Reservation, a tool to reduce the balking effect and the probability of delay

Benjamin Legros

Reservation, a tool to reduce the balking effect and the probability of delay

Benjamin Legros

EM Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016, Paris, France

Abstract

We investigate a threshold reservation policy implemented within a single customers' class. From an explicit performance analysis, we prove that the potential of reservation is into the reduction of the balking effect together with a higher servers' utilization. However, it also may result in a higher expected waiting time and more abandonment. We conclude that reservation can be efficiently implemented in large systems, under high workload situations, with a low waiting aversion and a low impatience.

Keywords: queue; balking; reservation; threshold policy.

1 Introduction

In a system where all resources are occupied, the service of new arriving customers has to be delayed. The waiting aversion leads some customers to reconsider asking for service in case of a non-immediate service. This phenomenon is referred to as the *balking effect*. Balking can be significant in situations where customers know that they can easily find a similar service elsewhere, eventually without wait. For instance in a street with many restaurants, a significant proportion of customers may refuse waiting for an available table if they can find another one elsewhere. Another example is call centers; [4] show that the balking phenomenon results in a significant loss of customers.

Since the loss of customers is undesirable, it may be useful to develop strategies which allows the service provider to maintain an amount of resource availability if preemption is not possible. In a context where a unique group of servers has to serve urgent and non-urgent customers, it has been shown that a reservation strategy can significantly improve the service level of urgent customers. The idea of a reservation strategy is to force some servers to remain available for new arriving urgent customers even if some non-urgent customers are waiting. This strategy is sometimes implemented in health care systems (see [6]) and in call centers with inbound and outbound calls (see [9]).

An important stream of literature is devoted to the analysis of such reservation strategies with two customer's classes. These strategies are referred to as *blending policies*. Some papers focus on performance evaluation, and others address the analysis of blending policies or staffing decisions. [5] develop various continuous Markov chain models for a call center with inbound and outbound calls. The authors consider a threshold policy and characterize the rate of outbound calls and the waiting time distribution of inbound calls. Other papers address the analysis of reservation policies. [7] and [2] prove that a threshold policy on

the number of idle agents is optimal to maximize the outbound throughput under a service level constraint on the inbound waiting time. Similar results are also found in [13], for a non-stationary model where inbound calls arrive according to a non-homogeneous Poisson process or in [14] in a setting with a callback option. [15] consider a large call center with a reservation model and propose a logarithmic safety staffing rule, combined with a threshold control policy to ensure that agents' utilization is always close to one with always idle agents present.

In a broaden perspective, reservation is the idea that some resources should remain idle when some work is available. For instance in a context with heterogeneous servers and the objective to minimize the time spent in the system, it is optimal to consider non-workconserving policies where slow servers are used only if the queue length exceeds some given thresholds. A large literature investigates this question through the so-called "slow-server" problem (e.g., see [11], [12], [16]). Other illustrations are overflow policies where agents are initially reserved for only one customer type but can treat another customer type in case of high congestion (e.g., see [1], [10]). When there is a switching time or a switching cost to change the job type in service, it may also be preferred for a server to remain idle and ready to serve the same job type instead of serving the other job type. For this purpose, cycling reservation strategies are studied (e.g., see [17], [3]).

The particularity of our article compared to the existing ones on the same subject is to consider a threshold reservation strategy with a *unique class* of customers who may balk. The choice for a reservation threshold policy follows from the knowledge that this type of policies outperforms randomized policies since they take into account the system state. In addition, threshold policies are deterministic; for each state there is a unique action. This makes these policies easy to implement in practice. Finally, threshold policies are controlled by only one parameter which also makes them possible to analyze and implement.

We choose to adapt the threshold policy developed in the call blending references for a unique class of customers. We consider a multi-server single queue with infinite capacity and s identical, parallel servers. The arrival process of customers is Poisson with rate λ. Service times are independent and exponentially distributed with rate μ. At a customer's arrival, if at least one server is available then this customer is directly served, otherwise with probability r he/she accepts waiting and is routed to a first-come-first-served queue. The reservation policy is of threshold type. We define a threshold c on the number of servers. After a service completion, if the number of busy servers is higher than or equal to $s - c$ then no customer is routed to service. Otherwise, if the queue is non-empty then the first customer in line is routed to service. In other words, c servers are reserved for new arriving customers ($0 \leq c < s$).

The objective of this article is to evaluate the performance of this system and determine the effect of reservation on the proportion of lost customers and on the waiting time. The idea is to show the motivations and the risks in implementing such a policy. In Section 2, we present the Markov chain associated to this system and develop a method to obtain the explicit expressions of the stationary probabilities. Section 3 is devoted to the impact of reservation on balking. We prove that reservation can reduce the proportion of lost customers. This might be unexpected by service providers. By forcing some servers not to work, the

overall productivity may increase. The improvement can be significant in large systems, under high workload situations and with a low customers' waiting aversion. In Section 4, we develop a method to compute the Laplace Stieltjes Transform (LST) of the waiting time distribution of an arriving customer. This allows us to obtain explicit expressions of the first and second moments of the waiting time. As expected, although reservation helps to reduce the probability of delay, it deteriorates the expected waiting time due to the existence of situations where a server may remain idle while a customer is waiting. In addition, in Section 5 we include abandonment in the model. From a numerical analysis, we show that reservation may be counterproductive if customers' impatience is high.

2 Stationary probabilities

In this section, we derive explicitly the stationary probabilities. The system is modeled using a two dimensional continuous-time Markov chain. We denote by (x, y) a state of the system for $0 \leq x \leq s$ and $y \geq 0$, where x represents the number of busy servers and y represents the number of customers in the queue. The state space S is $S = \{(0,0),(0,1),\cdots,(0,s)\} \cup \{s-c, s-c+1, \cdots, s\} \times \mathbb{N}^*$. We denote by $p_{x,y}$ the stationary probability to be in state (x, y) and by a the ratio λ/μ.

We next describe the 4 possible transitions in the Markov chain.

1. An arrival with rate λ while a server is available ($0 \leq x < s$), which changes the state to $(x+1, y)$. The number of busy server is increased by one.

2. An arrival with rate $r\lambda$ while all servers are busy ($x = s$), which changes the state to $(x, y+1)$. The number of customers in the queue is increased by 1.

3. A service completion with rate $(s-c)\mu$ while the queue is not empty ($y > 0$) and the number of busy servers is equal to $s-c$ ($x = s-c$), which changes the state to $(x, y-1)$. The number of customers in the queue is reduced by 1.

4. A service completion with rate $\min(s, x)\mu$ while the queue is empty ($y = 0$) or the queue is not empty but the number of busy servers is strictly higher than $s-c$ ($x > s-c$, $y > 0$), which changes the state to $(x-1, y)$. The number of busy servers is reduced by 1.

The Markov chain is depicted in Figure 1 of Section 1 of the online supplement. In Theorem 1, we give the stationary probabilities and the stability condition. The proof of Theorem 1 is given in Section 2 of the online supplement.

Theorem 1 *Under the stability condition* $r\frac{(s-c-1)!}{s!}a^{c+1} < 1$*, we have*

$$p_{x,0} = \frac{a^x}{x!}p_{0,0}, \text{ for } 0 \le x \le s-c, \tag{1}$$

$$p_{x,0} = \frac{a^x}{x!}\frac{1 + r\sum\limits_{k=0}^{s-x-1}\frac{(x+k)!}{s!}a^{s-k-x}}{1 + r\sum\limits_{k=0}^{c-1}\frac{(s-c+k)!}{s!}a^{c-k}}p_{0,0}, \text{ for } s-c \le x \le s, \tag{2}$$

$$p_{x,y} = \frac{a^x\left(1 - a^{s-(x+c+1)}\left[1 - r\frac{(s-c-1)!}{s!}a^{c+1}\right]\frac{\sum\limits_{k=0}^{s-x-1}(x+k)!a^{-k}}{\sum\limits_{k=0}^{c}(s-c-1+k)!a^{-k}}\right)}{rx!\sum\limits_{k=0}^{c}\frac{(s-c-1+k)!}{s!}a^{c+1-k}}\left(\frac{r\sum\limits_{k=0}^{c}\frac{(s-c-1+k)!}{s!}a^{c+1-k}}{1 + r\sum\limits_{k=1}^{c}\frac{(s-c-1+k)!}{s!}a^{c+1-k}}\right)^{y+1}p_{0,0}, \tag{3}$$

for $s-c \le x \le s, y > 0$*, with*

$$p_{0,0} = \left[\sum_{x=0}^{s-c-1}\frac{a^x}{x!} + \frac{\sum\limits_{x=s-c}^{s}\frac{a^x}{x!}}{1 - r\frac{(s-c-1)!}{s!}a^{c+1}}\right]^{-1}. \tag{4}$$

3 Impact on Balking

We evaluate in Proposition 1 the impact of the reservation threshold on the proportion of customers who balk, P.

Proposition 1 *P is decreasing and convex in c.*

The consequence of this result may be unexpected by service providers. By forcing some servers not to work when some work is available, these servers will at the end work more. The reason is due to the reduction of the flow of customers who balk. This result is stated in Corollary 1 and follows Proposition 1 using the relation: Servers' utilization $= \frac{\lambda(1-P)}{s\mu}$.

Corollary 1 *Servers' utilization is increasing and concave in c.*

Proof. From Theorem 1, we deduce the explicit expression of P from $P = (1-r)\sum\limits_{y=0}^{\infty}p_{s,y}$. This leads to

$$P = \frac{(1-r)\frac{a^s}{s!}}{\left(1 - r\frac{(s-c-1)!}{s!}a^{c+1}\right)\sum\limits_{x=0}^{s-c-1}\frac{a^x}{x!} + \sum\limits_{x=s-c}^{s}\frac{a^x}{x!}}. \tag{5}$$

The proportion of customers who balk, P, can be rewritten as

$$P = (1-r)\left[\sum_{x=0}^{s}\frac{s!}{x!a^{s-x}} - r\sum_{x=0}^{s-c-1}\frac{(s-c-1)!}{x!a^{s-x-c-1}}\right]^{-1} = (1-r)\left[B(s,a)^{-1} - rB(s-c-1,a)^{-1}\right]^{-1},$$

where $B(t,a) = \left(\sum\limits_{x=0}^{t}\frac{t!}{x!a^{t-x}}\right)^{-1}$ is the Erlang loss function for $t \in \mathbb{N}$.

For analytic purposes, it will be convenient to extend the domain of the Erlang loss function to non-integer values. Following [8], the continuous extension of the classic Erlang loss function is defined for any $t \in [0, +\infty)$ and $a > 0$ by

$$B(t,a) = \left(\int_0^{+\infty} a e^{-ax} (1+x)^t \mathrm{d}x \right)^{-1}$$

As observed by [8], the above integral expression coincides with the initial definition of $B(t,a)$ for integer values of t.

This allows us to obtain the monotonicity properties by derivation. We have

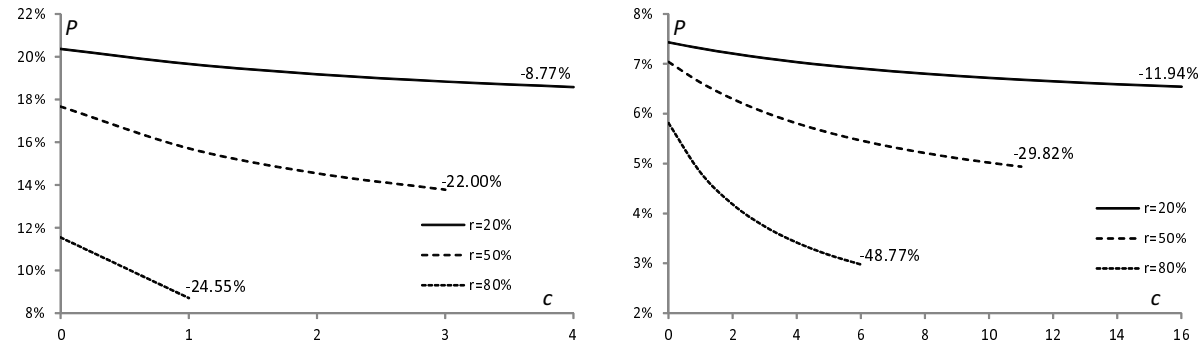
$$\frac{\partial P}{\partial c} = -(1-r)r \frac{\int_0^{+\infty} a e^{-ax} \ln(1+x)(1+x)^{s-c-1} \mathrm{d}x}{\left[B(s,a)^{-1} - rB(s-c-1,a)^{-1}\right]^2} < 0.$$

So, P is decreasing in c. We also have

$$\frac{\partial^2 P}{\partial c^2} = (1-r)r \frac{\frac{\partial^2 B(s-c-1,a)^{-1}}{\partial c^2} \left[B(s,a)^{-1} - rB(s-c-1,a)^{-1}\right] + 2r \left(\frac{\partial B(s-c-1,a)^{-1}}{\partial c}\right)^2}{\left[B(s,a)^{-1} - rB(s-c-1,a)^{-1}\right]^3}.$$

Since $\frac{\partial^2 B(s-c-1,a)^{-1}}{\partial c^2} = \int_0^{+\infty} a e^{-ax} (\ln(1+x))^2 (1+x)^{s-c-1} \mathrm{d}x > 0$, we have $\frac{\partial^2 P}{\partial c^2} > 0$ and P is convex in c. □

In Figure 1, we represent the proportion of customers who balk as a function of the reservation threshold for different values of the system parameters. For each curve, we only evaluate P when the system is stable. At the end of each curve, for the highest possible value of the reservation threshold, we compute the relative improvement obtained in comparison with a situation without reservation, where the relative improvement is defined by relative improvement $= \frac{P(c=c_{max}) - P(c=0)}{P(c=0)}$. This computation is done to show the maximal potential in applying a reservation policy. In addition to illustrate Proposition 1, we observe



(a) $s = 10$ (b) $s = 100$

Figure 1: P as a function of c ($\mu = 1$, $\lambda = s\mu$)

that the highest improvement is obtained for the highest value of r. In other words, the fewer customers are tempted to balk, the more balking can be reduced by a reservation strategy. This makes sense intuitively; for low values of r most customers balk, so only a small proportion of customers accept to wait. This means that the chance to find an available server is high at customer's arrival even without reservation. So, reservation has only a limited effect in this case. The opposite holds when r is high.

We also observe that we get higher improvement in larger systems. Recall that balking is due to situations where all servers are busy at a new customer's arrival. The idea of reservation is to create idle capacities to absorb the new arriving customers. From a situation where all servers are busy, the first service completion which allows the system to have a first idle server occurs after an expected duration of $1/(s\mu)$. This duration reduces with the system size. So, large systems have a stronger ability to recreate idle capacity which in turn leads to fewer customers who balk.

4 Consequences on the waiting time

We provide here a method to compute the Laplace Stieltjes transform (LST) of the waiting time distribution of an arriving customer in order to evaluate the consequences of reservation on the waiting time.

Recall that an arriving customer waits only if all servers are busy. We focus then on an arriving customer in state (s, y), signifying that there are s busy servers and y other customers in the queue ($y \geq 0$). The state of the system is denoted by (x, y) where x is the number of busy servers and y is the number of customers in front of the considered customer.

The state transition diagram for the finite two-dimensional birth-and-death process modeling the behavior of a tagged customer in the queue is shown in Figure 2. This process has $(c+1) \times (y+1)$ transient states and one absorbing state called "Service". Let $W_{x,y}(t)$ be the unconditional LST of the density function for

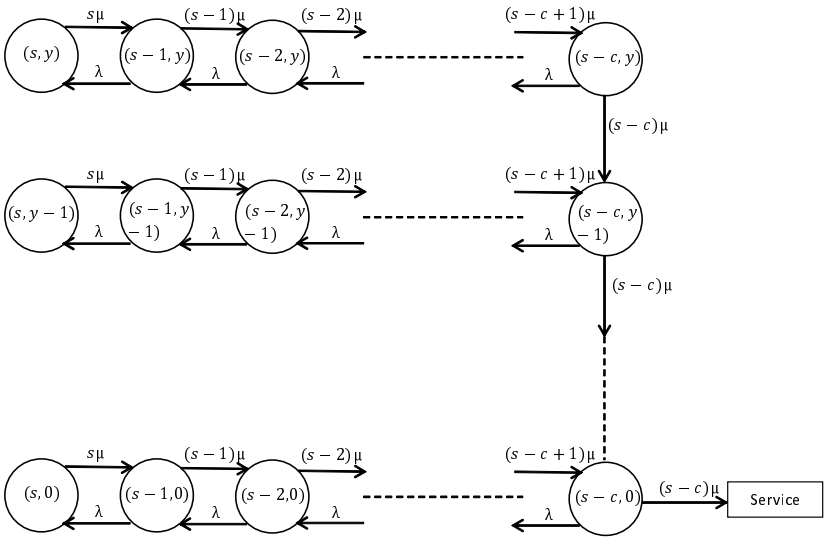


Figure 2: Transition diagram

the passage time from state (x, y) to state "Service". In Theorem 2, we provide a method which allows us to obtain the expression of $W_{s,y}(t)$ after a finite number of computations. The proof is given in Section 3 of the online supplement.

Theorem 2 *For $y \geq 0$, we have*

$$W_{s,y}(t) = \frac{1}{f_{c+1}(t)} \left(\frac{f_c(t)}{f_{c+1}(t)} \right)^y, \tag{6}$$

where the functions $f_k(t)$ for $0 \leq k \leq c+1$ are given by $f_0(t) = 1$, $f_1(t) = 1 + \frac{t}{s\mu}$ and

$$f_k(t) = \left(1 + \frac{\lambda + t}{(s-(k-1))\mu} \right) f_{k-1}(t) - \frac{\lambda}{(s-(k-1))\mu} f_{k-2}(t), \tag{7}$$

for $2 \leq k \leq c+1$.

In Corollary 2, we give the expected value and the variance of the waiting time for an arriving customer in state (s, y) for $y \geq 0$; $E(W_y)$ and $V(W_y)$. The proof is given in Section 3 of the online supplement.

Corollary 2 *We have*

$$E(W_y) = \frac{1}{\mu} \left((y+1) \sum_{k=0}^{c} \sum_{i=0}^{k} \frac{a^i (s-k-1)!}{(s-k+i)!} - y \sum_{k=0}^{c-1} \sum_{i=0}^{k} \frac{a^i (s-k-1)!}{(s-k+i)!} \right),$$

$$V(W_y) = \frac{1}{\mu^2} \left[2(y+1) \left(\sum_{k=0}^{c} \sum_{i=0}^{k} \frac{a^i (s-k-1)!}{(s-k+i)!} \right)^2 - (2y+1) \left(\sum_{k=0}^{c} \sum_{i=0}^{k} \frac{a^i (s-k-1)!}{(s-k+i)!} \right) \left(\sum_{k=0}^{c-1} \sum_{i=0}^{k} \frac{a^i (s-k-1)!}{(s-k+i)!} \right) \right.$$
$$\left. + y \left(\sum_{k=1}^{c} \sum_{i=1}^{k-1} \sum_{j=0}^{i-1} \sum_{m=0}^{j} \frac{a^{m+k-i+1} (s-k)!(s-j-1)!}{(s-1)!(s-j+m)!} \right) - (y+1) \left(\sum_{k=1}^{c+1} \sum_{i=1}^{k-1} \sum_{j=0}^{i-1} \sum_{m=0}^{j} \frac{a^{m+k-i+1} (s-k)!(s-j-1)!}{(s-1)!(s-j+m)!} \right) \right]$$

From Corollary 2, one can compute the first and the second moments of the waiting time for an arbitrary customer who do not balk by considering the system at arrival instants using the stationary probabilities in Theorem 1. For the expected waiting time, $E(W)$, we get

$$E(W) = \frac{r}{\mu} \frac{\sum_{k=0}^{c} \frac{(s-c-1+k)!}{s!} a^{c-k} \sum_{x=s-c}^{s} \frac{a^x}{x!} - \left(1 - r \frac{(s-c-1)!}{s!} a^{c+1} \right) \sum_{x=s-c}^{s} \sum_{k=0}^{s-x-1} \frac{(x+k)!}{s!x!} a^{s-1-k}}{\left(1 - r \frac{(s-c-1)!}{s!} a^{c+1} \right) \left(\sum_{x=0}^{s-1} \frac{a^x}{x!} - r \frac{(s-c-1)!}{s!} a^{c+1} \sum_{x=0}^{s-c-2} \frac{a^x}{x!} \right)}. \tag{8}$$

Another way to obtain $E(W)$ is to apply Little's law by first computing the expected number of customers in the queue and next by dividing this quantity by $\lambda(1-P)$. The complexity of $E(W)$ does not allow us to prove the first and second order monotonicity results in c. Yet, from an extensive numerical study, it seems clear the $E(W)$ is increasing and convex in c (see Figure 3).

Reservation produces two phenomena which may degrade the waiting performance measures. First, as shown in Section 3, reservation reduces the proportion of customers who balk and therefore increases the workload. This is somehow positive since this also allows the service provider to serve more customers. Yet, it is clear that a higher workload will result in longer waiting times for customers who wait. Second, with reservation one may encounter situations with one idle server and a customer waiting in the queue, this may also have a negative effect on the waiting performance measures. To isolate this second phenomenon, we present in Figure 4 different waiting time performance measures in the case $r = 1$ (no balking); the

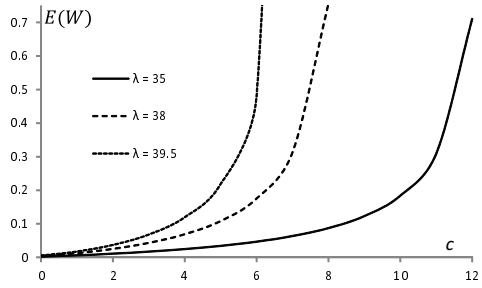
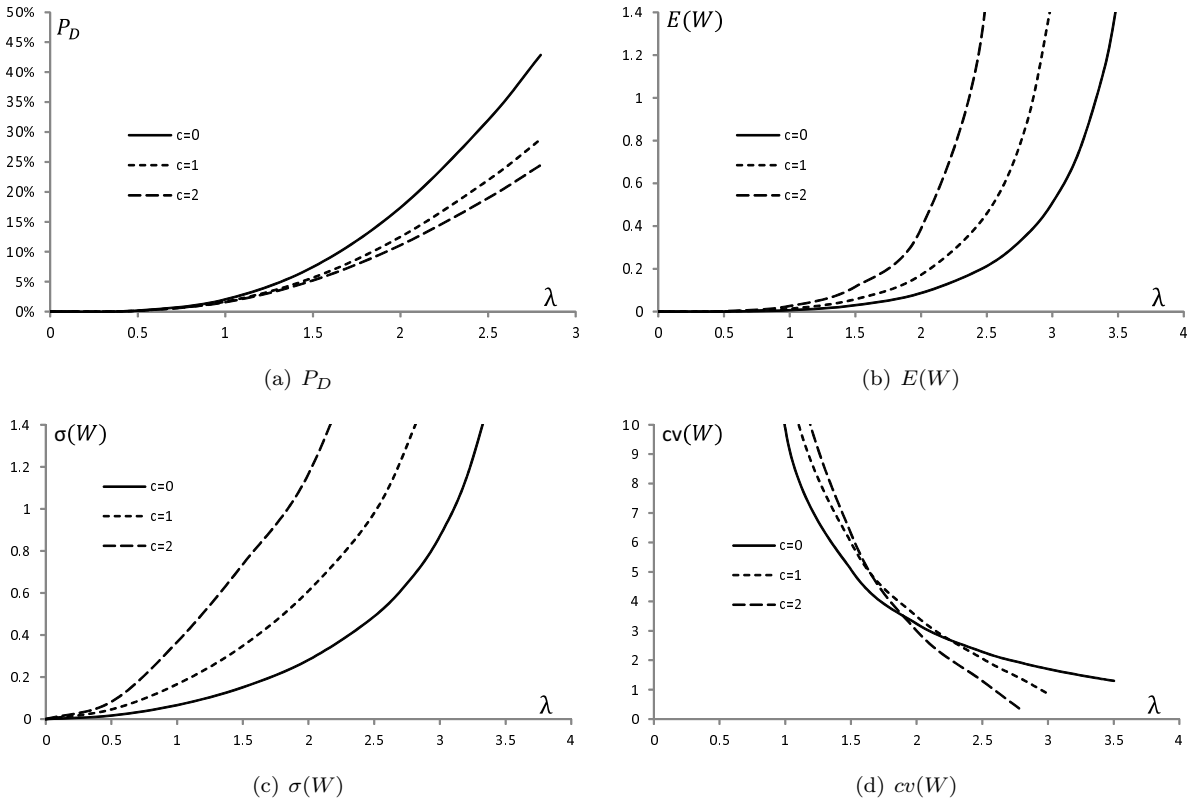


Figure 3: $E(W)$ as a function of c ($\mu = 1$, $s = 40$, $r = 50\%$)

probability of delay (P_D), the expected waiting time ($E(W)$), the standard deviation of the waiting time ($\sigma(W)$), and the coefficient of variability ($cv(W) = \sigma(W)/E(W)$). The idea is to show the main impacts of reservation on the waiting time performance measures.



(a) P_D (b) $E(W)$

(c) $\sigma(W)$ (d) $cv(W)$

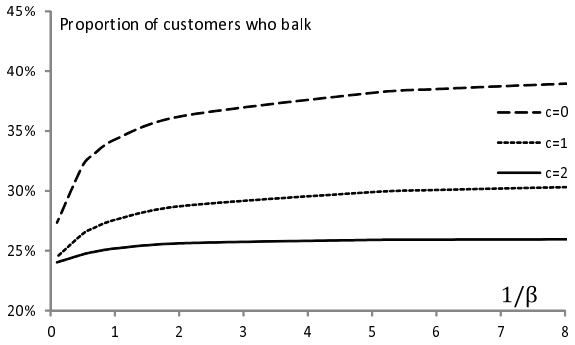
Figure 4: Waiting time performance measures as a function of λ ($\mu = 1$, $s = 4$, $r = 1$)

The first observation in Figure 4(a) is that the probability of delay reduces with reservation. This is particularly apparent in high workload situations. This result is already proven in Proposition 1 since the probability of delay is proportional with the proportion of customers who balk; $P = (1 - r)P_D$. The second observation is that the expected value and the standard deviation of the waiting time increase with reservation (Figures 4(b) and 4(c)). As mentioned above, the reason is related to the loss of work-conservation. It is however interesting to note that the variability of the waiting time measured by the coefficient of variability reduces with reservation in high workload situations. This is mainly due to the high proportion of customers

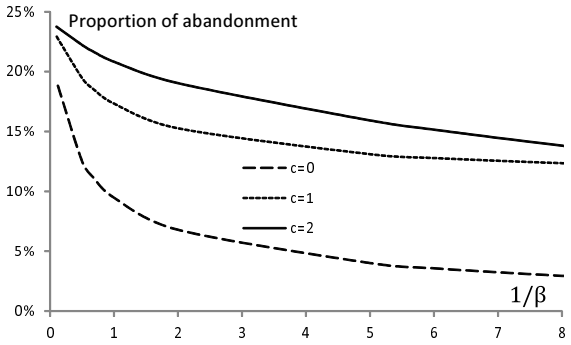
who do not wait.

5 Benefits of reservation with abandonment?

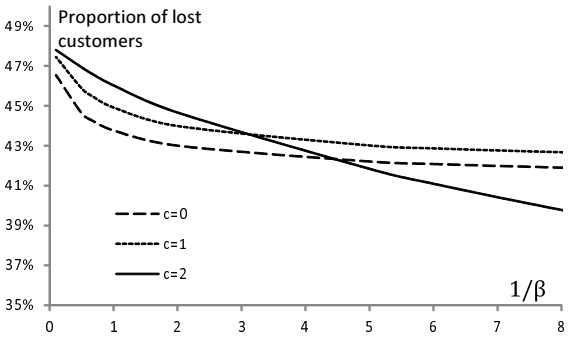
In Section 4 of the main document, we have shown the degradation of the waiting time with reservation. Given that high waiting times may induce a high abandonment, one may wonder if reservation instead of allowing to serve more customers would instead lead to more customers' loss. We therefore propose here to include abandonment in the modeling. For this purpose, we consider an exponential abandonment with parameter β. The complexity added by abandonment does not allow us to obtain simple closed-form expressions for the stationary probabilities. However, in Section 4 of the online supplement, we propose a numerical method to derive them.



(a) Proportion of customers who balk



(b) Proportion of abandonment



(c) Proportion of lost customers

Figure 5: Impact of abandonment ($s = 5$, $\lambda = 8$, $\mu = 1$, $r = 0.5$)

In Figure 5, we illustrate the impact of abandonment on the proportion of lost customers. As expected, the proportion of customers who balk increases with the expected patience ($1/\beta$) while the proportion of abandonment reduces (Figures 5(a) and 5(b)). This creates a compensation phenomenon. The value of reservation then depends on the customers patience in the queue. From Figure 5(c), we observe that reservation deteriorates the overall proportion of lost customers (the sum of the proportion who balk and the proportion of customers who abandon) when customers are very impatient.

Acknowledgment. The author wants to express his gratitude to the reviewer and the associated editor for their useful comments, that significantly improved this paper. The author would like also to thank Sébastien Thorel from INTERACTIV GROUP for our helpful discussions.

References

[1] W. Barth, M. Manitz, and R. Stolletz. Analysis of two-level support systems with time-dependent overflow-a banking application. *Production and Operations Management*, 19(6):757–768, 2010.

[2] S. Bhulai and G. Koole. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, 48(8):1434–1438, 2003.

[3] O.J. Boxma, H. Levy, and U. Yechiali. Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research*, 35(3):187–208, 1992.

[4] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50, 2005.

[5] A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A.N. Avramidis. Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645, 2007.

[6] G. Dobson, S. Hasija, and E.J. Pinker. Reserving capacity for urgent patients in primary care. *Production and Operations Management*, 20(3):456–473, 2011.

[7] N. Gans and Y.-P. Zhou. A call-routing problem with service-level constraints. *Operations Research*, 51:255–271, 2003.

[8] A.A. Jagers and E.A. Van Doorn. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Review*, 33(2):281, 1991.

[9] G. Koole. *Call Center Optimization*. MG Books, 2013.

[10] G. Koole, B. Nielsen, and T. Nielsen. Optimization of overflow policies in call centers. *Probability in the Engineering and Informational Sciences*, 29(3):461–471, 2015.

[11] B Krishnamoorthi. On poisson queue with two heterogeneous servers. *Operations Research*, 11(3):321–330, 1963.

[12] R.L. Larsen and A.K. Agrawala. Control of a heterogeneous two-server exponential queueing system. *IEEE Transactions on Software Engineering*, (4):522–526, 1983.

[13] B. Legros, O. Jouini, and G. Koole. Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 47(4):414–430, 2015.

[14] B. Legros, O. Jouini, and G. Koole. Optimal scheduling in call centers with a callback option. *Performance Evaluation*, 95:1–40, 2016.

[15] G. Pang and O. Perry. A logarithmic safety staffing rule for contact centers with call blending. *Management Science*, 61(1):73–91, 2014.

[16] V.V. Rykov. Monotone control of queueing systems with heterogeneous servers. *Queueing Systems*, 37(4):391–403, 2001.

[17] P. Tran-Gia and R. Dittmann. A discrete-time analysis of the cyclic reservation multiple access protocol. *Performance Evaluation*, 16(1):185–200, 1992.