DECISION SCIENCES
A JOURNAL OF THE DECISION SCIENCES INSTITUTE

Decision Sciences
Volume 50 Number 3
June 2019

The Value of the Customer's Waiting Time for General Queues*

Kalyan Singhal and Jaya Singhal
Merrick School of Business, University of Baltimore, Baltimore, MD, 21201,
e-mail: Ksinghal@ubalt.edu, Jsinghal@ubalt.edu

Subodha Kumar†
Fox School of Business, Temple University, Philadelphia, PA, 19122,
e-mail: subodha@temple.edu

ABSTRACT

We have developed a measure of the value of the customer's waiting time that is applicable to all queuing systems. Since the birth of the modern queuing theory over 100 years ago, this measure is the first addition to the list of the measures of performance of general queues that includes the servers' utilization factor, the expected queue length, the expected waiting time, and some variations of the last two. The curves for trade-offs between the servers' utilization factor and the value of the customer's time can be used to supplement or replace the curves for trade-offs between the servers' utilization factor and the customer's expected queue length (or expected waiting time) that have been a fundamental part of the modern queuing theory. The decisions made with the value of the customer's waiting time will mirror the decision maker's goals more closely than the decisions made with the customer's expected queue length or expected waiting time that are surrogates for the value of the customer's waiting time. Although our definition of the value of the customer's time is deceptively simple, its implications can be significant and far reaching. It could change the way we pursue research in the queuing theory, the way we teach the queuing theory, and the way we design queuing systems in practice.
[Submitted: July 15, 2017. Revised: June 3, 2018. Accepted: September 12, 2018.]

Subject Areas: Optimization of Queues, Queuing Theory, Value of the Customer's Time, and and Waiting Lines.

*We want to thank the anonymous associate editor and the two referees for their insightful comments on the preceding version of the article. We also want to thank Gabriel Bitran, Rachel Chen, Morris Cohen, Brian Denton, Laurens Debo, Craig Froehle, Michael Fu, Steve Graves, Rafi Hassin, Fred Hillier, Wallace Hopp, Seyed Iravani, Costis Maglaras, Michael Pinedo, Suresh Radhakrishnan, Sergei Savin, Sridhar Seshadri, George Shanthikumar, Max Shen, Shaler Stidham, Ward Whitt, and David Yao for their comments and insights on earlier versions of the article.

†Corresponding author.

INTRODUCTION

The Role and Importance of Waiting Time

The role and importance of waiting time have been widely discussed in the economics, marketing, and operations literature. Most service operations, including banks, call centers, fast-food drive-through restaurants, health care facilities, and retail stores, compete based on waiting time, price, and quality attributes. Hopp (2008, p. 79) emphasizes that "waiting remains a major source of inefficiency in many public- and private-sector systems."

Tong and Rajagopalan (2014, p. 689) also emphasized the importance of service time: "In many services, for example, website or landscape design, the value or quality derived by a customer depends upon the service time, and this valuation differs across customers. Customers procure the service based on the expected value to be delivered, prices charged, and the timeliness of service." Service time and the resulting customer patience play a critical role in the design of queuing systems. Wang, Lan, and Jiang (2016) "explore the impact of customer impatience on the performance of a production service system that consists of one production inventory subsystem and one service subsystem." Zacharias and Pinedo (2017, p. 639) analyzed a discrete multiserver model for scheduling customer arrivals under no-shows where they "assign customers to time slots so that the service system utilizes its resources efficiently and customers experience short waiting times."

Researchers have extensively documented that for waiting time in queues, time is indeed money. Furthermore, although money can be transferred or exchanged, time saved and time lost cannot (Leclerc, Schmitt, & Dubé, 1995, p. 119). Becker (1965, p. 494) emphasized the value of time observing that "the full costs of" obtaining services "would equals the sum of market prices and the foregone value of the time used up." The Food Marketing Institute (1985, 1986), in its annual updates, reported that the consumers were willing to pay a higher price if they had to wait for less time in checkout lines. Waiting time has the same effect as price on consumer choices and on market share (Deacon & Sonstelie, 1985; Siferd, Benton, & Ritzman, 1992; Allon, Federgruen, & Pierson, 2011). A maxim in the fast-food drive-through industry is that for "every seven-second reduction in total service time, sales will increase by 1% over time" (Hughlett, 2008). Allon et al. (2011, p. 503) empirically validated that this maxim was "on average" correct.

The value of the customer's waiting time depends on the context of waiting and on the degree to which waiting is pleasurable (Kahneman & Tversky, 1984; Larson, 1987; Maister, 2005). For example, Allon et al. (2011, p. 501) found that the value for the customer of waiting time in the fast-food drive-through industry was at least three times the value of time spent driving to the restaurant.

Long delays can impose a cost on the service provider because its customers can decide not to seek services from the service provider in the future, or even if they do seek the services, they can renege the queue or balk. Because these decisions are driven by the customers' perception of the cost of their time, this cost is an important parameter in designing queueing systems.

The Current State of the Art in Analyses of Waiting Lines and Its Limitations

Most analyses and evaluations of queuing systems include computations of the utilization factor (expected fraction of time the servers are busy), the expected queue length, the expected waiting time, and some variations of the last two. These measures are used so widely that they are the basis of at least one publicly available online calculator for *M/M/s* queueing models.[i]

To determine the design parameters of a queuing system, decision-makers compare available alternatives based on these measures and make a trade-off between the servers' utilization factor and the expected queue length (or the expected waiting time). Although they generally know the cost of the servers' utilization factor, they treat expected queue length and expected waiting time as surrogates for the cost of the customer's waiting time which varies from customer to customer. Another approach for determining the design parameters of a queuing system is to minimize the sum of the costs of the servers' time and the estimated cost of the customers' time (Gross & Harris, 1998, p. 9; Hillier & Lieberman, 2005, pp. 813–818). Although the cost of the servers' time is easy to estimate, the cost of the customer's waiting time is not easy to estimate partly because different customers have different costs of their waiting times.

Organization of This Article

In "The Value of the Customer's Time" section, we define the value of the customer waiting time for general queues. In "The *M/M/s* Models" section, we analyze the value of the customer waiting time for the *M/M/s* models. In "Computations for the *M*/*M*/*s* Models" section, we provide computations for the *M/M/s* models and the curves for trade-offs between the servers' utilization and the value of the customer waiting time for nine different values of the number of servers. We also discuss their role in managerial decisions. In "A Numerical Example" section 5, we describe a numerical example. In "The Value of the Customer's Time and Optimization of Queues" section, we discuss optimization of queues and the relationship between the optimized values and the value of the customer waiting time. In the concluding section, we discuss implications of our work to practice research, and teaching.

THE VALUE OF THE CUSTOMER'S TIME

Notations

s = Number of servers.

L_{qs} = Average length of the queue of waiting customers. From the Little's Law, this length is equal to the product of customers' arrival rate multiplied by their average waiting time, and thus it also represents the mean waiting time for all customers per unit of time.

ρ = Server's utilization factor, $0 < \rho < 1$.

$S_s = (1 - \rho)s$ = Mean server idle time for all servers per unit of time.

If we compared two queues with unequal waiting times in the same system or two different systems, the one with the longer waiting time would imply that

[i] http://www.free-onlinecalculator.com/

the system has assigned a lower value of the customer's time to it than to the other queue. Thus, every queue has a value of the customer's waiting time. This is the underlying rationale for our definition. We define α, the value of the customer's waiting time, as the ratio of the marginal changes in S_s and the marginal changes in L_{qs} where S_s represents the mean server idle time for all servers per unit of time and L_{qs} represents the average length of the queue of waiting customers, and it is equal to the mean waiting time for all customers per unit of time:

$$\alpha = -\frac{dS_s}{dL_{qs}} = -\frac{\frac{dS_s}{d\rho}}{\frac{dL_{qs}}{d\rho}}.$$

The negative sign in the expression shows that an increase (decrease) in the denominator results in a decrease (increase) in the numerator. Substituting the value of $S_s = (1 - \rho)\, s$ in the equation above, we get

$$\alpha = -\frac{\frac{dS_s}{d\rho}}{\frac{dL_{qs}}{d\rho}} = -\frac{\frac{d[(1-\rho)s]}{d\rho}}{\frac{dL_{qs}}{d\rho}} = s\left[\frac{dL_{qs}}{d\rho}\right]^{-1}. \tag{1}$$

Equation (1) will hold if L_{qs} is differentiable with respect to ρ. Theoretically, the empirical variant of (1) given below will also hold if the function for L_{qs} was empirical and not differentiable.

$$\alpha = s\left[\frac{\Delta L_{qs}}{\Delta \rho}\right]^{-1}. \tag{2}$$

The definition in (1) holds for all possible designs of any queuing system. In both (1) and (2), α gives the value of the customer's waiting time as a multiple of the cost of the server's idle time. A numerical value equal to α for the value of the customer's waiting time in a queuing system means that the system is designed on the assumption that the value of the customer's waiting time is α times the cost of the server's idle time.

The *value of the customer's waiting time* and the *cost of the customer's waiting time* are two different concepts. The former is an attribute or a characteristic (or a measure of performance) of a queuing system that is implicit in its design, and the latter is the opportunity cost of each customer's time, regardless of who estimates it. The value of the customer's waiting time for a queue will be the same for each customer, but the cost of the customer's waiting time will vary from the customer to customer. If there is a change in the number of servers or in the service rate per server, the value of the customer's waiting time assigned by the system will change in the same direction, but the cost of each customer's waiting time will remain unchanged. Conversely, if we replace a customer population that has a low average cost of waiting time with a customer population that has a high average cost of waiting time, the value of the customer's waiting time assigned by the system will remain unchanged.

One can use (1) to compute the value of the customer's waiting time for a range of options for the design of a queue for further evaluation, and thus one can generate curves for trade-offs between the value of the customer's time and the utilization factor ρ for a set of values of s, the number of servers. These curves will

be analogous to the curves for trade-offs between the servers' utilization factor and the customer's expected queue length (or expected waiting time). The decisions made with the value of the customer's time should mirror the decision maker's goals more closely than the decisions made with the surrogates of the value of the customer's time: the customer's expected queue length or their expected waiting time.

Robinson and Chen (2011) developed a method for estimating the value of the customer's time for an optimal transient the $GI/G/m$ queue in appointment scheduling. Their model is a special case of ours, and its results can be obtained from our model. Their other contribution is in finding the upper and the lower bounds on the value of α for the $GI/G/m$ queues because several parameters in the expression for α in the $GI/G/m$ queues are unknown.

THE *M/M/s* MODELS

We derive an expression for the value of customer's waiting time for the *M/M/s* models that are the most elementary Markovian birth–death queuing models widely deployed in practice. The arrival process is Poisson, and each of the s servers is independent, and they have identical exponential service-time distributions. These models have been widely discussed in the literature (Sobel, 1969; Purdue, 1974; Levhari & Luski, 1978; Pegden & Rosenshine, 1987; L'Ecuyer, Giroux, & Glynn, 1994; Hopp, 2008).

Hillier and Lieberman (2005, p. 788) provide the following expression for L_{qs} for the *M/M/s* queues:

$$L_{qs} = \frac{\left(\frac{\lambda}{\mu}\right)^s \rho}{s!(1-\rho)^2 \left\{ \left[\sum_{n=0}^{s-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \right] + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!} \frac{1}{1-(\lambda/s\mu)} \right\}}.$$

Substituting $(\lambda/\mu) = s\rho$,

$$\begin{aligned} L_{qs} &= \frac{(s\rho)^s \rho}{s!(1-\rho)^2 \left\{ \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} \right] + \frac{(s\rho)^s}{s!} \frac{1}{1-\rho} \right\}} \text{ or} \\ L_{qs} &= \frac{(s\rho)^s \rho}{s!(1-\rho)^2 \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} \right] + (1-\rho)(s\rho)^s}. \end{aligned} \tag{3}$$

The Value of the Customer's Waiting Time for the M/M/s Queues

Theorem 1: *'M/M/s queues,*

$$\alpha = \frac{s\left\{ s!(1-\rho)^2 \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} \right] + (1-\rho)(s\rho)^s \right\}^2}{(s\rho)^s \left\{ (1-\rho)s! \left[s+1-(2s-1)\rho+s\rho^2 \right] \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} \right] + (s\rho)^s \left[s+1-2s\rho+s\rho^2 \right] \right\}}$$

Proof

Differentiating (3),

$$\frac{dL_{qs}}{d\rho} = \frac{\left[ss(s\rho)^{s-1}\rho+(s\rho)^s\right]\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2} - \frac{(s\rho)^s\rho\left\{s!2(\rho-1)\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+s!(1-\rho)^2s\left[\sum_{n=0}^{s-2}\frac{(s\rho)^n}{n!}\right]+ss(s\rho)^{s-1}(1-\rho)-(s\rho)^s\right\}}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2}. \tag{4}$$

We simplify only the numerators in the two parts of (4). We first simplify the first square bracket in the first part to get

$$\frac{dL_{qs}}{d\rho} = \frac{(s+1)(s\rho)^s\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2} - \frac{(s\rho)^s\rho\left\{s!2(\rho-1)\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+s!(1-\rho)^2s\left[\sum_{n=0}^{s-2}\frac{(s\rho)^n}{n!}\right]+s^2(s\rho)^{s-1}(1-\rho)-(s\rho)^s\right\}}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2}.$$

We combine the two parts by taking out the common denominator and $(s\rho)^s$ and then combining similar expressions in the remainders of the two parts:

$$\frac{dL_{qs}}{d\rho} = \frac{(s\rho)^s}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2} \times \left\{\begin{array}{l}\left[(s+1)\ s!(1-\rho)^2+s!2\rho(1-\rho)\right]\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]-s!(1-\rho)^2s\rho\left[\sum_{n=0}^{s-2}\frac{(s\rho)^n}{n!}\right]+ \\ (s+1)(1-\rho)(s\rho)^s-\rho s^2(s\rho)^{s-1}(1-\rho)+\rho(s\rho)^s\end{array}\right\}.$$

We combine and simplify the last three expressions in the numerator:

$$\frac{dL_{qs}}{d\rho} = \frac{(s\rho)^s}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2} \times \left\{(1-\rho)s![(s+1)(1-\rho)+2\rho]\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]-s!(1-\rho)^2s\rho\left[\sum_{n=0}^{s-2}\frac{(s\rho)^n}{n!}\right]+(s\rho)^s\right\}.$$

We add and subtract $s!(1-\rho)^2s\rho[\frac{(s\rho)^{s-1}}{(s-1)!}]$ to the expression after the multiplication sign:

$$\frac{dL_{qs}}{d\rho} = \frac{(s\rho)^s}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2} \times \left\{\begin{array}{l}(1-\rho)s![(s+1)(1-\rho)+2\rho]\left[\sum_{n=0}^{s-1}\frac{(s)^n}{n!}\right]-s!(1-\rho)^2s\rho\left[\sum_{n=0}^{s-2}\frac{(s\rho)^n}{n!}\right]- \\ s!(1-\rho)^2s\rho\left[\frac{(s\rho)^{s-1}}{(s-1)!}\right]+s!(1-\rho)^2s\rho\left[\frac{(s\rho)^{s-1}}{(s-1)!}\right]+(s\rho)^s\end{array}\right\}.$$

We can combine the expression with the summation from $n = 0$ to $n = (s-2)$ with the expression that follows it so that the resulting expression will have summation from $n = 0$ to $n = (s-1)$:

$$\frac{dL_{qs}}{d\rho} = \frac{(s\rho)^s}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right] + (1-\rho)(s\rho)^s\right\}^2} \times \left\{\begin{array}{l}(1-\rho)s![(s+1)(1-\rho)+2\rho]\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right] - \\ s!(1-\rho)^2 s\rho\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right] + s!(1-\rho)^2 s\rho\left[\frac{(s\rho)^{s-1}}{(s-1)!}\right] + (s\rho)^s\end{array}\right\}.$$

We can combine the first two expressions in the curly bracket, and then combine the remaining two expressions in the curly bracket:

$$\frac{dL_{qs}}{d\rho} = \frac{(s\rho)^s\left\{(1-\rho)s!\left[s+1-(2s-1)\rho+s\rho^2\right]\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(s\rho)^s\left[s+1-2s\rho+s\rho^2\right]\right\}}{\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2}. \quad (5)$$

From (1),

$$\alpha = s\left[\frac{dL_{qs}}{d\rho}\right]^{-1}$$

Substituting the value of $dL_{qs}/d\rho$ from (5) into the equation above,

$$\alpha = \frac{s\left\{s!(1-\rho)^2\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(1-\rho)(s\rho)^s\right\}^2}{(s\rho)^s\left\{(1-\rho)s!\left[s+1-(2s-1)\rho+s\rho^2\right]\left[\sum_{n=0}^{s-1}\frac{(s\rho)^n}{n!}\right]+(s\rho)^s\left[s+1-2s\rho+s\rho^2\right]\right\}}. \text{ QED.}$$

COMPUTATIONS FOR THE *M/M/s* MODELS

Table 1 contains computations for the relationship between the utilization factor, ρ, and the value of the customer's time, α, for 13 different values of s.

Figure 1 is a plot of the computations in Table 1 for nine values of $s = 1, 2, 3, 5, 7, 10, 15, 20,$ and 25, and it shows how α changes with ρ for various values of s. As in the case of trade-offs between the server's utilization and the average queue length (or average waiting time), both Table 1 and Figure 1 show the economies of scale with the increasing number of servers. With these curves for trade-offs between ρ, the utilization factor, and α, the value of the customer's waiting time as a multiple of the cost of the server's idle time, the decision maker can easily converge on a desirable value of α.

Benchmark Utilization Factor and Iso-Customer–Value Curves

We designate $\alpha = 1$ *the benchmark* for the values of α, denote by ρ_{sb} the value of ρ for which $\alpha = 1$, and call this value of ρ the *benchmark utilization factor*. Table 2 contains the values of ρ_{sb} for 13 different values of s when $\alpha = 1$, that is, when the value of the customer's time is equal to the value of the server's time. For $\alpha = 1$, each value of ρ_{sb} in Table 2 represents utilization for specified values of s. As one would expect, with the economies of scale from the increasing number of servers, the utilization of servers increases asymptotically toward 1 as the number

Table 1: The relationship between the utilization factor, ρ, and the value of the customer's time, α.

	α												
ρ	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$	$s=10$	$s=15$	$s=20$	$s=25$
0.1	4.263	32.779	185.88	939.73	4,481	20,636	92,893	411,513	1,801,580	7,816,228	11,045,245,092	14,448,577,382,920	18,161,483,664,738,300
0.2	1.778	7.784	24.997	71.100	189.85	488.03	122	3,013	7,327	17,640	1,301,672	88,363,852	5,748,433,381
0.3	0.961	3.162	7.636	16.283	32.49	62.26	116.09	212.34	382.89	682.94	11,143	165,908	2,358,400
0.4	0.563	1.553	3.157	5.663	9.49	15.26	23.82	36.45	54.91	81.74	534.70	3159.9	17,740
0.5	0.333	0.818	1.486	2.383	3.571	5.128	7.150	9.760	13.11	17.389	63.32	206.23	634.15
0.6	0.190	0.431	0.725	1.078	1.500	2.000	2.589	3.280	4.089	5.031	12.530	27.707	57.524
0.7	0.099	0.211	0.337	0.477	0.632	0.802	0.988	1.193	1.416	1.660	3.241	5.602	9.063
0.8	0.042	0.086	0.132	0.181	0.232	0.285	0.341	0.399	0.459	0.522	0.876	1.304	1.817
0.9	0.010	0.020	0.031	0.041	0.052	0.063	0.074	0.085	0.096	0.107	0.165	0.226	0.291

Figure 1: The relationship between the utilization factor, ρ, and the value of the customer's time, α for various values of s.

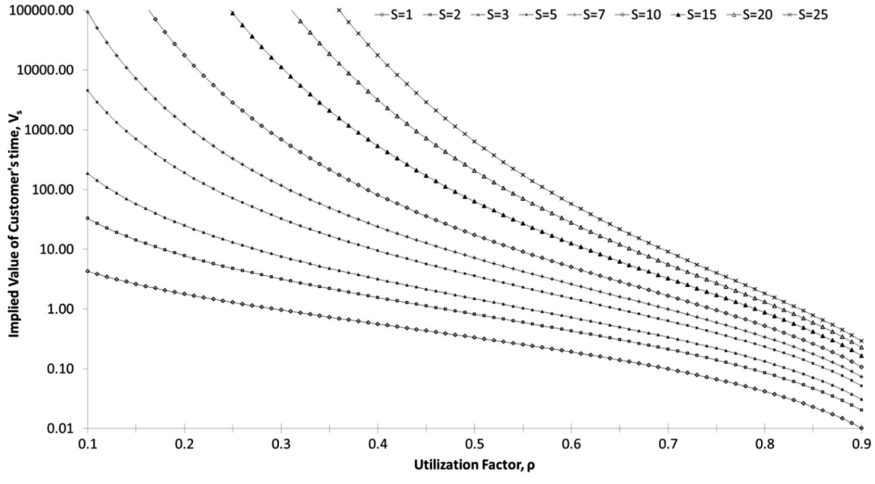


Table 2: The relationship between the number of servers s and the benchmark utilization factor ρ_{sb} for $\alpha = 1$.

s	1	2	3	4	5	6	7
ρ_{sb}	0.2929	0.4682	0.5552	0.6095	0.6476	0.6761	0.6988
s	8	9	10	15	20	25	
ρ_{sb}	0.7171	0.7326	0.7455	0.7908	0.8178	0.8361	

of servers approaches ∞. Figure 2 is a plot of Table 2 where we have connected the various values of ρ_{sb}. We call this piecewise linear function the *iso-customer–value curve* for $\alpha = 1$. One can generate a set of *iso-customer–value curves* for other values of α. The higher the value of α, the lower will be the location of the curve and vice versa, but all curves will increase asymptotically toward one as the number of servers approaches ∞.

The Value of the Customer's Waiting Time for the M/M/1 Queue

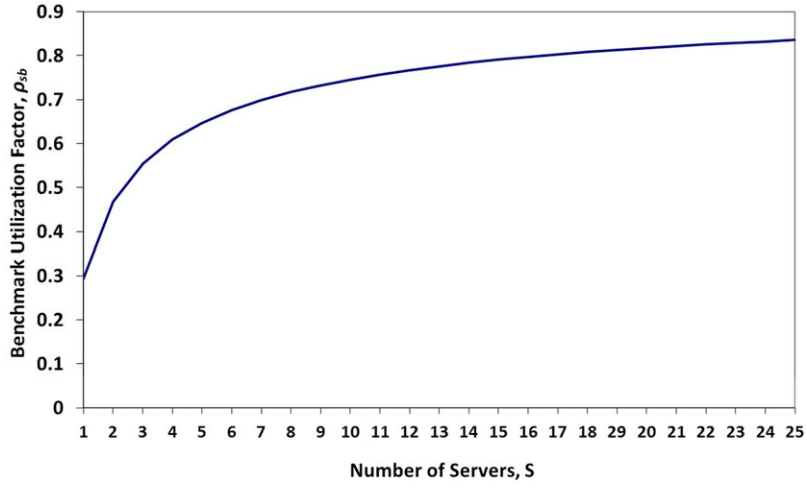
The *M/M/1* queue is the "classical Poisson-input, exponential-service, single-server queue" (Gross & Harris, 1998, p. 53). It is widely deployed in practice, and a visible example is the waiting line at fast-food drive-through restaurants.

The average number of customers waiting in the queue:

$$L_{q1} = \frac{\rho^2}{1-\rho}\,. \tag{6}$$

Thus, $\frac{dL_{q1}}{d\rho} = \frac{\rho(2-\rho)}{(1-\rho)^2}$.

Figure 2: An iso-customer–value curve for $\alpha = 1$: The relationship between the number of servers s and the benchmark utilization factor ρ_{sb} for $\alpha = 1$.



This result gives

$$\alpha = (s = 1) \left[\frac{dL_{q1}}{d\rho}\right]^{-1} = \frac{(1-\rho)^2}{\rho\,(2-\rho)}\,. \tag{7}$$

This expression is as simple as the expressions for the expected queue length or the expected waiting time in an *M/M/1* queue.

Managerial Decisions

Currently, the decision-makers compare available alternatives for the designs of a queue and make a trade-off between the servers' utilization factor and the expected queue length (or the expected waiting time), which they treat as surrogates for the value of the customer's time. The literature contains both sets of trade-off curves. The trade-off curves for the value of the customer's waiting time are the new set of curves which the managers can use in addition to or in place of the trade-off curves for the expected queue length or the expected waiting time. The decisions made with the value of the customer's waiting time will mirror the decision maker's goals more closely than the decisions made with the customer's expected queue length or expected waiting time, which are surrogates for the value of the customer's waiting time.

Table 3: Two scenarios for the County Hospital.

Measures of Performance	One Doctor	Two Doctors
ρ, the utilization factor	2/3	1/3
L_q, the expected queue length	4/3	1/12
L, the expected number of patients in the queuing system	2	3/4
W_q, the expected waiting time in the queue for each patient	40 minutes	2.5 minutes
W, the expected waiting time in the system for each patient	60 minutes	22.5 minutes
α, the value per unit of patients' waiting time as a multiple of the cost per unit of doctors' idle time	0.125	2.4615

A NUMERICAL EXAMPLE

The Value of the Patient's Time in an M/M/1 Queue and an M/M/2 Queue

Hillier and Lieberman (2005, pp. 766 and 790–791) provide an example of an emergency room of the "County Hospital" where emergency cases arrive at random and thus follow a Poisson input process, leading to an exponential distribution for interarrival times. The time a doctor spends treating emergency patients can be approximated by an exponential distribution. Currently, at any hour, the emergency room has one doctor, but, because of an increase in emergency cases, the hospital is exploring the possibility of increasing to two doctors at any hour. The average emergency-case arrival rate is two per hour, that is, $\lambda = 2$, and the average service rate is three patients per hour, that is, $\mu = 3$. Table 3 contains various measures of performance, including the value per unit of patients' waiting time as a multiple of the cost per unit of doctors' idle time in the two scenarios.

Insights from the numerical example: If one were to decide without considering the value of the customer time, one would use either W_q, the expected waiting time in the queue for each patient for a trade-off between W_q and utilization or L_q, the expected queue length for a trade-off between L_q and utilization. The values of W_q and L_q for one doctor are 16 times those for two doctors. However, if one were to use the value per unit of the patient's waiting time as a multiple of the cost per unit of a doctor's idle time, with the addition of another doctor, this value increases 19.7 times, from 0.125 to 2.4615. We focus on three things. First, although 16 and 19.7 differ by about 20%, their relative numerical values confirm that both are sensitive to the waiting time. Second, whereas the waiting time is only a surrogate for the dollar value of the customer's time, the value per unit of the patient's waiting time has real dollar value because it is a multiple of the cost per unit of a doctor's idle time that has a dollar value. Third, whereas the decision in this binary choice is likely to be the same under either criterion, this is not likely to be so when the range of choice is much wider and a decision with real money value would obviously meet the goals of the organization more effectively than a decision based on a surrogate criterion.

THE VALUE OF THE CUSTOMER'S TIME AND OPTIMIZATION OF QUEUES

Singhal, Singhal, and Kumar (2018) minimize the sum of the servers' idle time costs and the customers' waiting costs. They show that the value of the customer's waiting time, α, is equal to the ratio $\alpha_* = C_w / C_i$, that is, the ratio of the cost of customer's waiting time and the cost of each server's idle time when the sum of the costs of the servers' time and the estimated cost of the customers' time is minimized. However, cost minimization cannot be used for external customers because the customers' waiting costs vary from customer to customer and a service provider does not know what these costs are. In any case, a service provider would not want to minimize the sum of the two costs because the goal of a service provider is to minimize its costs or maximize its profits while taking into consideration the customers' demand function regarding their costs of waiting. Furthermore, the system cost function for the sum of the two costs is not always convex or even quasi-convex (unimodal).

However, one could minimize the sum of the two costs when the firm has internal customers and the cost function is convex or quasi-convex. Singhal et al. (2018) describe an application of minimization of the system costs for a machine shop that has internal customers.

CONCLUSIONS

A New Fundamental Measure of Performance

Since the birth of the modern queuing theory over 100 years ago, our measure of the value of the customer's waiting time is the first addition to the list of the measures of performance of general queues that includes the servers' utilization factor, the expected queue length, the expected waiting time, and some variations of the last two. Although the measures like the utilization of the server's time are related to the server and the expected waiting time and the expected queue length are related the customer, the value of the customer's waiting time as a multiple of the cost of the server's idle time is related to both.

A New Set of Trade-off Curves for Choosing Parameters of a Queuing System

The curves for trade-offs between the servers' utilization factor and the customer's expected queue length (or the expected waiting time) for determining the number of servers and for making investments in enhancing the service rate of each server have been a fundamental part of the modern queuing theory since its development over a century ago. Our curves for trade-offs between the value of the customer's waiting time and the utilization factor can be used to supplement or replace those curves. The decisions made with the value of the customer's waiting time will mirror the decision maker's goals more closely than the decisions made with the customer's expected queue length or expected waiting time, which are surrogates for the value of the customer's waiting time.

Research Opportunities

Researchers have an opportunity to explore the possibility of deriving expressions for the value of the customer's waiting time when they compute the customer's expected queue length (or expected waiting time) for any queuing system. As a first step, they have an opportunity to generate the curves for trade-offs between the value of the customer's time and the utilization factor ρ for a set of values of s, the number of servers for some of the basic queuing systems, similar to what we have done for the *M/M/s* queues.

Although our definition of the value of the customer's time is deceptively simple, its implications can be significant and far reaching. It could change the way we pursue research in the queuing theory, the way we teach the queuing theory, and the way we design queuing systems in practice.

REFERENCES

Allon, G., Federgruen, A., & Pierson, M. (2011). How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-through industry based on structural estimation methods. *Manufacturing & Service Operations Management*, *13*(4), 489–507.

Becker, G. (1965). A theory of allocation of time. *The Economic Journal*, *75*(3), 493–517.

Deacon, R. T., & Sonstelie, J. (1985). Rationing by waiting and the value of time: Results from a natural experiment. *Journal of Political Economy*, *93*(4), 627–647.

Food Marketing Institute. (1985). *Trends—consumer attitudes and the supermarket, 1985 update*. Washington, DC: Food Marketing Institute.

Food Marketing Institute. (1986). *Trends—consumer attitudes and the supermarket, 1986 update*. Washington, DC: Food Marketing Institute.

Gross, D., & Harris, C. M. (1998). *Fundamentals of queueing theory*. New York: Wiley.

Hillier, F. S., & Lieberman, G. J. (2005). *Introduction to operations research*, 8th ed. New York, NY: McGraw-Hill.

Hopp, W. J. (2008). Single server queueing models. In D. Chhajed & T. J. Lowe (Eds.), *Building intuition: Insights from basic operations management models and principles*. New York: Springer.

Hughlett, M. (2008). Drive-throughs done right ring up returns. *Chicago Tribune* (November 28). Retrieved from https://articles.chicagotribune.com/2008-11-28/news/0811270365_1_drive-through-restaurant-technologies-competitive-advantage

Kahneman, D., & Tversky, A. (1984). Choice, values, and frames. *American Psychologist*, *39*(4), 341–350.

Larson, R. C. (1987). Perspectives on queues: Social justice and the psychology of queueing. *Operations Research*, *35*(6), 895–905.

Leclerc, F., Schmitt, B. H., & Dubé, L. (1995). Waiting time and decision making: Is time like money? *Journal of Consumer Research*, *22*(1), 110–119.

L'Ecuyer, P., Giroux, N., & Glynn, P. W. (1994). Stochastic optimization by simulation: Numerical experiments with the M/M/1 queue in steady state. *Management Science*, *40*(10), 1245–1261.

Levhari, D., & Luski, I. (1978). Duopoly pricing and waiting lines. *European Economic Review*, *11*(1), 17–35.

Maister, D. H. (2005). The psychology of waiting lines. Retrieved from https://davidmaister.com/articles/the-psychology-of-waiting-lines/

Pegden, C. D., & Rosenshine, M. (1987). Some new results for the *M/M/1* queue. *Management Science*, *28*(7), 821–828.

Purdue, P. (1974). The M/M/1 queue in a Markovian environment. *Operations Research*, *22*(3), 562–569.

Robinson, L. W., & Chen, R. R. (2011). Estimating the implied value of the customer's waiting time. *Manufacturing and Service Operations Management*, *13*(1), 53–57.

Siferd, S., Benton, W., & Ritzman, L. (1992). Strategies for service systems. *European Journal of Operational Research*, *56*, 291–303.

Singhal, K., Singhal, J., & Kumar, S. (2018). The value of the customer's time and optimization of queues. Working paper, Temple University.

Sobel, M. (1969). Optimal average-cost policy for a queue with start-up and shut-down costs. *Operations Research*, *17*(1), 114–162.

Tong, C., & Rajagopalan, S. (2014). Pricing and operational performance in discretionary services. *Production and Operations Management*, *23*(4), 689–703.

Wang, K., Lan, S., & Jiang, Z. (2016). Impact of customer impatience on a production service system. *International Journal of Production Research*, *54*(9), 2731–2749.

Zacharias, C., & Pinedo, M. (2017). Managing customer arrivals in service systems with multiple identical servers. *Manufacturing & Service Operations Management*, *19*(4), 639–656.

Kalyan Singhal is Doris and Robert McCurdy professor of operations and supply-chain management at the Merrick School of Business, University of Baltimore. He is a fellow of INFORMS and a fellow of POMS. He founded the Production and Operations Management Society (POMS) in 1989. He also founded The Institute of Management Science's College on Production and Operations Management in 1987, which merged with the Operations Research Society of America's Special Interest Group in Manufacturing in 1994 to become the Manufacturing and Service Operations Management (MSOM) Society of INFORMS. He has published in the *European Journal of Operational Research, Harvard Business Review, INFORMS Journal on Computing, Interfaces, the International Journal of Production Research, the Journal of Operations Management, Management Science, Operations Research, Production and Operations Management,* and the *TIMS Studies in the Management Sciences.* He has served on the faculty of the University of Arizona,

the University of Houston, and the Indian Institute of Management, Bangalore; worked with Union Carbide; and consulted with Baltimore Gas & Electric Company, the Government of India, Standard Oil of California, and Westinghouse Electric Corporation.

Jaya Singhal holds the Frank Baker Chair for Research Excellence at the Merrick School of Business, University of Baltimore. She earned MSc in nuclear physics from Marathwada University and PhD in management science from the University of Arizona. She has published in *Management Science*, *INFORMS Journal on Computing*, *European Journal of Operational Research*, *Journal of Operations Management*, *Production and Operations Management*, *INFOR*, and *Omega*. She is also the lead author of the mathematical software, ZOOM (Zero-One Optimization Methods) that has been widely used for decisions involving hundreds of millions of dollars. In 2016, she was awarded the prestigious Elkins Professorship by the University System of Maryland.

Subodha Kumar is the Paul Anderson distinguished chair professor of supply chain, marketing, information systems, and statistical science at Temple University's Fox School of Business. He also serves as the director of the Center for Data Analytics. He has published several papers in reputed journals and refereed conferences. In addition, he has authored a book, and coauthored book chapters, Harvard Business School cases, and Ivey Business School cases. He also has a patent. He has featured on several media outlets including *NBC*, *CBS,* etc. He is the deputy editor and a department editor of *Production and Operations Management* (POM), and the deputy editor-in-chief of *Management and Business Review*. He has served as a senior editor of *Decision Sciences* (DSI) and an associate editor of *Information Systems Research*. Additionally, he serves on other editorial boards. He was the conference chair for POMS 2018 and DSI 2018, and has cochaired several other conferences. He has been keynote speaker and track/cluster chairs at leading conferences. He is an associate executive director of POMS, the web editor of *POMS*, and the vice president of INFORMS Information Systems Society. He serves on the Advisory Boards of *Insightzz* and the *Srini Raju Centre for IT and The Networked Economy* at the Indian School of Business. He has served on the faculty of University of Washington and Texas A&M University.