Analysis and Comparison of Queues with Different Levels of Delay Information

Pengfei Guo • Paul Zipkin

Fuqua School of Business, Duke University, Durham, NC 27708, USA

Information flow between customers and providers is an important element of most service systems. Developments in technology and managerial practice can enhance this information flow. In particular, in many situations, it is now possible for the provider to acquire and convey to customers fairly accurate information about anticipated delays due to congestion. Such information can directly affect customer satisfaction and also influence customers’ behavior. In our work, we focus on the arriving customer’s decision to wait for service or to leave (balk).

Information can take many forms, with different degrees of precision. Different levels of information have different effects on customers’ decisions and thus the overall arrival process. We consider three typical types of delay information: none, the system occupancy, and the exact waiting time. With no information, customers still estimate their waiting times, but these estimates are based only on long-term (equilibrium) experience, not real-time information. The occupancy provides partial information; the remaining uncertainty comprises the actual service times of the waiting customers. The exact waiting time gives the customer full information.

The Models

We assume Poisson arrivals and independent service times. There is a single server using the FCFS discipline.

To assess the effects of information consistently, we posit a customer-decision mechanism common to all levels of information. A given function, \( c(w) \), measures the basic cost of delay. Different customers, however, value time differently. Each customer arrives with a specific parameter \( \theta \), which scales the basic cost function. Upon arrival the customer receives information, which affects his estimate of the distribution of delay. Based on his scale parameter and the information, the customer computes his expected delay cost. If that is more than the reward, \( r \), which he anticipates from receiving service, he balks, and if not, he stays. In summary, customers’ expected utility function is \( U = r - \theta E[c(W)] \). In this scheme, then, different levels of information lead to different delay distributions in the expected-cost calculations, and those in turn affect everything else.

Related Literature

An overview of customer psychology in waiting situations, including the impact of uncertainty, can be found in Maister (1984). See also Carmon et al. (1995). There is some empirical evidence about customers’ reactions to delay information. Taylor (1994) showed that delays affect customers’ service evaluations in an experiment involving airline flights.
Hui and Tse (1996) conducted a survey on the relationship between information and customer satisfaction.

There is of course a substantial literature on queues with impatient customers. Models with balking and reneging (leaving after waiting for some time) can be found in many books, e.g. Kulkarni (1995). Recent works on this topic include Bae, Kim and Lee (2001), Zohar et al. (2002), and Ward and Glynn (2003).

The literature on customers influenced by delay information begins with Naor (1969), who studies a system like ours with partial information, but with identical customers and linear waiting cost. The focus here is the role of prices in changing customer behavior. Additional research along these lines includes Yechiali (1971), Knudsen (1972), Edelson and Hildebrandt (1975), Schroeter (1982), and Hassin (1986).

More recently, Shimkin and Mandelbaum (2004) discussed equilibria in a setting similar to our no-information model. Whitt (1999) studied two $M/M/s/r$ systems, corresponding roughly to our models with no and partial information. His customer-choice mechanism, however, is quite different from ours, and so are his findings about the impact of information on performance. His partial-information system’s occupancy is (first-order) stochastically less than his no-information system’s. Its throughput, therefore, is also less. As noted previously, we find instead a second-order relation and higher throughput. Thus, despite their similar motivation, his models represent very different behavior from ours.

Armony and Maglaras (2004a, 2004b) analyzed systems where an arriving customer learns some delay information and then can choose to balk, wait, or leave a message, in which case the provider calls back within a guaranteed time. That guaranteed time is an estimate based on a heavy-traffic approximation. In that heavy-traffic regime, indeed, such estimates become nearly precise. In Armony and Maglaras (2004a) customers’ choices are based on the equilibrium waiting time, as in our no-information system. Customers employ a utility function to assess delays, but the function’s argument is the expected delay, a constant. This approach is justified in the heavy-traffic limit, but it thereby suppresses the risk-reduction role of information. In Armony and Maglaras (2004b) each arriving customer receives more information, a point estimate of delay based on the system occupancy. The customer treats the estimate as exact, again based on the heavy-traffic limit. Comparing the results with the other system, they show that more information improves performance on several dimensions. Thus, while their findings are broadly consistent with ours, their modeling approach is quite different.

Our Results

We show how to compute the primary determinants of performance for each of the three levels of information. For certain cases (e.g., exponential processing times and linear basic cost) the results are in closed form.

We also obtain qualitative results about the impact of information. It is common to distinguish between individual and social optimization (e.g., Hassin and Haviv 2003). Social optimization aims to achieve the best outcome for everyone. It usually requires some central coordination scheme, such as admissions control. Our model represents individual decisions. Each customer maximizes only his own benefit. This control mechanism need not maximize average customer utility, because customers do not consider the effects of their own service times on others’ delays.
Under social optimization, more information is always better than less (absent information-processing costs), regardless of the objective. The controller can choose to ignore additional information, so the less-information solution is feasible for the more-information case. Here, the matter is less obvious.

We show that, in this setting, the objectives of the provider and the customers are perfectly aligned. A single measure, namely throughput, is proportional to both objectives. We find, moreover, that for exponential service times the impact of information on throughput is positive – more information leads to better performance.

Conclusions and Subsequent Works

In this work, we considered service systems with three levels of customer-delay information. Customers use that information to determine their expected waiting costs, and so to decide whether to stay and receive service or leave (balk). We obtained closed-form solutions for some cases and nearly closed-form solutions for others. We showed that, for delay-cost functions of a certain form, the only important criterion is system throughput; this measure reflects both the system’s average profit and customers’ average utility. We demonstrated that, for exponential service times, more precise information improves performance. Numerical examples indicated that this effect is stronger when customers are more sensitive to risk.

This utility-based approach forces us to revise our notions of good performance. Most peculiar is the concept that customers actually prefer one system to another when its probability of delay is larger. Of course, this does not mean that the customers want to wait. Rather, high utilization is an indication that, overall, customers are better satisfied. Also, according to this approach, the better system may have lower average occupancy and so lower average waiting time, but not always; in some cases those measures are higher. Finally, more information and better performance are associated with second-order stochastic dominance relations between state variables.

Numerous extensions are worth pursuing. (1) Does the positive effect of information on performance remain valid for general service times? For heavy-tailed service-time distributions? For multi-server systems? (2) The analysis assumes a delay cost of form $\theta c(w)$, so customers differ only by the uniformly-distributed scale factor $\theta$. What happens when the cost is $h(\theta)c(w)$ for some function $h$? Or, more generally, $c(\theta,w)$? (3) We focus here on the FCFS discipline. Other disciplines, such as fixed priorities, can improve system performance according to conventional measures. How do such disciplines fare in a utility-based approach? (4) It would be interesting to extend the analysis to a make-to-stock system, which employs inventory to reduce delays. (5) Finally, there are many other types and levels of information besides the three we examine here. It would be useful and interesting to model them and explore their effects.

References


