matter. For these arriving customers, the difference in the expected waiting time could be up to m and would be of that order. (As first $m \uparrow \infty$ and then $p \downarrow 0$, the expected difference will approach $2m/3$; see the final section below.) To have the choice matter to those special arriving customers, simply choose m sufficiently large. Also note that the alternative decision rule does not tend to cause significant extra delays for other customers. The alternative rule tends to be better for all.

5 How to Close the Deal?

Of course, the full situation is more complicated, making a proper analysis difficult. After the event described above, where the queues first differ by at least 2, there will be new arrivals and service completions, and new customers with service time m entering service, so that we no longer are in the scenario above. Thus, even though we understand the idea in the example above, we may not yet be convinced that the alternative rule is actually better than joining the shorter queue. Even if we are convinced, we are left with the mathematical challenge of providing a rigorous proof.

In order to rigorously establish the claimed non-optimality of joining the shorter queue, in [**?**] we resorted to "light-traffic asymptotic analysis" in which, for any specified m, we let p decrease toward 0. In that way, it is possible to show that the subsequent complex scenarios that would dictate that it is better to join the shorter queue are highly unlikely compared to the scenario we have described. It is then possible to prove that the alternative policy produces lower expected steady-state delay than joining the shorter queue for sufficiently small p. The proof strategy is quite general, so that it can be used in other settings.

6 The Asymptotic Analysis: A Sketch

The first idea is to consider a stationary model for which steady-state is well defined. As in [**?**], that is achieved, first, by assuming that customers arrive according to a Poisson process and, second, by choosing the parameter p of our service-time distribution sufficiently small so that the intervals between successive times at which an arrival comes to find an empty system are independent and identically distributed random variables with finite mean.

The existence of these cycles make the stochastic process representing the state of the system a regenerative stochastic process, for which the steady-state distribution can be expressed in terms of the behavior over a single cycle. With that framework, we can show that the probability any arrivals have to wait in a cycle is of order $O(p^2)$ as $p \downarrow 0$, because we need two arrivals during the cycle that have service times m. We then find that the probability of the special event we have described is of order $O(p^3)$, because it requires yet another customer in the cycle to have service time m. However, the probability of all the more complex scenarios that might possibly favor joining the shorter queue are of order $O(p^4)$, because they require one more customer in the cycle to have service time m. As $p \downarrow 0$, the difference between the two policies becomes of order $O(p^3)$, and in that small order the alternative rule has a lower mean steady-state delay.

7 The Advantage of Joining the Longer Queue

Let $D \equiv D(m,p)$ be the additional delay experienced by an arrival if that arrival joins the shorter queue instead of the longer queue when the difference between the two queues is found to differ by two or more. We will establish a positive limit as $p \downarrow 0$.

An important role is played by the *equilibrium lifetime* (EL) distribution associated with the uniform distribution. For a nonnegative random variable X with cumulative distribution function (cdf) F having finite mean EX and probability density function (pdf) f, let X_L be a random variable with the associated EL cdf associated with F, having cdf F_L and pdf f_L, where

$$F_L(x) = \frac{1}{EX}\int_0^t x\, dF(x) \quad \text{and} \quad f_L(x) \equiv \frac{xf(x)}{EX};$$

e.g., see §3.4 of [**?**] and §5 of [**?**]. Let U_L have the EL cdf associated with U, where $U \equiv U(0,1)$ is uniformly distributed on $[0,1]$; it has cdf and mean:

$$P(U_L \le t) = t^2, \quad 0 \le t \le 1, \quad \text{and} \quad E[U_L] = \frac{2}{3}. \tag{7.1}$$

Theorem 7.1 (quantifying the advantage) *For the two-server queueing system with Poisson arrival process with arrival rate λ and the special two-point service-time distribution with parameter pair (m,p) specified above,*

$$D(m,p) \Rightarrow mU_L \quad \textit{and} \quad E[D(m,p)] \to \frac{2m}{3} \quad \textit{as} \quad p \downarrow 0$$

for U_L in (7.1).

The proof is based on two lemmas. Let $R \equiv R(m,p)$ be the remaining service time of the server to finish first after an epoch when the two servers are first both busy.

Lemma 7.1 (the remaining service time when the two servers are first both busy) *In the setting of Theorem* 7.1*,*

$$R(m,p) \Rightarrow mU \quad \textit{as} \quad p \downarrow 0,$$

where U is uniformly distributed on $[0,1]$.

Proof of Lemma 7.1. We start by considering consecutive epochs at which one of the queues first becomes busy after the system has been empty. The arriving customer with service time m will be in the system for time m. Since p is very small, most of the time no customers with service

time m will arrive during this interval, so that the system will become empty again. Thus the system will alternate between intervals where there is one customer present (called busy periods) and intervals where the system is empty (called idle periods). These busy and idle periods will be repeated until eventually there are first two customers in the system.

To determine the distribution of R, we can look at the arrival process only during the busy periods, not counting the arrivals that initiated the busy periods. Thus we are counting the *additional* arrivals with service time m that make the number in the system be at least 2. Let $A_m(t)$ count the number of these additional arrivals with service time m during the first t units of busy time. The key mathematical property (by Poisson thinning) is that the stochastic process $\{A_m(t) : t \geq 0\}$ is a Poisson process with rate λp. If we only look during the busy time, then the time until the first additional customer with service time m arrives, say $T \equiv T(\lambda, m, p)$ is exponentially distributed with mean $1/\lambda p$. Let k be such that $km \leq T < (k+1)m$. Then $R = m(k+1) - T$. As a consequence, for $0 \leq t \leq 1$,

$$\begin{aligned} P(R(m,p) > tm) &= \sum_{k=0}^{\infty} \left(e^{-\lambda pm(k+t)} - e^{-\lambda pm(k+1)} \right) = \frac{e^{-\lambda pmt} - e^{-\lambda pm}}{1 - e^{-\lambda pm}} \\ &= \frac{\lambda pm(1-t) + o(p)}{\lambda pm + o(p)} = (1-t) + o(p) \quad \text{as} \quad p \downarrow 0, \end{aligned}$$

so that, as $p \downarrow 0$, asymptotically R is uniformly distributed over the interval $[0, m]$. ∎

Now let $V \equiv V(m,p)$ be the remaining service time of the server to finish first after an epoch when both queues are positive with a difference of at least 2. Of course, V will coincide with $m - R$, where the R is the remaining time during which one of the customers in that queue has a positive service time, bringing the difference to at least 2. However, V is not simply distributed as $m - R$, because the queue sizes prior to this epoch will depend upon R. Thus, if R is larger, then queues will tend to be larger, which will increase the likelihood that one of these customers has a positive service time. However, the queue size will tend to be directly proportional to R. That enables us to determine the distribution of V as we choose p sufficiently small.

Lemma 7.2 (the remaining service time when the queue difference first is at least two) *In the setting of Theorem* 7.1*,*

$$V(m,p) \Rightarrow m(1 - U_L) \quad \textit{as} \quad p \downarrow 0,$$

for U_L *in* (7.1)*, so that* $E[V(m,p) \to 1/3$*.*

Proof of Lemma 7.2. The proof can parallel the proof of Lemma 7.1. A key step is the initial construction of the random variable $V \equiv V(m,p)$. We recognize that, for sufficiently small p, the system will go through many cycles before a third customer has a positive service time, causing the queues to differ by at least two. Now it suffices to consider the renewal process with i.i.d. interrenewal times X_j distributed as R. Let the associated partial sums and counting process be $S_n \equiv X_1 + \cdots + X_n$, $n \geq 1$, $S_0 \equiv 0$ and $N(t) \equiv \max\{t \geq 0 : S_n \leq t\}$, $t \geq 0$. Let the lifetime of the renewal process at time t be

$$L(t) \equiv S_{N(t)+1} - S_{N(t)}, \quad t \geq 0.$$

As in the proof of Lemma 7.1, let $T \equiv T(\lambda, p)$ be an exponential random variable with mean $1/\lambda p$. We observe that $V(m,p)$ is distributed as $L(T)$.

We next observe that, because of the Poisson arrival process, the distribution of $R(m,p)$ for all m and p and in the limit as $p \downarrow 0$ is nonlattice. Thus we have the familiar limit

$$L(t) \Rightarrow R_L \quad \text{as} \quad t \to \infty$$

for all m and p; e.g., see §3.4 of [**?**]. We also can immediately apply Lemma 7.1 to deduce that $R_L(m,p) \Rightarrow mU_L$ as $p \downarrow 0$, because the random variables $R(m,p)$ all have support on the finite interval m. For such random variables, the lifetime operator is continuous: If $X_n \Rightarrow X$, then $X_{L,n} \Rightarrow X_L$.

We then can combine the results to conclude that, as $p \downarrow 0$, $pT(\lambda,p) \Rightarrow Z(\lambda)$, where $Z(\lambda)$ is an exponential random variable with mean $1/\lambda$, so that $T(\lambda,p) \Rightarrow \infty$. We then deduce that

$$V(m,p) \stackrel{\mathrm{d}}{=} L^{(m,p)}(T(\lambda,p)) \Rightarrow U_L \quad \text{as} \quad p \downarrow 0. \quad \blacksquare$$

Proof of Theorem 7.1. Let time 0 be the time when the queues first differ by at least 2. The delay experienced by new arrivals after time 0 depends on which queue they join. If p is sufficiently small, then all customers that join the longer queue will leave the system at time $V(m,p)$, while all customers that join the shorter queue will leave the system at time m. However, the delays experienced by each successive arrival that joins a specified queue will be less, because time passes between successive arrivals. Nevertheless, during the time interval $[0, V(m,p)]$, the difference between the delay experienced by each arrival that elects to join the shorter queue instead of the longer queue, which is the delay advantage experienced by this customer, is always

$$D(m,p) = m - V(m,p)$$

Hence the conclusion of Theorem 7.1 follows from Lemma 7.2. ∎

Remark 7.1 (a difference in the mean steady-state delay) Theorem 7.1 can be applied to deduce that there will be a positive difference in the mean steady-state delay for *all* arrivals if we let $m \uparrow \infty$ as $p \downarrow 0$, provided that we let $mp^3 \to c > 0$ and $mp_4 \to 0$. If we let $mp^3 \to \infty$, then the advantage will grow without bound.