Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Reducing Delay in Retrial Queues by Simultaneously Differentiating Service and Retrial Rates

Jinting Wang, Zhongbin Wang, Yunan Liu

To cite this article:
Jinting Wang, Zhongbin Wang, Yunan Liu (2020) Reducing Delay in Retrial Queues by Simultaneously Differentiating Service and Retrial Rates. Operations Research

Published online in Articles in Advance 18 Sep 2020

. https://doi.org/10.1287/opre.2019.1933

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Reducing Delay in Retrial Queues by Simultaneously Differentiating Service and Retrial Rates

Jinting Wang,[a] Zhongbin Wang,[b,c] Yunan Liu[d,*]

[a] School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100081, China; [b] Business School, Nankai University, Tianjin 300071, China; [c] Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China; [d] Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695

*Corresponding author

Contact: jtwang@bjtu.edu.cn, https://orcid.org/0000-0003-4946-2719 (JW); wangzhongbin@bjtu.edu.cn, https://orcid.org/0000-0002-1154-5861 (ZW); yliu48@ncsu.edu, https://orcid.org/0000-0001-9961-2610 (YL)

Received: November 29, 2016
Revised: April 1, 2018; December 18, 2018
Accepted: July 22, 2019
Published Online in Articles in Advance: September 18, 2020

Subject Classifications: queues: algorithms, applications, optimization
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2019.1933

Abstract. In this article, we introduce a service grade differentiation policy for queueing models with customer retrials. We show that the average waiting time can be reduced through strategically allocating the rates of service and retrial times without needing additional service capacity. Countering to the intuition that higher service variability usually yields a larger delay, we show that the benefits of our simultaneous service-and-retrial differentiation policy outweigh the impact of the increased service variability. We present a necessary and sufficient condition under which the proposed policy reduces the waiting time and a closed-form expression for the optimal allocation policy. In heavy traffic, our policy can asymptotically reduce both the delay and the number of customer retrials before entering service by a significant factor, which is a function of the ratio of the service rate to the retrial rate.

Funding: We acknowledge support from the National Natural Science Foundation of China (NFSC) [Grants 71871008 and 71571014] (J. Wang and Z. Wang) and NSF [Grant CMMI 1362310] (Y. Liu).
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2019.1933.

Keywords: retrial queues • simultaneous service-and-retrial differentiation • service differentiation

1. Introduction

Customer retrials commonly occur in many service systems, such as healthcare, call centers, mobile networks, computer systems, and inventory systems (see Tran-Gia and Mandjes 1997, Wang et al. 2001, Ren and Zhou 2008, Yom-Tov and Mandelbaum 2014, Li et al. 2015b, and Wang et al. 2017). In general, there are two types of customer retrials: (i) In certain service systems, such as call centers, excessive queueing delays or a lack of waiting space can cause waiting customers to temporarily leave the system before entering service and return at a future time (Falin and Templeton 1997; Mandelbaum et al. 1999, 2002; Ding et al. 2015). (ii) After completing service, customers may return at a later time because their initial service was "unsatisfactory." For instance, the treatment of a patient by a doctor in a hospital may naturally occur in stages, starting with an initial screening and continuing later after tests have been ordered and completed (Artalejo et al. 2006; Liu and Whitt 2014, 2017; Yom-Tov and Mandelbaum 2014). In this paper, we consider retrials of the first type (i.e., retrials before service). We develop a service differentiation policy that is beneficial for models of this type because it reduces the average waiting time without necessitating an increase in the overall service capacity.

In conventional retrial models, customer behavior is usually assumed to be homogeneous; such a scenario is characterized by *independent and identically distributed* (i.i.d.) service and retrial times with common service and retrial rates. In contrast, our service differentiation policy works as follows: we randomly classify the originally homogeneous customers into *heterogeneous* customer groups with stochastically shorter and longer service and retrial times while maintaining a constant overall service capacity (the average service time and retrial time are left unchanged). At first glance, such a policy seems unappealing because of the general consensus that customer differentiation increases service variability, which is expected to prolong the waiting time, thus leading to excessive system congestion. However, this differentiation policy creates priorities among the originally homogeneous customers, which can be used to shorten the average total orbit time. We prove that, if this differentiation policy is correctly implemented (that is, if the heterogeneous service and retrial rates are properly selected and paired), then the overall

average delay can be significantly reduced while holding the average service time unchanged.

We emphasize that our differentiation policy is different from dynamic service control policies that make use of system information, for example, a policy in which the service provider speeds up (slows down) the service process when the system is more (less) congested (George and Harrison 2001). In some settings, dynamic service control policies may become impractical if the timely acquisition of the required system state information becomes difficult or costly. By contrast, from this perspective, our differentiation policy is quite easy to implement in practice because it is independent of the system state.

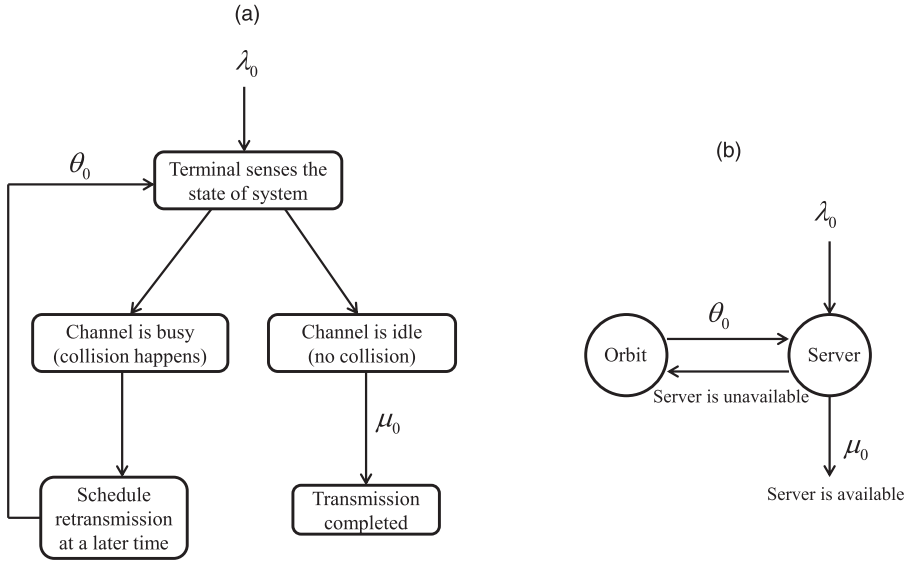
One potential application of the retrial model and the aforementioned differentiation policy is a *local area network* (LAN). In a LAN, a large number of terminals are connected to a centralized internet service provider. Upon receiving a job, a terminal attempts to determine the state of the server (channel). If the server is idle, the job is transmitted to the server from the terminal; if the server is busy, the job is temporarily stored at the terminal and waits for a service attempt at a later time. One widely adopted communication protocol in LANs is *carrier sense multiple access with collision detection* (CSMA/CD), which has been proven effective in rescheduling packets for future transmission when collisions occur (see Fayolle et al. 1977, Tobagi and Hunt 1980, Choi et al. 1992, Li et al. 2015a). CSMA/CD can be further classified into two types: (i) nonpersistent and (ii) persistent. In the nonpersistent CSMA/CD protocol, if the channel is currently idle, the terminal initiates the transmission of a job; otherwise, the terminal schedules a retransmission at a future time. See Figure 1(a) for an illustration. In contrast, in the persistent CSMA/CD protocol, if the channel is found to be busy, the terminal continuously attempts to transmit until the channel eventually becomes free. See Fayolle et al. (1977), Choi et al. (1992), Li et al. (2015a), and Tobagi and Hunt (1980) for details of the persistent and nonpersistent CSMA/CD protocols. In the case of retrial systems with different retrial rates, we can regard the persistent (nonpersistent) CSMA/CD protocol as representing a customer group that makes more (less) frequent service attempts, meaning that these customers have a higher (lower) retrial rate. With the aim of improving the overall system performance, we propose a new service differentiation policy that considers a mixture of frequent and infrequent retrials.

We focus on studying how service differentiation influences the performance of retrial models. Our findings reveal that the overall customer delay can be reduced only when *both the service rate and retrial rate are simultaneously differentiated, not either one alone.* To obtain structural results and useful insights, we mainly focus on a single-server $M/G/1$ retrial model with Poisson arrivals (M), i.i.d. service times (G), i.i.d. orbit times (times until the next retrial) that are exponentially distributed, and a buffer size of zero. We later extend our analysis to a retrial system with a finite buffer size, which is potentially suitable for modeling more realistic service systems in practice.

1.1. Literature Review

1.1.1. Queueing Models with Retrials. There is a large body of research on retrial queueing models. The optimal retrial rate and routing policy for an $M/M/1$

Figure 1. Dynamics of the CSMA/CD Protocol with Homogeneous Service



Dynamics of the CSMA/CD protocol

$M/G/1$ retrial model with homogeneous service

retrial queue have been studied by Elcan (1994, 1999), Hassin and Haviv (1996), Avrachenkov et al. (2015), and Liang and Kulkarni (1999). Aissani and Phung-Duc (2015) developed a call center retrial model with two-way communications (i.e., including incoming and outgoing calls). Artalejo (1997) considered an $M/G/1$ retrial queue with service vacations and obtained the optimal control policy for this queue under the so-called N-policy (i.e., the server is turned off when the system becomes empty and is turned on again when the queue length reaches N). Gharbi et al. (2009) investigated systems with multiclass retrial customers. See Falin and Templeton (1997) and Artalejo and Gómez-Corral (2008) for reviews of retrial models.

1.1.2. Service Rate Control Policies. There are two types of service rate control policies: dynamic, in which decisions are made based on the real-time system state, and static, which are independent of the system state. Dynamic rate control problems have been proposed and studied by George and Harrison (2001), Ata and Shneorson (2006), and Adusumilli and Hasenbein (2010); these authors focused on reducing the waiting time by adjusting the service rates using system information, such as queue lengths. Another related line of research concerns how to balance the waiting time and service value in a quality-based queueing system; see Hopp et al. (2007), Anand et al. (2011), and the references therein. Recently, Xu et al. (2015) proposed a static service differentiation policy for a single-server $M/G/1$ queue. Their idea is to randomly assign customers different service rates, thus creating heterogeneous service grade information, which enables the implementation of the *shortest expected processing time* (SEPT) policy; see Schrage and Miller (1966) for a further discussion of this policy. With this differentiation policy, customers are scheduled based on their absolute priorities; therefore, this policy can be regarded as a special case of the c-μ rule (Smith 1956, Mendelson and Whang 1990). Xu et al. (2015) showed that such a differentiation rule, if properly implemented, can reduce the average waiting time without affecting the mean service time; in addition, they discovered that the performance of the system can be improved by increasing the number of service grades (a 5% improvement can be asymptotically achieved as the number of service grades approaches infinity).

In this paper, we develop a differentiation rule for a service system with customer retrials. Our policy simultaneously differentiates the service rate and retrial rate, thereby creating *relative* priorities (rather than absolute priorities as in Xu et al. (2015)) among customer groups. To the best of our knowledge, such a service-and-retrial differentiation policy has never been studied before in the retrial queueing literature.

1.2. Our Contributions

- For the first time to our knowledge, we study a service-and-retrial differentiation policy for retrial queueing systems. We prove that the overall mean waiting time can be reduced by offering heterogeneous service and retrial rates to originally homogeneous customers while holding the total service capacity unchanged. We present a necessary and sufficient condition under which our differentiation policy dominates the homogeneous service policy.
- In contrast to the results of Xu et al. (2015), in which the performance improves as the number of service grades increases, we show that the minimum customer delay can be achieved by differentiating customers into exactly two grades; moreover, the system benefits from a high level of the differentiation between these two customer grades. See Theorem 2 for the optimal structure of our policy, including the differentiation probabilities, workloads, service rates, and retrial rates for both service grades. See also Remark 5 for further discussion and insight.
- We quantify the asymptotic performance gain when the system is in heavy traffic (as the traffic intensity approaches one). In addition to the mean delay, our service differentiation rule can help reduce the *number of trials* and the *slowdown*, which are also considered useful service-level indicators. For example, we show that it is possible to reduce both the mean delay and the number of trials by a factor of $(c_v^2+1)/(c_v^2+1+2\mu_0/\theta_0)$ when the system is in heavy traffic, where μ_0/θ_0 is the ratio of the service rate to the retrial rate and c_v^2 is the squared coefficient of variation of the service times.
- We present numerical examples to verify the effectiveness of our differentiation policy. We report a sensitivity analysis of the system performance (delay, number of trials, etc.) with respect to the model parameters under the optimal service differentiation policy. To show that the insights gained from analyzing the simple $M/G/1$ retrial model are of practical value, we also consider some more general model settings, such as the cases of a finite waiting room and a convex delay cost; we discover that our service differentiation rule can continue to be beneficial for improving the system performance in these cases.

1.2.1. Organization of the Paper. In Section 2, we describe the single-server retrial queueing model and introduce our new service differentiation policy. In Section 3, we present a necessary and sufficient condition under which our differentiation policy dominates the homogeneous service policy. In Section 4, we prove the suboptimality of the m-grade case ($m \geq 3$)

and present the optimal structure of the differentiation policy with two service grades. In Section 5, we quantify the asymptotic performance gain when the system is in heavy traffic, using performance metrics such as the mean delay and the number of trials. In Section 6, we conduct numerical experiments and sensitivity analysis. In Section 7, we extend our analysis to some more general model settings. We make concluding remarks in Section 8. Supplementary materials are provided in the e-companion.

2. The Model

Our base model is a single-server $M/G/1$ retrial queueing system having Poisson arrivals with rate λ_0, i.i.d. service times with mean service time $1/\mu_0$, and zero waiting capacity. If an arriving customer finds the server busy serving another customer, this customer immediately enters an orbit queue of infinite capacity. Customers in the orbit queue attempt to reenter service at a rate of θ_0, where the specific orbit times after which they attempt reentry are i.i.d. following an exponential distribution. This model is depicted in Figure 1(b). In the base model with homogeneous service, let S_0 be a generic random service time with mean $E[S_0] = 1/\mu_0$ and *squared coefficient of variation* (SCV) of $c_v^2 = \mathrm{Var}(S_0)/E[S_0]^2$.

2.1. Simultaneous Service-and-Retrial Differentiation

Our service differentiation policy is described as follows. We allow the server to offer a discrete set of service rates $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_m)$ and orbit rates $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_m)$. Upon arrival, a customer is immediately assigned with service and orbit rates (μ_k, θ_k) *with probability* (w.p.) p_k, $1 \le k \le m$ such that $\sum_{k=1}^m p_k = 1$. In this way, the originally homogeneous customers with a single service rate μ_0 and a single orbit rate θ_0 are manually divided into heterogenous customer groups with different service grades, in which the k^{th} service grade is characterized by the pair (μ_k, θ_k). Without loss of generality, we assume that $0 < \mu_1 < \cdots < \mu_m$. As we see later, the optimal policy requires that $0 < \theta_1 < \cdots < \theta_m$. This should not be too surprising: in the context of quality-based service systems, it is reasonable to offer a longer waiting time (i.e., a lower orbit rate θ_k) to a customer demanding higher service quality, which is equivalent to a longer service time (i.e., a lower service rate μ_k). In addition, we remark that the value of the rate θ_k can be interpreted as a measure of the relative priority level for customers of class k (Haviv and Van Der Wal 1997).

To achieve differentiated service grades, we scale the base random service time S_0 defined previously while maintaining the form of its distribution (Debo et al. 2008, Xu et al. 2015). Under the m-grade simultaneous service-and-retrial differentiation (SSRD) policy, a customer's service time S is given by

$$S \sim \begin{cases} S_1 \equiv \left(\dfrac{\mu_0}{\mu_1}\right) S_0, & \text{w.p. } p_1, \\ \quad\vdots & \\ S_m \equiv \left(\dfrac{\mu_0}{\mu_m}\right) S_0, & \text{w.p. } p_m, \end{cases} \tag{1}$$

where S_i is the generic service time of class i. To ensure that the overall expected service time and orbit time remain unchanged, we impose the following *fixed-capacity* constraints:

$$E[S] = \sum_{i=1}^m \frac{\mu_0}{\mu_i} E[S_0] p_i = \sum_{i=1}^m \frac{p_i}{\mu_i} = \frac{1}{\mu_0} = E[S_0] \quad \text{and}$$
$$\sum_{k=1}^m \frac{1}{\theta_k} p_k = \frac{1}{\theta_0}, \quad \text{with} \quad \sum_{k=1}^m p_k = 1, \tag{2}$$

where $1/\mu_0$ and $1/\theta_0$ are the overall mean service time and orbit time, respectively.

We emphasize that our service differentiation rule is *static*; that is, the grade of a customer is determined immediately upon that customer's arrival, independently of the state of the system. See Figure 2(b) for an illustration of the $M/G/1$ retrial queue under the SSRD policy (see also Figure 2(a) for the corresponding LAN example with both frequent and infrequent retrials).

Next, we show that the SSRD policy dominates the homogeneous service policy ($m = 1$) as long as the parameters (μ_k, θ_k) are properly chosen.

3. Dominance of the SSRD Policy

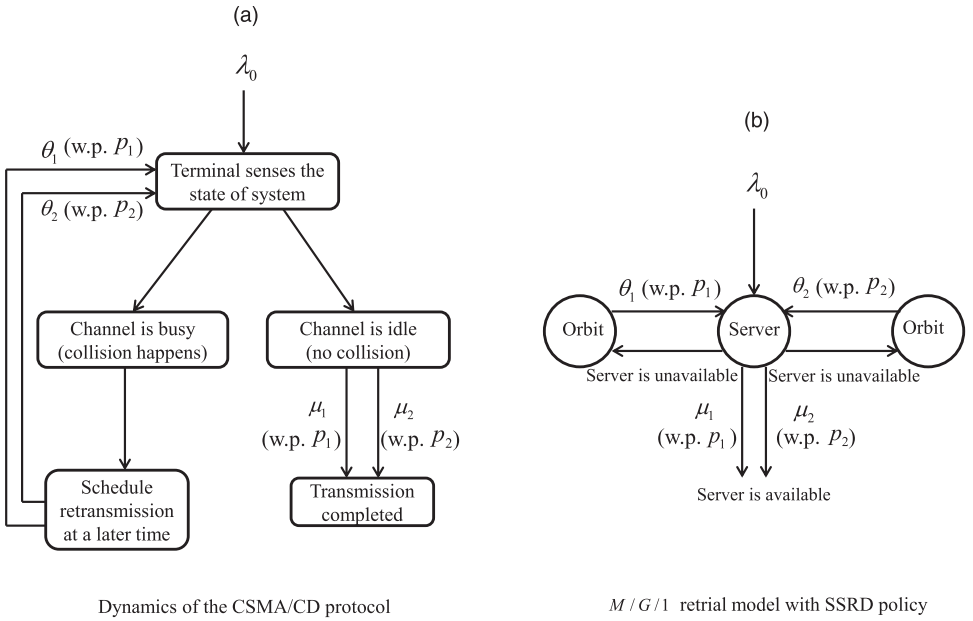
In this section, we derive the dominance condition for the SSRD policy. The treatment of the mean delay under the SSRD policy requires computing the second moment of the service time and the mean residual workload in service, which are given by

$$E[S^2] = \frac{c_v^2 + 1}{2} \beta_2 \quad \text{and} \quad \sum_{i=1}^m \rho_i \frac{E[S_i^2]}{2E[S_i]} = \frac{\lambda_0 E[S^2]}{2}$$
$$= \frac{\lambda_0 \beta_2 (c_v^2 + 1)}{4}, \quad \text{with} \quad \beta_2 \equiv \sum_{i=1}^m \frac{2p_i}{\mu_i^2}.$$

Let the traffic intensity be $\rho \equiv \lambda_0/\mu_0$. We assume for the rest of the paper that $\rho < 1$ (note that the system is stable if and only if $\rho < 1$ regardless of the value of θ_0; see Falin and Templeton 1997). According to Falin and Templeton (1997), the steady-state mean number of customers of grade i in the orbit queue is

$$N_i \equiv \frac{\lambda_i \rho}{\theta_i (1 - \rho)} + \frac{\lambda_i \lambda_0 \beta_2 (c_v^2 + 1)}{4} x_i, \quad i = 1, \ldots, m, \tag{3}$$

Figure 2. Dynamics of the CSMA/CD Protocol with SSRD Policy



where $\lambda_i \equiv \lambda_0 p_i$ is the arrival rate of customers of class i and $\mathbf{x} = (x_1, \ldots, x_m)^T$ is the unique solution to the linear equation

$$\mathbf{Ax} = -\mathbf{e} \qquad \text{or equivalently} \qquad \mathbf{x} = -\mathbf{A}^{-1}\mathbf{e}, \tag{4}$$

where $\mathbf{e} = (1, \ldots, 1)^T$, the matrix $\mathbf{A} = (A_{i,j})_{1 \le i,j \le m}$ has off-diagonal entries $A_{i,j} = a_{i,j}$ and diagonal entries $A_{j,j} = \sum_{i=1}^{m} a_{j,i} + a_{j,j} - 1, a_{i,j} = \theta_j \rho_j / (\theta_i + \theta_j)$ for $1 \le i, j \le m$, and $\rho_i \equiv \lambda_i / \mu_i$ is the traffic intensity for the i^{th} grade. We next compute the expected waiting time before a customer enters service (i.e., the expected total orbit time). Unlike in traditional work-conserving queueing systems, an idle server and waiting (orbiting) customers may coexist in retrial models. During the total waiting time w_i of a customer of class i (i.e., the total time the customer spends orbiting before entering service), the state of the server alternates between busy and idle. Let w_i^I (w_i^B) represent the expected total waiting time of a customer of class i when the server is idle (busy). Little's law and (3) imply that the expected waiting time for a customer of grade i is

$$w_i = \frac{\rho}{\theta_i(1-\rho)} + \frac{\lambda_0 \beta_2 (c_v^2 + 1)}{4} x_i = w_i^I + w_i^B, \tag{5}$$

where the second equality follows from a similar analysis for a retrial model with homogeneous service (theorem 1 in Artalejo and Falin (1994)). Because $\lambda_0 \beta_2 (c_v^2 + 1)/4$ is the mean remaining workload in service, x_i represents the expected number of service completions for which a customer of class i needs to wait before entering service. In addition, the first term $w_i^I \to 0$ as the retrial rate $\theta_i \to \infty$ (continuous retrial for service).

Now, we are ready to derive the conditions under which the two-grade SSRD policy ($m = 2$) dominates the homogeneous service policy ($m = 1$) (note that the two-grade SSRD policy is a special case of the general m-grade SSRD policy with $m \ge 2$) with the base model parameters λ_0, μ_0, and θ_0 held fixed. Because the expected service time $1/\mu_0$ and arrival rate λ_0 are fixed, the SSRD and homogeneous service cases have the same traffic intensity $\rho = \lambda_0/\mu_0$, which is also the probability that the server is busy. For the case of homogeneous service, the average waiting time is

$$w_0 = w_0^I + w_0^B, \quad \text{where} \quad w_0^I \equiv \frac{\rho}{\theta_0(1-\rho)} \quad \text{and}$$
$$w_0^B \equiv \frac{\rho(c_v^2 + 1)}{2\mu_0(1-\rho)}. \tag{6}$$

For two-grade SSRD, solving (4) and (5) yields the expected delays for grade 1 and grade 2 customers:

$$w_1 = \frac{\rho}{\theta_1(1-\rho)} + \frac{\lambda_0 \beta_2 (c_v^2 + 1)}{4(1-\rho)} \cdot \frac{(1-\rho)\theta_1 + \theta_2}{(1-\rho_1)\theta_1 + (1-\rho_2)\theta_2}, \tag{7}$$

$$w_2 = \frac{\rho}{\theta_2(1-\rho)} + \frac{\lambda_0 \beta_2 (c_v^2 + 1)}{4(1-\rho)} \cdot \frac{\theta_1 + (1-\rho)\theta_2}{(1-\rho_1)\theta_1 + (1-\rho_2)\theta_2}. \tag{8}$$

Hence, the overall expected delay is

$$w_{SSRD} \equiv p_1 w_1 + p_2 w_2 = \frac{\rho}{(1-\rho)\theta_0} + \frac{\lambda_0 \beta_2 (c_v^2 + 1)}{4(1-\rho)} \cdot \frac{\theta_1 + \theta_2 - \rho(p_1\theta_1 + p_2\theta_2)}{\theta_1 + \theta_2 - (\rho_1\theta_1 + \rho_2\theta_2)}. \tag{9}$$

A careful comparison of (6) and (9) leads to the following necessary and sufficient condition. The proof is given in EC.1.

Theorem 1 (Dominance of SSRD). *Considering an $M/G/1$ retrial model with fixed model parameters λ_0, θ_0, and μ_0, the SSRD policy with two service grades satisfying (2) results in a shorter waiting time than the homogeneous service policy does if and only if*

$$\text{i.}\quad 1 < \frac{\mu_2}{\mu_1} < \frac{1}{1-\rho} \qquad \textit{and}$$
$$\text{ii.}\quad \frac{\mu_2/\mu_1 + \rho - 1}{1-(1-\rho)\mu_2/\mu_1} < \frac{\theta_2}{\theta_1} < \infty. \tag{10}$$

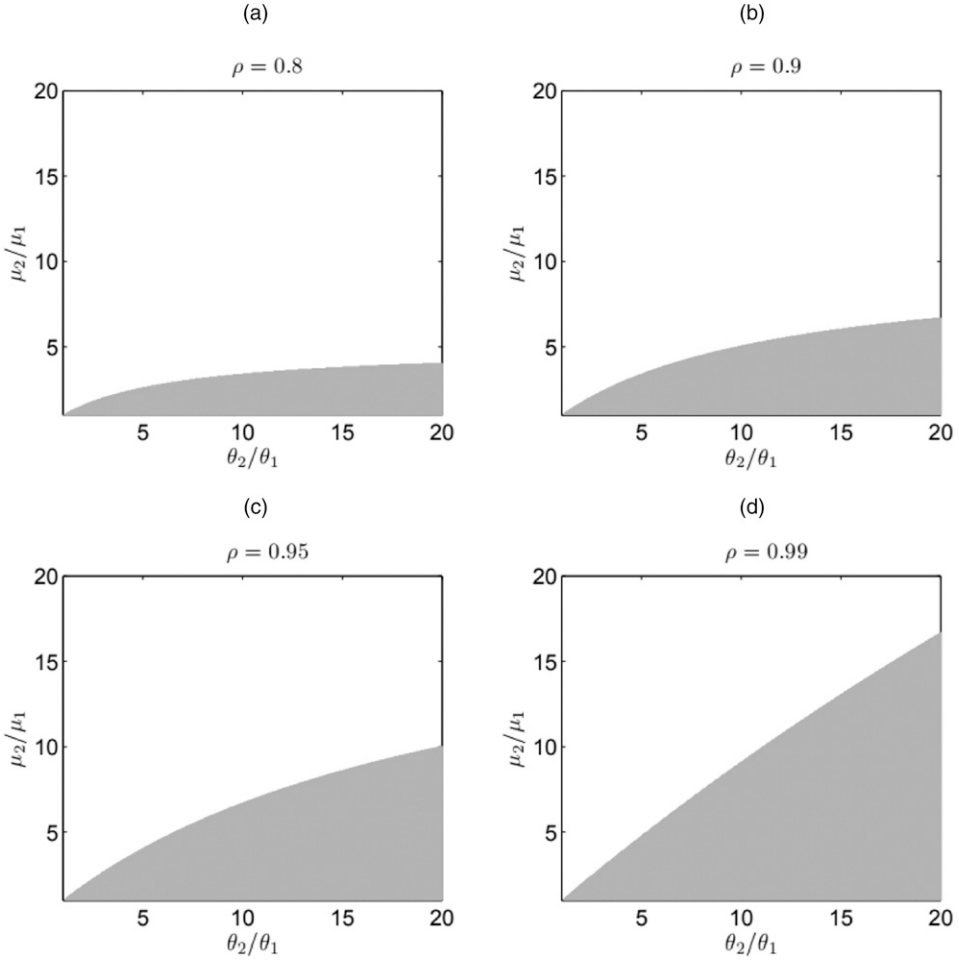
Remark 1 (Understanding Condition (10)). First, condition (i) of (10) gives an upper bound on the ratio μ_2/μ_1, which measures the level of service differentiation. It is evident that the difference in service level cannot be too large; otherwise, the service time variability (and, thus, the mean residual workload in service $\lambda_0\beta_2(c_v^2+1)/4$) would be too large to be compensated for. Because $1/(1-\rho) \to \infty$ as ρ increases, condition (i) becomes less restrictive in a more congested system (see Proposition 3 for the heavy-traffic analysis of the SSRD policy as $\rho \to 1$).

Next, condition (ii) of (10) indicates that, for a given service ratio μ_2/μ_1, the differentiation level of *customer priority* (measured by the ratio θ_2/θ_1) must be sufficiently high. Indeed, the SSRD policy achieves a smaller delay for the entire customer pool by assigning a significantly higher priority to customers with shorter service times. Because the lower bound in condition (ii) is increasing in μ_2/μ_1, a higher priority difference is required for a greater service difference.

Another interesting observation is that the conditions in (10) are independent of the allocation probabilities p_1 and p_2 (i.e., for any p_1 and p_2, we can always find suitable parameters $(\mu_1, \mu_2, \theta_1, \theta_2)$ such that (2) and (10) are satisfied). In addition, the conditions in (10) are insensitive to the service time distribution because the form of this distribution is preserved (although scaled) when the SSRD policy is adopted.

In Figure 3, we present an example to visualize the conditions in Theorem 1. For $\mu_0 = 1$, $\theta_0 = 1$, and $\lambda_0 = \mu_0\rho = \rho$, we plot the dominance regions characterized

Figure 3. Dominance Region of SSRD over Homogenous Service for an $M/G/1$ Retrial Model with $\lambda_0 = \rho$; $\mu_0 = \theta_0 = 1$; and $\rho = 0.8, 0.9, 0.95$ and 0.99



by the conditions in (10) as functions of μ_2/μ_1 and θ_2/θ_1 (see the shadowed areas) for traffic intensities $\rho = 0.8$, 0.9, 0.95, and 0.99. The dominance region expands as the traffic intensity increases. In particular, when the system is in heavy traffic (i.e., $\rho \to 1$), the conditions in (10) degenerate to the lower triangle of the first quadrant (i.e., the area below the 45° line $\theta_2/\theta_1 > \mu_2/\mu_1$) as seen in panel (d) of Figure 3.

Corollary 1 (Necessity of Simultaneously Differentiating Service and Retrial Rates)**.** *If $\mu_1 = \mu_2$, then the SSRD and homogeneous service policies achieve the same expected waiting time regardless of the values of θ_1 and θ_2. If $\theta_1 = \theta_2$, the homogeneous service policy dominates the SSRD policy.*

Remark 2 (Necessity of Simultaneous Differentiation of Service and Orbit Rates)**.** Corollary 1 shows that the benefit of SSRD is not gained when either the service rate or the retrial rate alone is differentiated but instead relies on the combined effect of differentiating both. It is straightforward to see that $w^I_{SSRD} \equiv p_1 w^I_1 + p_2 w^I_2 = w^I_0$ (this condition holds with any feasible SSRD parameters). Therefore, for SSRD to be beneficial, we need $w^B_{SSRD} \equiv p_1 w^B_1 + p_2 w^B_2 < w^B_0$. On the one hand, if the service rate is homogeneous, we have $\theta_1 + \theta_2 - \rho(p_1\theta_1 + p_2\theta_2) = \theta_1 + \theta_2 - (\rho_1\theta_1 + \rho_2\theta_2)$; in this case, SSRD does not help because w^B_{SSRD} is independent of θ_1 and θ_2 (see (9)). This explains why differentiating the orbit rate alone is not helpful. On the other hand, if the retrial rate is homogeneous, then w^B_{SSRD} increases because of the increased mean remaining workload in service, which is consistent with the conventional wisdom that higher service variability leads to excessive system congestion.

4. The Optimal SSRD Policy

In Section 3, we have shown that the SSRD policy, if correctly implemented, can help reduce the waiting time. In this section, we obtain the optimal SSRD parameters for achieving the maximum delay reduction. We do so in two steps. First, we show that it is sufficient to create two service grades; in other words, generating any additional k^{th} grade ($k = 3, \ldots, m$) does not further reduce the overall customer delay. Second, we compute the optimal parameters of the two-grade SSRD policy.

4.1. Road Map of the Main Steps

For the general case with $m \geq 3$, we begin by developing the optimal allocation probabilities $\mathbf{p}^*(C, \boldsymbol{\rho})$ for fixed values of $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ and $\mathbf{C} = (C_1, \ldots, C_m)$, where $\rho_i = \lambda_i/\mu_i$ is the traffic intensity for class i and $C_i = \theta_i/\theta_1$ is the *orbit rate differentiation ratio* (ODR), $i = 1, \ldots, m$; see Proposition 1. Next, using the conditionally optimal allocation probabilities $\mathbf{p}^*(\mathbf{C}, \boldsymbol{\rho})$, we derive the optimal workload $\boldsymbol{\rho}^*(\mathbf{C})$; see Theorem 2. Finally, we show that the expected delay decreases as the ODR increases; see Proposition 2. All proofs are provided in the e-companion.

Using the expected waiting time formula in (5), we now solve an optimization problem with the objective of minimizing the overall delay $w_{SSRD} = \sum_{i=1}^m w_i p_i$ subject to constraint (2). Namely we have

$$\begin{aligned} \min_{\boldsymbol{\mu}, \mathbf{p}} \quad & w_{SSRD} = \sum_{i=1}^m \left(\frac{\rho}{\theta_i(1-\rho)} + \frac{\lambda_0 \beta_2 (c_v^2 + 1)}{4} x_i \right) p_i \\ \text{s.t.} \quad & \sum_{i=1}^m \frac{p_i}{\mu_i} = \frac{1}{\mu_0}, \sum_{i=1}^m \frac{p_i}{\theta_i} = \frac{1}{\theta_0}, \sum_{i=1}^m p_i = 1 \\ & \mathbf{A}\mathbf{x} = -\mathbf{e}, \ \frac{\theta_i}{\theta_1} = C_i \\ & p_i, x_i \geq 0, i = 1, \ldots, m, \end{aligned} \tag{11}$$

where $\mathbf{A}$ is defined in (4). Note that c_v^2 is a constant for any given service time distribution. Therefore, by substituting $\boldsymbol{\rho}$ for $\boldsymbol{\mu}$ in (11) (i.e., $\mu_i = \lambda_0 p_i/\rho_i$, $\sum_{i=1}^m p_i/\mu_i^2 = \rho_i^2/p_i$) and using the equation $\sum_{i=1}^m p_i/\theta_i = 1/\theta_0$, we simplify (11) to

$$\begin{aligned} \min_{\boldsymbol{\rho}, \mathbf{p}} \quad & \left(\sum_{i=1}^m \frac{\rho_i^2}{p_i} \right) \cdot \left(\sum_{i=1}^m x_i p_i \right) \\ \text{s.t.} \quad & \sum_{i=1}^m \rho_i = \rho, \sum_{i=1}^m p_i = 1 \\ & \mathbf{A}\mathbf{x} = -\mathbf{e}, \frac{\theta_i}{\theta_1} = C_i \\ & p_i, x_i \geq 0, i = 1, \ldots, m. \end{aligned} \tag{12}$$

Because $\mathbf{x}$ can be determined from $\boldsymbol{\rho}$ and $\mathbf{C}$ and is irrelevant to $\mathbf{p}$, (12) can be solved directly using the Cauchy–Schwarz inequality

$$\sum_{i=1}^m a_i^2 \sum_{i=1}^m b_i^2 \geq \left(\sum_{i=1}^m a_i b_i \right)^2,$$

where $a_i = \rho_i/\sqrt{p_i}$ and $b_i = \sqrt{x_i p_i}$. In the following proposition, we present the optimal allocation probabilities conditional on $\boldsymbol{\rho}$ and $\mathbf{C}$.

Proposition 1 (m-Grade Optimal Allocation Probabilities Conditional on $\boldsymbol{\rho}$ and $\mathbf{C}$)**.** *Considering an $M/G/1$ retrial queue under the m-grade SSRD policy, for a given $\boldsymbol{\rho}$ and $\mathbf{C}$, the optimal allocation probabilities are*

$$p_i^*(\mathbf{C}, \boldsymbol{\rho}) = \frac{\rho_i/\sqrt{x_i}}{\rho_1/\sqrt{x_1} + \cdots + \rho_m/\sqrt{x_m}}, \qquad i = 1, \ldots, m, \tag{13}$$

where $\mathbf{x} = (x_1, \ldots, x_m)$ uniquely solves (4).

Computing the unconditional optimal parameters for m-grade SSRD is less straightforward than it is for

the two-grade case. Plugging $p^*(\mathbf{C}, \boldsymbol{\rho})$ in (13) into (12) yields

$$\begin{aligned} \min_{\rho} \quad & \sum_{i=1}^{m} \rho_i \sqrt{x_i} \\ \text{s.t.} \quad & \sum_{i=1}^{m} \rho_i = \rho < 1, \mathbf{A}\mathbf{x} = -\mathbf{e} \\ & \rho_i, x_i \geq 0, i = 1, \ldots, m. \end{aligned} \tag{14}$$

Under the assumption that $C_m = \theta_m/\theta_1$ is fixed, we show that allocating any customer to grade i such that $i = 2, \ldots, m-1$ is not beneficial to the system performance. In other words, it is optimal only to allocate the originally homogeneous customers to grades 1 and m, not to any grade in between (i.e., $\rho_2^* = \cdots = \rho_{m-1}^* = 0$). The following lemma serves as a building block for our main result.

Lemma 1. *For any m-grade SSRD workloads $\rho_1, \ldots, \rho_m$ and retrial rates $\theta_1 < \ldots < \theta_m$, the following statements hold:*

i. *The inverse of the matrix $\mathbf{A}$ is negative, that is, $\mathbf{A}^{-1} < 0$.*

ii. *The solution $\mathbf{x} = (x_1, \cdots, x_m)^T$ to $\mathbf{A}\mathbf{x} = -\mathbf{e}$ satisfies the following properties:*

a. *First, $x_i > 1$ for $i = 1, \ldots, m$.*

b. *Second, $x_1 > x_2 > \cdots > x_m$.*

c. *Last, $x_1 < C_2 x_2 < C_i x_i$ for $i = 3, \ldots, m$.*

Although no explicit expression for $\mathbf{x} = (x_1, \ldots, x_m)$ is available, Lemma 1 exhibits useful structural properties of $\mathbf{x}$, which facilitate the derivation of the optimal workload allocation for m-grade SSRD. First, x_i, the expected number of service completions before a customer of class i enters service, is strictly higher than one. Next, x_i is decreasing in i, which coincides with our intuition that class j has a higher priority than that of class k if $j > k$. According to (5), the mean number of trials for customers of class i is $\theta_i w_i = \rho/(1-\rho) + C_i x_i[\lambda_0 \beta_2 \theta_0 (c_v^2 + 1)/2]$. Therefore, part (ii.c) of Lemma 1 implies that a higher retrial rate results in a shorter waiting time but a larger number of trials, especially when $i \geq 3$. This observation seems to suggest that SSRD achieves a delay reduction at the expense of increasing the number of trials. (We further investigate the impact of SSRD on the total number of trials in the next section.)

Theorem 2 (Optimal Service-and-Retrial Differentiation Policy)**.** *For given traffic intensity ρ, retrial rate θ_0, and the condition of $1 < C_2 < \cdots < C_m$, we consider an $M/G/1$ retrial queue operated under the m-grade ($m \geq 3$) SSRD policy and have the following results.*

i. *The optimal m-grade SSRD degenerates to the two-grade SSRD, that is, we have that $p_i^*(\mathbf{C}, \rho) = \rho_i^*(\mathbf{C}, \rho) = 0$ for $i = 2, \ldots, m-1$.*

ii. *For grade-1 and grade-m customers, the optimal allocation policy is given as follows:*

The optimal probabilities are

$$\left(p_1^*(\mathbf{C}, \rho), p_m^*(\mathbf{C}, \rho)\right) \equiv \left(\frac{1}{1+r_p}, \frac{r_p}{1+r_p}\right). \tag{15}$$

The optimal workloads, service rates, and retrial rates are

$$\begin{aligned} &\left(\rho_1^*(\mathbf{C}, \rho), \rho_m^*(\mathbf{C}, \rho)\right) \equiv \left(\frac{\rho}{1+\sqrt{r_p}}, \frac{\rho\sqrt{r_p}}{1+\sqrt{r_p}}\right), \\ &\left(\mu_1^*(\mathbf{C}, \rho), \mu_m^*(\mathbf{C}, \rho)\right) \equiv \left(\mu_0\left(\frac{1+\sqrt{r_p}}{1+r_p}\right), \mu_0\left(\frac{r_p+\sqrt{r_p}}{1+r_p}\right)\right), \\ &\left(\theta_1^*(\mathbf{C}, \rho), \theta_m^*(\mathbf{C}, \rho)\right) \equiv \left(\theta_0\left(\frac{C_m + r_p}{(1+r_p)C_m}\right), \theta_0\left(\frac{C_m + r_p}{1+r_p}\right)\right). \end{aligned} \tag{16}$$

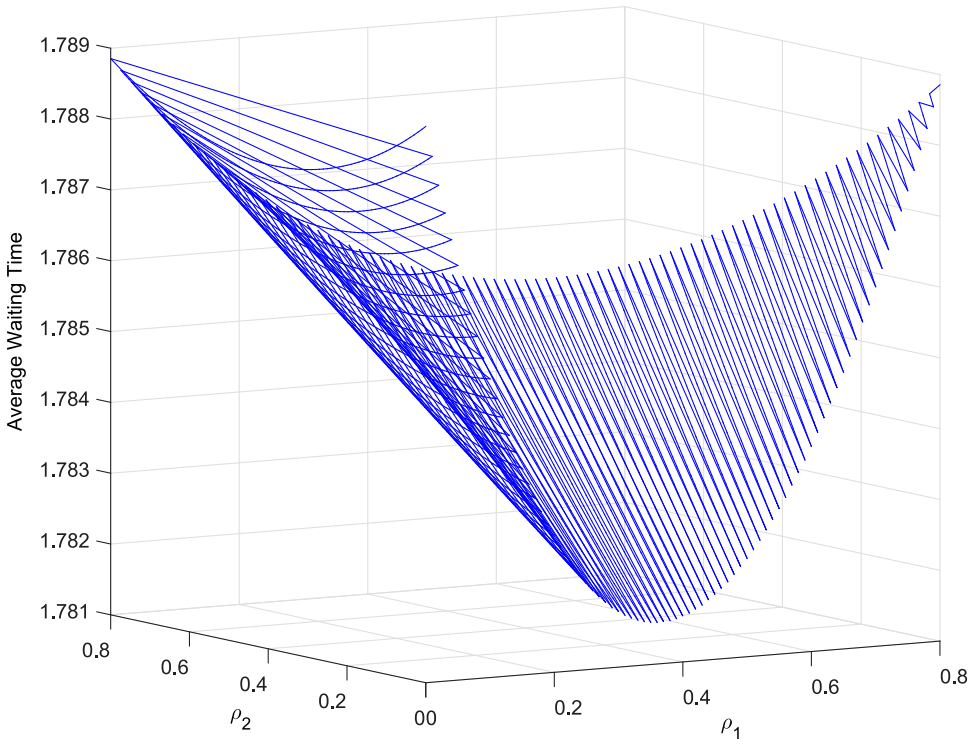
The expected waiting time under the optimal SSRD policy is

$$\begin{aligned} w^*(\mathbf{C}, \rho) &\equiv w_{SSRD}(\mathbf{p}^*, \boldsymbol{\mu}^*, \boldsymbol{\theta}^*, \mathbf{C}) \\ &= w_0^B \cdot \left(\frac{4\sqrt{r_p}}{(1+\sqrt{r_p})^2}\right) + w_0^I < w_0, \\ &\text{where} \qquad r_p \equiv \frac{p_m^*(\mathbf{C}, \rho)}{p_1^*(\mathbf{C}, \rho)} = \frac{1 + C_m - \rho}{1 + C_m - C_m \rho}. \end{aligned} \tag{17}$$

Remark 3 (The Optimal SSRD Structure Has Two Service Grades)**.** In the $M/G/1$ priority model described in Xu et al. (2015), increasing the number of service grades increases the level of differentiation; consequently, the optimal policy is one with infinite service grades. In contrast to Xu et al. (2015), Theorem 2 indicates that it is sufficient to differentiate customers into two groups because creating a third customer group does not improve system performance. In other words, given $m \geq 2$ service levels with retrial rates of $\theta_1 < \cdots < \theta_m$, it is optimal to assign all customers to either class 1 (with the lowest retrial rate) or class m (with the highest retrial rate) and none to any other class k, $2 \leq k \leq m-1$.

We present a numerical example with $m = 3$ to illustrate the results of Theorem 2. In Figure 4, we plot the average waiting time as a function of the workloads ρ_1 and ρ_2 with $m = 3$, $\rho = 0.8$, $C_1 = 1$, $C_2 = 2$, and $C_3 = 3$. Figure 4 shows that, for each fixed ρ_1, the waiting time increases in ρ_2, meaning that the minimum delay is attained when $\rho_2 = 0$. This observation confirms that only two grades are needed to achieve the minimum delay. See also Section EC.3 for additional discussions.

Figure 4. (Color online) The Impact of ρ_i on Average Waiting Time



Because the optimal SSRD policy has two grades, we let $C \equiv C_m = \theta_m/\theta_1$. From now on we restrict our attention to two-grade SSRD. It should be noted that the optimal SSRD parameters are independent of the service time SCV c_v^2 (they depend only on the service rate). In addition, the service time distribution beyond its first two moments (characterized by μ_0 and c_v^2) makes no impact on the optimal expected waiting time. We know from Corollary 1 that differentiating the service rate while leaving the retrial rate homogeneous results in a longer waiting time, that is, $w_{SSRD}(\mathbf{p}^*, \boldsymbol{\mu}^*, \theta_0, C) > w_0$. Next, we quantify this increase in the waiting time.

Corollary 2 (Quantifying the Influence of Differentiating the Retrial Rate)**.** *Under the SSRD policy with parameters* $(\mathbf{p}^*, \boldsymbol{\mu}^*, \boldsymbol{\theta}^*)$ *given in Theorem 2, we have*

$$w_0 - w_{SSRD}(\mathbf{p}^*, \boldsymbol{\mu}^*, \boldsymbol{\theta}^*, C) = w_{SSRD}(\mathbf{p}^*, \boldsymbol{\mu}^*, \theta_0, C) - w_0. \tag{18}$$

Corollary 2 supplements Corollary 1 to quantify the delay-saving effect of SSRD and the delay increase incurred when differentiating only the service rate with the case of homogeneous service being treated as the benchmark. Interestingly, the delay reduction under the optimal SSRD policy is equal to the delay increase when the service rate alone is differentiated (with the retrial rate remaining homogeneous). Corollary 2 suggests that there is a first-order benefit to differentiating the retrial rates, whereas the cost of differentiating the service rates is of second order. Next, we show that the delay-saving effect of SSRD in (18) can be further improved by increasing the ODR C.

Proposition 2 (Monotonicity in C and Asymptotic Limit)**.** *Consider the M/G/1 retrial queue under the optimal two-grade SSRD policy defined in Theorem 2.*

i. *For a fixed* $\rho < 1$, *the expected waiting time in* (17) *decreases in* C.

ii. *As* $C \to \infty$, *the limiting service and retrial rates, allocation probabilities, and mean delay are*

$$\begin{aligned}
&\left(\theta_1^*(\rho), \theta_2^*(\rho)\right) = \left(\frac{\theta_0(1-\rho)}{2-\rho}, \infty\right),\\
&\left(\mu_1^*(\rho), \mu_2^*(\rho)\right) = \left(\frac{\mu_0\left((1-\rho)+\sqrt{1-\rho}\right)}{2-\rho}, \frac{\mu_0\left(\sqrt{1-\rho}+1\right)}{2-\rho}\right),
\end{aligned} \tag{19}$$

$$\left(p_1^*(\rho), p_2^*(\rho)\right) = \left(\frac{1-\rho}{2-\rho}, \frac{1}{2-\rho}\right), \quad \text{and}$$

$$w^*(\rho) = \frac{\rho}{\theta_0(1-\rho)} + \frac{\rho\left(c_v^2+1\right)}{2\mu_0\sqrt{1-\rho}\left(\sqrt{1-\rho}+1\right)^2}. \tag{20}$$

Remark 4 (Further Increasing the Level of Differentiation)**.** As previously explained (see Theorem 2 and Remark 3), for given ratios $C_i = \theta_i/\theta_1$ $(1 \le i \le m)$, the two-grade case structurally "maximizes" the level of differentiation. Proposition 2 shows that, for the two-grade SSRD policy, increasing $C = \theta_2/\theta_1$ further increases the level

of differentiation. Unlike the policy considered in Xu et al. (2015), the SSRD policy controls operations in the $M/G/1$ retrial model based on a relative priority rule. In other words, although unlikely, it is possible that customers with lower priorities may be served before those with higher priorities. Because low-priority customers are assigned a lower service rate (and, thus, have a longer mean service time), their services cause longer waiting times for other customers. To prevent this from happening, we can further increase the differentiation level of customers' retrial rates such that most high-priority customers are served before low-priority customers. Indeed, according to (15), the probability $p_2^* = r_p/(1+r_p)$ is increasing in C; therefore, more customers are assigned the higher retrial rate. In the next section, we quantify the delay reduction achieved with SSRD (compared with the homogeneous case) in heavy traffic (as $\rho \to 1$).

Remark 5 (Additional Comparisons with Xu et al. (2015))**.** We compare our results with those of Xu et al. (2015) from the following perspectives.

a. *Relative priority versus absolute priority.* Unlike the absolute prioritization policy (i.e., SEPT) presented in Xu et al. (2015), the SSRD policy considered here is based on the creation of relative priorities (with $1 = C_1 < C_2 < \cdots < C_m$) because there is no guarantee that a class with a higher retrial rate is always served before those with lower retrial rates. Obtaining an absolute prioritization such as that considered in Xu et al. (2015) would require that $C_{i+1}/C_i \to \infty$ for all $1 \le i \le m-1$. In other words, customers with a higher service priority would need to be able to retry infinitely more frequently than those with a lower priority.

b. *Two-point versus continuous distribution.* It was shown in Xu et al. (2015) that it is beneficial to increase the number of service grades m. In fact, a closer look at proposition 4 in Xu et al. (2015) reveals that, as $m \to \infty$, it is optimal to offer every customer a continuously distributed *random service rate* with a finite support; see Proposition EC.3 in the e-companion of this manuscript for details. (In Proposition EC.3, we show that the continuous service rate has a simple linear probability density function.) In contrast, Theorem 2 and Proposition 2 show that, under the SSRD policy, it is optimal to offer every customer a two-point-distributed *random service rate*, with which the values of the two points and their probabilities are given by (19) and (20). In addition, the relative priority for class 2 customers now becomes an absolute priority because $\theta_2^*(\rho) = \infty$ (with the orbit queue becoming a waiting line in this limit).

c. *Monotonicity of variability.* Essentially, both the SSRD policy presented here and the differentiation policy presented in Xu et al. (2015) successfully achieve delay reduction by introducing additional variability into the service process, which increases the variance of the delay. We now compare our results with those of Xu et al. (2015) by computing the variance of the delay and studying its monotonicity with respect to the number of classes m and the ODR C; see Section EC.3 for the related numerical experiments. First, following section 4.1 of Kella and Yechiali (1988), we can compute the variance of the delay for the $M/G/1$ model under the SEPT policy as in Xu et al. (2015). Specifically, Figure EC.3 shows that the variance of the delay increases as the number of service grades m increases. Second, for our $M/G/1$ retrial model, we investigate the impact of m and C on the variance of the delay. Our numerical examples reveal that this variance is increasing in C (see Figure EC.5) and decreasing in m (see Figure EC.4). These findings provide some insight into why, for the SSRD policy, which achieves a low expected delay when the variance is high, it is more beneficial to increase the ODR C and not the number of grades m.

5. System Performance in Heavy Traffic

Theorem 2 suggests that (i) it is sufficient to differentiate customers into two groups and (ii) for a given $\rho < 1$, the overall expected waiting time is monotonically decreasing in the ODR C. In this section, we quantify the performance improvement when the system is in heavy traffic, that is, as $\rho \to 1$.

5.1. Reduction in Delay

We define the *relative reduction in delay* (RRD) with respect to the homogeneous service case as

$$R_D(C,\rho) \equiv \frac{w_0 - w^*(C,\rho)}{w_0}. \tag{21}$$

We define $h(x) = O(g(x))$ if $0 < K_1 < h(x)/g(x) < K_2 < \infty$ for all $x > x_0 > 0$. The following proposition gives the maximum value of the RRD as well as the asymptotic order of the waiting times in heavy traffic.

Proposition 3 (Maximum RRD in Heavy Traffic)**.** *Consider the $M/G/1$ retrial model.*

a. *The expected waiting time under the optimal SSRD policy can be expressed as*

$$w^*(C,\rho) = w_0^I + w_0^B\sqrt{1-\rho}\cdot\left(\frac{1}{\left(\sqrt{1-\rho}+1\right)^2} + O(1/C)\right),$$

where w_0^B and w_0^I are given in (6).

b. *The RRD $R_D(C,\rho)$ in (21) is increasing in both ρ and C. In addition,*

$$\lim_{\rho\to 1, C\to\infty} R_D(C,\rho) = \frac{c_v^2+1}{c_v^2+1+2\mu_0/\theta_0}.$$

Proposition 3 shows that SSRD can significantly reduce the waiting time. As the traffic intensity ρ increases, both $w^*(\rho, C)$ and w_0 increase. However, SSRD becomes more advantageous (achieving a larger RRD) when ρ and C are large. In particular, the first term of $w^*(C, \rho)$ is equal to w_0^I (the first term of w_0). However, the second term of $w^*(C, \rho)$ is $O(1/\sqrt{1-\rho})$ and, thus, grows more slowly than $w_0^B = O(1/(1-\rho))$ (the second term of w_0) does. In addition, the error term (which is $O(1/C)$) vanishes rapidly, meaning that SSRD quickly starts to outperform the homogeneous case as C increases. For a fixed service time distribution and c_v^2, when the system is in heavy traffic, the RRD R_D depends only on μ_0/θ_0; RRD increases when μ_0/θ_0 is small (i.e., θ_0 is large). In the extreme case of $\mu_0/\theta_0 \approx 0$ (i.e., θ_0 is large), $R_D \approx 100\%$. In addition, the heavy-traffic RRD increases as the service time SCV c_v^2 increases. We present numerical examples in Section 6 to demonstrate the behavior of RRD.

Remark 6 (The Optimal Two-Class Case in Xu et al. (2015))**.** The level of prioritization is measured by C. Indeed, class-2 customers receive an absolute priority as $C \to \infty$ (equivalent to forming a waiting line in front of the server). In this case, the delays are

$$w_1 = \frac{(c_v^2+1)(2-\rho)\rho}{(1-\rho)^{3/2}\left(\sqrt{1-\rho}+1\right)^2\mu_0} + \frac{(2-\rho)\rho}{(1-\rho)^2\theta_0} = w_1^B + w_1^I$$

$$\text{and} \quad w_2 = \frac{(c_v^2+1)(2-\rho)\rho}{\sqrt{1-\rho}\left(\sqrt{1-\rho}+1\right)^2\mu_0} = w_2^B.$$

Now, class-1 customers are the only orbiting customers with an orbit delay of w_1^I. Next, as $\theta_0 \to \infty$, we find that $w_1^I \to 0$ (such that no one is orbiting any longer) and $w_2/w_1 \to 1-\rho$. This behavior is consistent with the results of the optimal two-class policy reported in Xu et al. (2015).

5.2. Reduction in the Number of Trials

In a retrial queueing system, the number of trials a customer undergoes before service begins is another important metric of system congestion (Artalejo and Lopez-Herrero 2007). For the $M/G/1$ retrial queue with homogeneous service, the expected number of trials can be derived by applying Little's law and Wald's identity (Artalejo and Gómez-Corral 2008). In particular,

$$r_0 = w_0\theta_0 = \frac{\rho}{1-\rho}\left(\frac{1}{2\mu_0/\left(\theta_0(c_v^2+1)\right)}+1\right). \tag{22}$$

The general understanding of retrial models implies that the delay and number of trials are somewhat "negatively correlated." For example, when ρ is fixed, if the retrial rate θ_0 increases, the expected waiting time w_0 decreases (see (6)), and the expected number of trials before entering service r_0 increases (see (22)). However, we claim that SSRD cannot only shorten the average delay but also reduce the expected number of trials.

The mean number of trials under the optimal SSRD policy is

$$r^*(C,\rho) \equiv r_{SSRD}(\mathbf{p}^*, \boldsymbol{\mu}^*, \boldsymbol{\theta}^*, C) = w_1^*\theta_1^* p_1^* + w_2^*\theta_2^* p_2^*$$
$$= \frac{\rho^2\theta_0(c_v^2+1)}{(1-\rho)\lambda_0}\left(\frac{\sqrt{r_p}(C+r_p)(C+1)}{C(1+r_p)(1+\sqrt{r_p})^2}\right) + \frac{\rho}{1-\rho}. \tag{23}$$

Analogously to (21), we define the *relative reduction in the number of trials* (RRT) as

$$R_T(C,\rho) \equiv \frac{r_0 - r^*(C,\rho)}{r_0}. \tag{24}$$

Proposition 4 (Maximum RRT in Heavy Traffic)**.** *Consider the $M/G/1$ retrial model.*

a. *The RRT $R_T(C,\rho)$ in (24) is increasing in $\rho \in (0,1)$, and it is strictly positive when $\rho > \overline{\rho}_C$, where $0 < \overline{\rho}_C < 1$ is the unique solution to the following equation:*

$$C(1+r_p)^2 = 2\sqrt{r_p}(r_p + C^2),$$

where $r_p = (1+C-\rho)/(1+C-C\rho)$.

b. *In addition, if the limit $C^{2/3}\cdot(1-\rho) \to \vartheta \in [0,\infty]$ exists as $\rho \to 1$ and $C \to \infty$, then we have*

$$\lim_{\rho\to 1, C\to\infty} R_T(C,\rho) = \frac{1-2\sqrt{\vartheta}}{1+2\mu_0/\left[\theta_0(c_v^2+1)\right]}.$$

Unlike the RRD in (21), which is monotonic in both C and ρ, the RRT in (24) is not monotonic in C; therefore, its limit (if it exists) depends on the manner in which ρ and C converge to one and ∞, respectively. In particular, if $(1-\rho) \to 0$ faster than $C^{2/3} \to \infty$ (e.g., $C^{2/3}\cdot(1-\rho) \to \vartheta < 1/4$), then $R_T(C,\rho)$ is asymptotically positive. In fact, RRT is asymptotically maximal at $\vartheta = 0$. Otherwise, $R_T(C,\rho)$ can be asymptotically negative; that is, SSRD may result in a larger number of trials than that in the homogeneous service case.

5.3. Reduction in Slowdown

Compared with the sojourn time (the sum of the waiting time and service time), the *slowdown* is often considered a more practical metric for the congestion level because it measures the relative sojourn time normalized with respect to the service time. Specifically, the slowdown is defined as the ratio of the sojourn time to the service time; see Hyytiä et al. (2012) and Harchol-Balter (2013) for discussions of the slowdown.

Under the assumption that $E[S_0^{-1}] < \infty$, the slowdown in the homogeneous service case is

$$\gamma_0 = E\left[\frac{W_0 + S_0}{S_0}\right] = 1 + w_0 E[S_0^{-1}]$$
$$= 1 + \frac{\rho}{1-\rho}\left(\frac{c_v^2+1}{2\mu_0} + \frac{1}{\theta_0}\right)E[S_0^{-1}], \quad (25)$$

where the second equality holds because the delay is independent of the service time; see also equation 1 in Hyytiä et al. (2012). For two-grade SSRD, the slowdown is

$$\gamma_{SSRD} = E\left[\frac{W_1 + S_1}{S_1}\right]p_1 + E\left[\frac{W_2 + S_2}{S_2}\right]p_2$$
$$= 1 + \left(w_1 p_1 \frac{\mu_1}{\mu_0} + w_2 p_2 \frac{\mu_2}{\mu_0}\right)E[S_0^{-1}]. \quad (26)$$

It is easy to verify that, under the optimal SSRD policy given in Theorem 2,

$$\gamma^*(C,\rho) \equiv \gamma_{SSRD}(\mathbf{p}^*, \boldsymbol{\mu}^*, \boldsymbol{\theta}^*, C) = 1 + \left[\frac{\left(C + r_p\sqrt{r_p}\right)\left(1+\sqrt{r_p}\right)\rho}{(1+r_p)(1-\rho)\theta_0(C+r_p)} + \frac{(c_v^2+1)\rho\sqrt{r_p}}{(1-\rho)\mu_0(1+r_p)}\right] \cdot E[S_0^{-1}].$$

We define the *relative reduction in slowdown* (RRS) as follows:

$$R_S(C,\rho) = \frac{\gamma_0 - \gamma^*(C,\rho)}{\gamma_0}.$$

The following proposition gives the asymptotic upper bound on the RRS in heavy traffic.

Proposition 5 (Maximum RRS Under Heavy Traffic)**.** *Consider the $M/G/1$ retrial model. If the limit $(1-\rho)C \to \xi \in [0,\infty]$ exists as $\rho \to 1$ and $C \to \infty$, then the RRS satisfies*

$$\lim_{\rho\to 1, C\to\infty} R_S(C,\rho) = \frac{c_v^2 + 1 + \frac{2\mu_0}{\theta_0}\cdot\frac{\xi+1}{\xi+2}}{c_v^2 + 1 + \frac{2\mu_0}{\theta_0}}.$$

If $\xi = \infty$ (i.e., $C \to \infty$ faster than $\rho \to 1$), then $R_S \to 1 = 100\%$; if $\xi = 0$ (i.e., $\rho \to 1$ faster than $C \to \infty$), then $R_S \to c_v^2 + 1 + \mu_0/\theta_0 / c_v^2 + 1 + 2\mu_0/\theta_0 \equiv R_S^{\downarrow} < 1$. In addition, one can easily verify that, if the limit of $(1-\rho)C$ does not exist, then $100\% = \limsup_{\rho\to 1, C\to\infty} R_S > \liminf_{\rho\to 1, C\to\infty} R_S = R_S^{\downarrow}$. In other words, R_S is asymptotically guaranteed to be at least $R_S^{\downarrow} = c_v^2 + 1 + \mu_0/\theta_0 / c_v^2 + 1 + 2\mu_0/\theta_0$.

A closer look at the results in Propositions 3–5 reveals that all three performance measures can be maximized simultaneously if C is on the order of $1/(1-\rho)^\alpha$ for some appropriate α.

Corollary 3. *The RRD, RRT, and RRS can all be maximized as $C \to \infty$ and $\rho \to 1$ if $C = O(1/(1-\rho)^\alpha)$ for $\alpha \in (1, 3/2)$.*

6. Numerical Analysis

In this section, we provide numerical examples to evaluate the effectiveness of SSRD. We compare the performance of SSRD with that of homogeneous service. We conduct a sensitivity analysis of the optimal system performance with respect to various model parameters.

6.1. Expected Delay

We study the RRD R_D of the optimal SSRD policy (as defined in Theorem 2) with respect to the case of homogeneous service. In particular, we consider an $M/M/1$ base model with arrival rate λ_0, service rate μ_0, and retrial rate θ_0, that is, $c_v^2 = 1$. In Figure 5, we first plot the RRD as a function of the ORD C, $1 \le C \le 100$. For $\rho = 0.9$, we consider six cases: $\theta_0 = 1$, 2, and 10 with $\mu_0 = 1$ (panel (a)) and $\mu_0 = 1$, 2, and 10 with $\theta_0 = 1$ (panel (b)). Consistent with the results in Propositions 2 and 3, we observe that the RRD is monotonically increasing in C and decreasing in μ_0/θ_0. Indeed, the SSRD policy achieves a significant RRD.

Next, we compute and report the asymptotic RRD (as $C \to \infty$) as a function of the traffic intensity ρ for $0 < \rho < 1$. In the bottom panels of Figure 5, we consider $\mu_0 = 1$ with $\theta_0 = 1$, 2, and 10 (panel (c)) and $\theta_0 = 1$ with $\mu_0 = 1$, 2, and 10 (panel (d)). Figure 5 shows that the RRD is increasing in ρ and approaches its upper bound as $\rho \to 1$. In particular, when $\mu_0 = 1$ and $\rho \to 1$, the RRD approaches 50%, 67%, and 91% for $\theta_0 = 1$, 2, and 10, respectively. In contrast, when $\theta_0 = 1$, RRD approaches 50%, 30%, and 9% for $\mu_0 = 1$, 2, and 10, respectively. These findings are consistent with the results in Proposition 3, namely that the asymptotic upper bound on the RRD is $\theta_0/(\theta_0 + \mu_0)$ when $\rho \to 1$ and $c_v^2 = 1$. In heavy traffic, the RRD is decreasing in the service rate μ_0 and increasing in the retrial rate θ_0. If, in addition, $\theta_0 \to \infty$, the RRD approaches 100%.

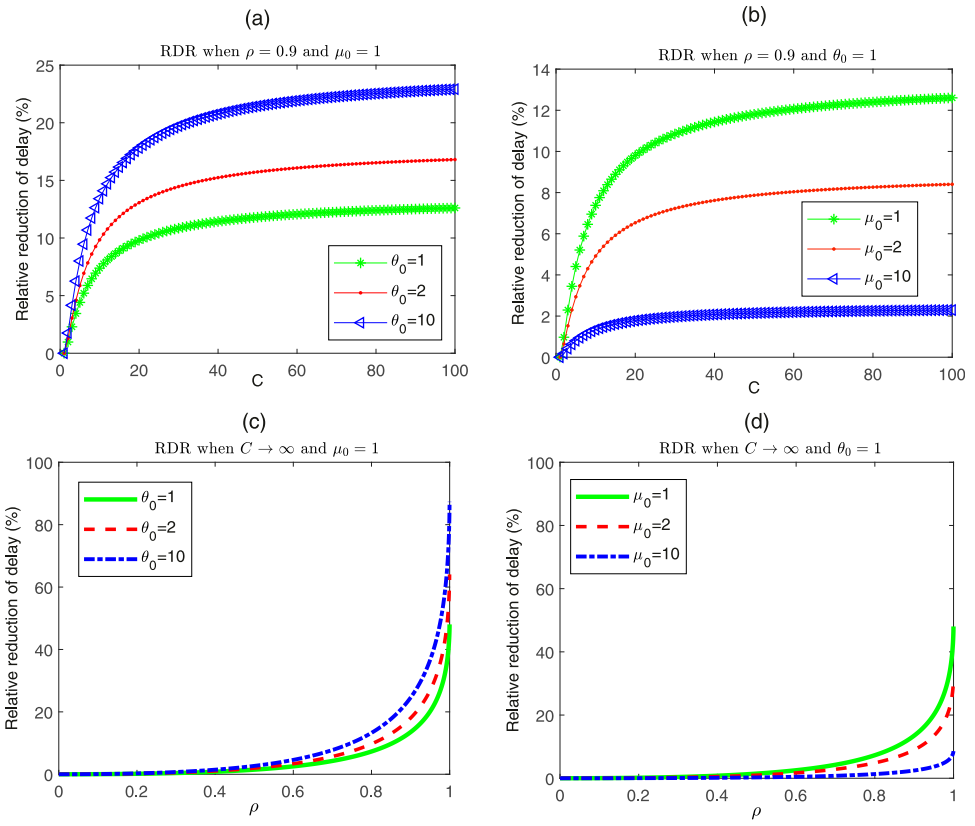
6.2. Expected Slowdown

We now consider the case of a lognormal service distribution $S_0 = e^{\hat{\mu}+\hat{\sigma}Z}$, where Z is a standard normal distribution. By varying the parameters $\hat{\mu}$ and $\hat{\sigma}$ while keeping the mean service time $E[S_0] = 1/\mu_0 = 1$ fixed, we plot the slowdown in Figure 6 as a function of ρ. We consider three cases of the service time SCV: (i) $c_v^2 = 0.5$ ($E[S_0^{-1}] = 1.5$), (ii) $c_v^2 = 1$ ($E[S_0^{-1}] = 2$), and (iii) $c_v^2 = 2$ ($E[S_0^{-1}] = 3$) with $\mu_0 = 1$ and $\theta_0 = 10$. Figure 6 shows that the SSRD policy significantly reduces the slowdown. In addition, the RRS is increasing in both $E[S_0^{-1}]$ and ρ. The upper bound of one is attained when $\rho \to 1$.

6.3. Expected Number of Trials

We next investigate the RRT performance under the optimal SSRD policy (Theorem 2) and compare it with the case of homogeneous service. In Figure 7, we first

Figure 5. (Color online) RRD R_D as a Function of C and ρ



plot the RRT as a function of the ORD C, $1 \le C \le 100$. For $\rho = 0.975$, we consider six cases: $\theta_0 = 1$, 2, and 10 with $\mu_0 = 1$ (panel (a)) and $\mu_0 = 1, 2$, and 10 with $\theta_0 = 1$ (panel (b)) as before. Figure 7 shows that the RRT is not monotonic in C. For a given traffic intensity (e.g., $\rho = 0.975$), the RRT first increases and later decreases, eventually becoming negative as C increases. Specifically, when $1 \le C \le 80$, the SSRD policy results in fewer trials. Consistent with the results in Proposition 4, for any $C > 1$, there exists a unique ρ' such that $R_T < 0$ when $\rho \in [0, \rho')$ and $R_T > 0$ when $\rho \in [\rho', 1)$; see panels (c) and (d) of Figure 7.

In Figure 8, we compare the mean number of trials in the case of homogeneous service (the straight line) with that in the case of SSRD with $\rho = 0.95, 0.96, 0.97$, and 0.98; $\mu_0 = \theta_0 = 1$; and $1 \le C \le 150$. Figure 8 shows that, as ρ increases toward one, SSRD outperforms homogeneous service (i.e., $r^*(C, \rho) < r_0$) over a wider range of C. In particular, when $\rho = 0.98$, SSRD achieves a smaller number of trials than the homogeneous service policy does when $C \in (1, 100)$. Consistent with Proposition 4, Figure 8 confirms that SSRD is effective in reducing both the delay and the number of trials when the system is under heavy traffic.

7. Extensions

We have shown that strategically differentiating the service and retrial rates can help reduce the overall customer delay. However, it is apparent that the SSRD policy achieves delay reduction at the expense of sacrificing fairness of the service among all customers; specifically, it significantly increases the waiting time and number of trials for certain customer classes. In this section, we discuss the limitations of SSRD and ways to address them.

7.1. Balancing the Delay and the Number of Trials

In Figure 8, we observe that SSRD can significantly increase the number of trials before a customer finally

Figure 6. (Color online) Reduction of Slowdown as a Function of ρ

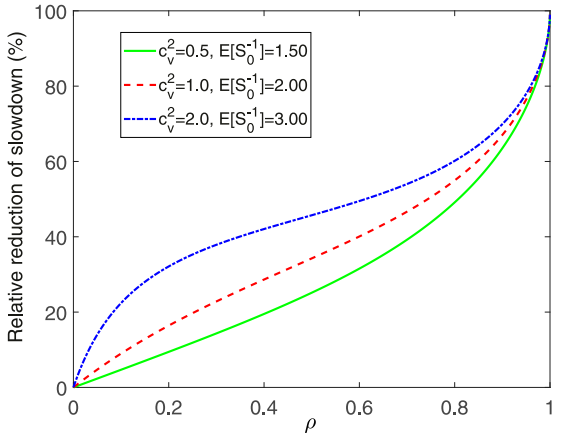
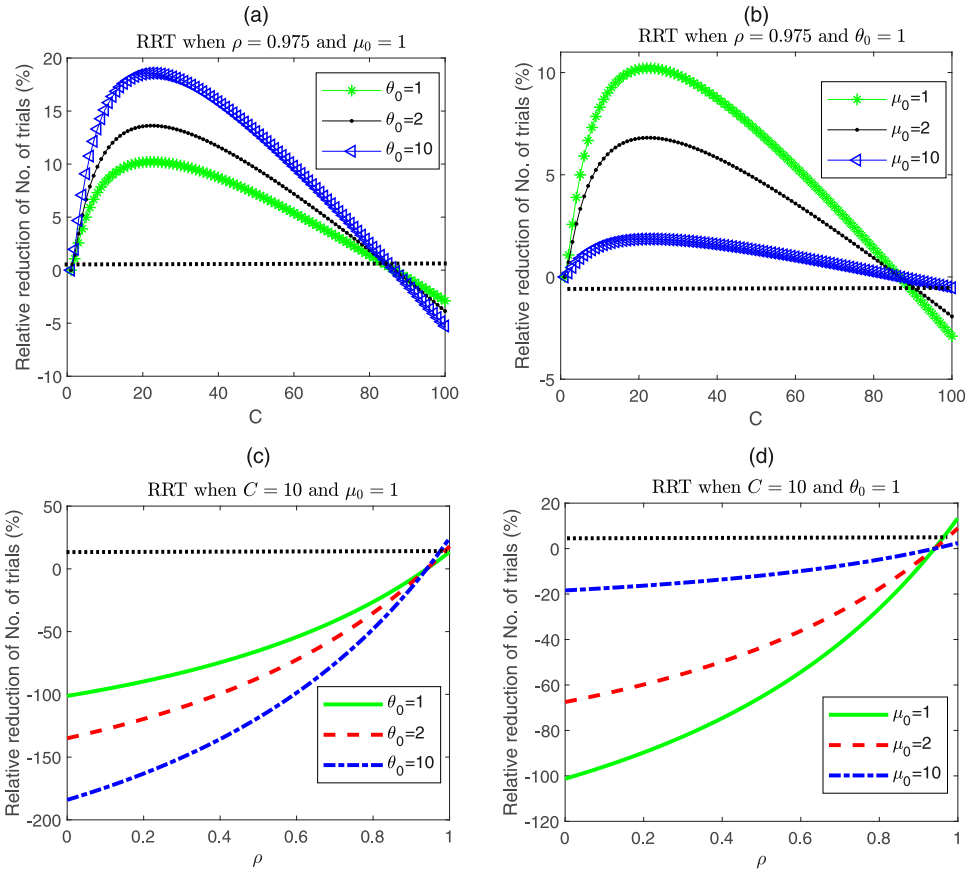


Figure 7. (Color online) RRT R_T as a Function of C and ρ



enters service when the system is not nearly critically loaded (i.e., when ρ is not close to one). Specifically, when $C = 100$, SSRD increases the average number of trials from 38 to 59 for $\rho = 0.95$ and from 48 to 65 for $\rho = 0.96$. Therefore, in practice, it may not always be optimal to choose a large C, especially from the viewpoint of minimizing the total number of trials. Motivated by Figure 8, we further explore the impact of SSRD on the number of trials by considering an alternative problem in which the objective is to minimize a weighted sum of the delay and the number of trials. To facilitate the analysis, we restrict our attention to the two-grade SSRD case. In particular, we consider the following problem:

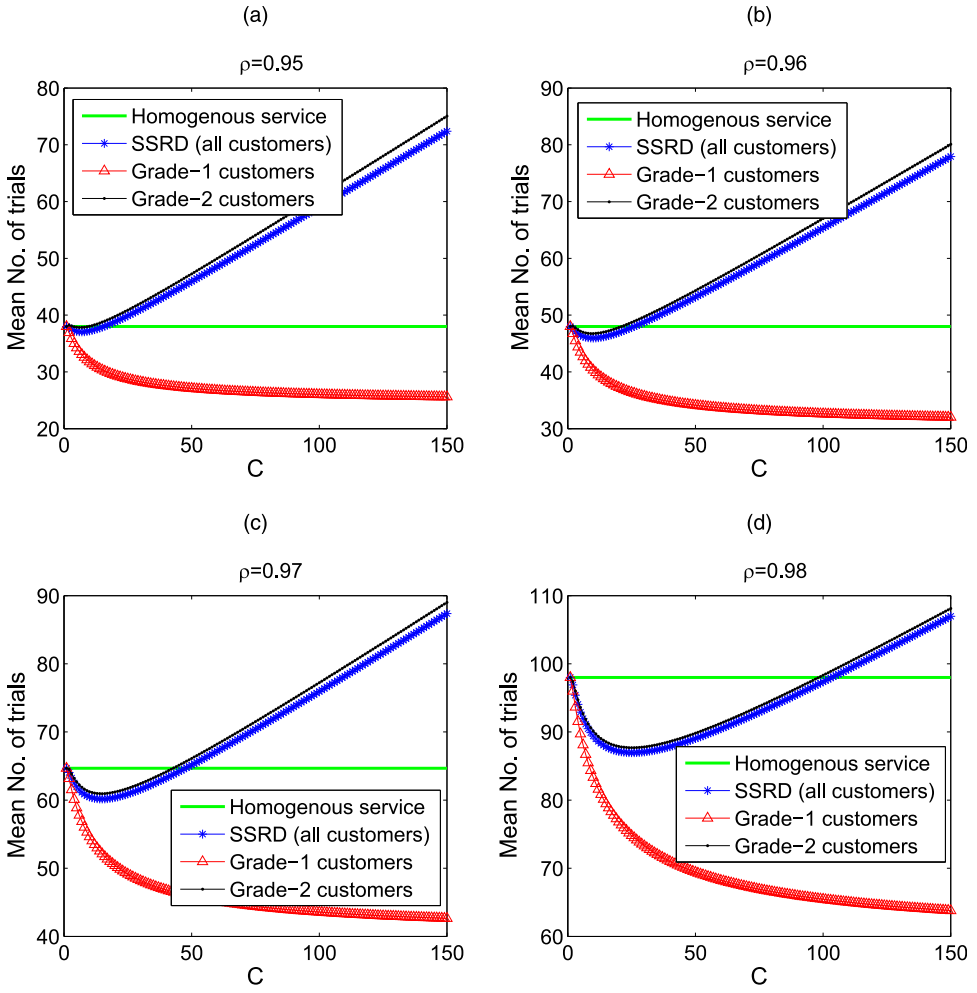
$$\begin{aligned}\min_{\boldsymbol{\mu},\mathbf{p}} \quad & p_1 w_1 + p_2 w_2 + (p_1 w_1 \theta_1 + p_2 w_2 \theta_2)\beta \\ \text{s.t.} \quad & \frac{p_1}{\mu_1} + \frac{p_2}{\mu_2} = \frac{1}{\mu_0}, \quad \frac{p_1}{\theta_1} + \frac{p_2}{\theta_2} = \frac{1}{\theta_0}, \quad p_1 + p_2 = 1, \\ & \frac{\theta_2}{\theta_1} = C, \end{aligned} \tag{27}$$

where the second term in the objective function is the average number of trials for all customers and the coefficient $\beta > 0$ is the normalized weight (cost) of the number of customer trials. When $\beta = 0$, this problem degenerates to our delay-based model. Plugging (7) and (8) into (27) yields the following equivalent problem:

$$\begin{aligned}\min_{\rho} \quad & \frac{\rho_1\sqrt{1-\rho+C}\sqrt{\theta_1\beta+1}}{\sqrt{1-\rho_1+(1-\rho_2)C}} + \frac{\rho_2\sqrt{1+(1-\rho)C}\sqrt{\theta_2\beta+1}}{\sqrt{1-\rho_1+(1-\rho_2)C}} \\ \text{s.t.} \quad & \frac{\rho_1\sqrt{1-\rho}\sqrt{1-\rho_1+(1-\rho_2)C}}{\theta_1\sqrt{\theta_1\beta+1}\sqrt{1-\rho+C}} + \frac{\rho_2\sqrt{1-\rho}\sqrt{1-\rho_1+(1-\rho_2)C}}{C\theta_1\sqrt{C\theta_1\beta+1}\sqrt{1+(1-\rho)C}} \\ & = \left(\frac{\rho_1\sqrt{1-\rho}\sqrt{1-\rho_1+(1-\rho_2)C}}{\sqrt{\theta_1\beta+1}\sqrt{1-\rho+C}} + \frac{\rho_2\sqrt{1-\rho}\sqrt{1-\rho_1+(1-\rho_2)C}}{\sqrt{C\theta_1\beta+1}\sqrt{1+(1-\rho)C}}\right)\frac{1}{\theta_0}, \rho_1+\rho_2 = \rho. \end{aligned} \tag{28}$$

The transformed problem (28) is a one-dimensional optimization problem. For a given $C \geq 1$, the optimal ρ_1^* and ρ_2^* can be numerically computed using a simple recursive algorithm (e.g., Newton's method).

Figure 8. (Color online) Mean Number of Trials as a Function of C When $\theta_0 = \mu_0 = 5$



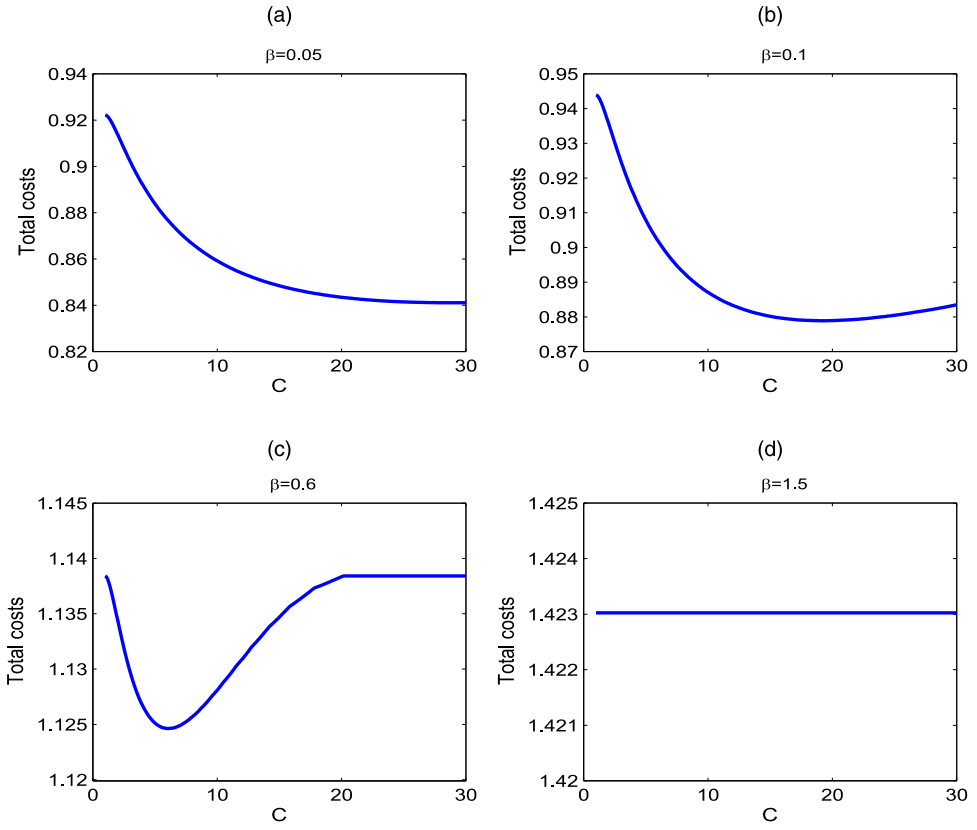
We present numerical results to illustrate the structure of the optimal policy. In Figure 9, we plot the optimal objective values for an $M/M/1$ retrial model as a function of the differentiation level $C \geq 1$ with $\lambda_0 = 0.8$, $\theta_0 = \mu_0 = 1$, and $0.05 \leq \beta \leq 1.5$. For each given β, the optimal cost is unimodal in C. The optimal differentiation levels are $C^* = 29.2$, 19.4, 6.2, and 1 when $\beta = 0.05$, 0.1, 0.6, and 1.5, respectively. In addition, the optimal cost becomes a constant if $C \geq C^{\uparrow}$, where $C^{\uparrow} > 0$ is some differentiation threshold (e.g., $C^{\uparrow} = 19.4$ when $\beta = 0.6$), indicating that SSRD is no longer beneficial to the system (because the optimal workload allocated to class 2 is $\rho_2^* = 0$ when $C > C^{\uparrow}$, meaning that the SSRD policy degenerates to the homogeneous case). Indeed, when $\beta = 1.5$ (Figure 9(d)), homogeneous service becomes the optimal policy. Figure 9 shows that service differentiation can help reduce the cost when β is small (i.e., when the number of trials makes an insignificant contribution to the objective function). However, for a large β, the rapid growth in the number of trials becomes too costly, outweighing the benefits of SSRD.

7.2. Convex Delay Cost

SSRD successfully reduces the overall customer delay at the cost of sacrificing service fairness across customer classes: while reducing the delay for high-priority classes, SSRD significantly increases the delay for other (low priority) classes. To account for fairness of service, we next extend our analysis to the case of a convex delay cost. Because a convex delay cost function more harshly penalizes longer delays, it helps reduce the variability of customer delay, thus maintaining a certain level of service fairness. See Guo and Zipkin (2007) and Mandelbaum and Stolyar (2004) for models with convex delay cost functions.

We consider a convex function $f(\cdot)$ that is nondecreasing and differentiable. When the changes of

Figure 9. (Color online) The Optimal Costs of an $M/M/1$ Retrial Model: $\lambda_0 = 0.8$; $\theta_0 = \mu_0 = 1$; and $\beta = 0.05, 0.1, 0.6, 1.5$



variables $p_1 = x$ and $\rho_1 = y$ are applied to (7) and (8), the mean waiting times for the two classes become

$$w_1(x,y) = \frac{1-\rho+C}{\lambda_0(1-\rho)(1-y+(1-\rho+y)C)} \cdot \left(\frac{y^2}{x} + \frac{(\rho-y)^2}{1-x}\right) + \frac{C\rho}{\theta_0(x(C-1)+1)(1-\rho)}, \tag{29}$$

$$w_2(x,y) = \frac{(1-\rho)C+1}{\lambda_0(1-\rho)(1-y+(1-\rho+y)C)} \cdot \left(\frac{y^2}{x} + \frac{(\rho-y)^2}{1-x}\right) + \frac{\rho}{\theta_0(x(C-1)+1)(1-\rho)}, \tag{30}$$

for a given $C \geq 1$. We now aim to minimize the overall delay cost

$$F(x,y,c) = p_1 f(w_1(x,y)) + p_2 f(w_2(x,y)). \tag{31}$$

Under SSRD, the optimal x^* and y^* that minimize $F(x,y)$ in (31) (if they exist) should necessarily satisfy the conditions

$$\frac{\partial F(x,y,C)}{\partial x} = \frac{\partial f(w_1(x,y))}{\partial x} x + \frac{\partial f(w_2(x,y))}{\partial x}(1-x) + f(w_1(x,y)) - f(w_2(x,y)) = 0, \tag{32}$$

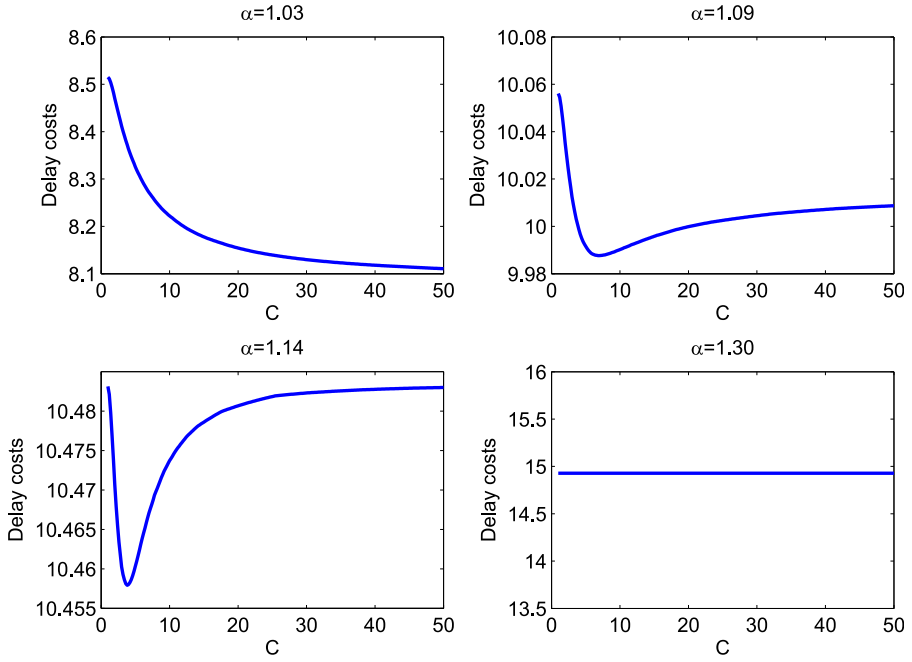
$$\frac{\partial F(x,y,C)}{\partial y} = \frac{\partial f(w_1(x,y))}{\partial y} x + \frac{\partial f(w_2(x,y))}{\partial y}(1-x) = 0, \tag{33}$$

where $0 \leq x^* \leq 1$ and $0 \leq y^* \leq \rho$. If such an x^* and y^* do not exist, the optimal performance is attained in the homogeneous case. For a given C, we define x_C and y_C as the minimizers of the delay cost.

Proposition 6 (Optimal C^* for a Convex Delay Cost)**.** *For an $M/M/1$ retrial queue under m-grade SSRD, a finite C^* that minimizes the delay cost in (31) can be found if there exists a $\overline{C}$ such that $x_C > (1-\rho)/(1+\rho)$ for $C > \overline{C}$.*

Proposition 6 is intuitive: it is no longer optimal to increase C to ∞ (as in Proposition 2) because the increased convex delay cost for those "sacrificed" customers quickly outweighs the benefits of SSRD. We test this result by considering an example with a convex delay cost function $f(x) = x^\alpha$, where the parameter $\alpha > 1$ quantifies the level of convexity (i.e., f is more convex when α is larger). Similar to Figure 9, Figure 10 shows the optimal delay cost as a function of C for $\alpha = 1.03$, 1.09, 1.14, and 1.30. According to Figure 10, the optimal C^* decreases when f is more convex (α increases). Indeed, when α is sufficiently large (e.g., $\alpha = 1.30$ in panel (d)), homogeneous

Figure 10. (Color online) The Optimal Convex Delay Cost for an $M/M/1$ Retrial Queue with $\lambda_0 = 0.8$; $\theta_0 = \mu_0 = 1$; and $\alpha = 1.03, 1.09, 1.14, 1.30$



service becomes the optimal policy, that is, SSRD is no longer beneficial.

7.3. Finite Waiting Buffer

We now extend our analysis to a more realistic $M/M/1/K$ retrial queue with a finite waiting buffer. Specifically, if an arriving customer finds a full waiting line (seeing K customers in the system), then the customer enters the orbit queue; otherwise, the customer enters the waiting line and be served later under the first-come first-served rule. See Figure 11 for an illustration. Retrial queues with a finite waiting buffer are quite complex and not well studied. In this section, we evaluate the performance of the two-grade SSRD policy in a retrial queue with a finite buffer: We first analyze the base model with homogeneous service. The performance in the homogeneous case is our benchmark. Next, we study the $M/M/1/K$ retrial queue under SSRD.

We consider a two-dimensional Markov process $\{(L(t), N(t)); t \geq 0\}$, where $L(t)$ and $N(t)$ are the numbers of customers in the main queue (in service and in queue) and in the orbit queue, respectively. Let $p_{(i,j)}$ denote the steady-state probability $P(L(\infty) = i, N(\infty) = j)$. Next, we give formulas for the queue length and delay.

Proposition 7. *Consider an $M/M/1/K$ retrial model with homogeneous service. The mean number of customers in the orbit queue is*

$$N_{orbit} = \frac{\mu_0(\Pi_1 - \rho_0(1-\rho_0))}{\theta_0(1-\rho_0)} + \sum_{i=1}^{K} \frac{(i\rho_0 - (i-1))\Pi_i}{1-\rho_0},$$

where Π_i is the steady-state probability that there are i customers in the waiting line. The mean customer delay is

$$w_0 = \frac{N_{orbit} + \sum_{i=1}^{K}(i/\mu)}{\lambda_0} - \frac{1}{\mu_0}.$$

An interesting observation is that N_{orbit} is determined only by the steady-state probabilities $\Pi_0, \Pi_1, \ldots, \Pi_K$ (independent of the orbit information). When $K = 1$, $N_{orbit} = (1/\theta_0 + 1/\mu_0)\lambda_0^2/(\mu_0 - \lambda_0)$, which coincides with the performance of our base retrial model with no waiting line. When $K \geq 2$, the probabilities $\Pi_i^* = \sum_{k=0}^{M} p_{(i,k)}$ can be numerically computed by imposing a (large) finite orbit capacity M. A numerical example with various values of N_{orbit} and w_0 is given in Table 1.

We next study the $M/M/1/K$ retrial model under the SSRD policy. The main difficulty here is the rapid increase in the number of states: when the buffer size is K, the number of states becomes 2^K (because we need to keep track of the number of customers in the queue and their types). We present the detailed

Figure 11. (Color online) The $M/M/1/K$ Retrial Model with SSRD

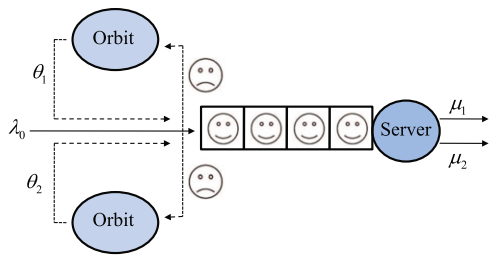


Table 1. The Mean Number of Customers in Orbit N_{orbit} with Truncated $M = 200$ when $\lambda_0 = 0.95$, $\theta_0 = \mu_0 = 1$

K	1	2	3	4	5	6
N_{orbit}	36.10	18.49	16.60	15.63	14.82	14.07
w_0	37.99	20.34	19.23	19.06	19.02	19.01
R_0	37.99	19.46	17.47	16.45	15.59	14.81

transition rate matrix in the appendix; see the proof of Proposition 8 in Secion EC.1 for more details.

Proposition 8. *For an $M/M/1/K$ retrial model, the steady-state probability that the server is serving a customer of type i is $\rho_i = \lambda_i/\mu_i$ for $i = 1, 2$, and $\rho_0 = \rho_1 + \rho_2$.*

Proposition 8 shows that the class-dependent workload is consistent with the case of $K = 1$; in other words, ρ_i is independent of the capacity K and the retrial rate. The total number of customers in the orbit queue is $N_{SSRD} = N_1 + N_2$, so the overall waiting time in the orbit queue is $w_{SSRD} = p_1[(N_1 + L_1)/\lambda_1 - 1/\mu_1] + p_2[(N_2 + L_2)/\lambda_2 - 1/\mu_2] = (N_{SSRD} + L_{SSRD})/\lambda_0 - 1/\mu_0$, where $L_{SSRD} = E[L(t)]$. The expected number of trials is $R_{SSRD} = p_1 N_1 \theta_1/\lambda_1 + p_2 N_2 \theta_2/\lambda_2 = (N_1\theta_1 + N_2\theta_2)/\lambda_0$. Here, the number of states is $(2^K + 1) \cdot (M + 1)^2$. Similar to the homogeneous case, we can derive the steady-state distribution under SSRD by truncating the states at some large M. We present an algorithm for computing the mean number of customers in the orbit queue (see Algorithm 1 in Section EC.1).

Remark 7 (Effectiveness of SSRD with a Finite Waiting Buffer)**.** Table 2 compares the impact of C on the system performance when $K = 1$, 2, 3 and $\rho = 0.95$. Here, the case of $C = 1$ represents the case of homogeneous service. We also show the relative reductions (abbreviated as "rel. red.") in all three performance metrics with respect to the homogeneous service case ($C = 1$). Table 2 shows that SSRD successfully reduces N_{SSRD} and w_{SSRD} and that a larger reduction is achieved as C increases. However, the number of trials R_{SSRD} is not necessarily monotonically decreasing in C: when $C < 6$, R_{SSRD} is decreasing in C while R_{SSRD} increases with C when $C > 6$. This observation is consistent with the case of $K = 1$ (no waiting line), see Figure 8. Table 2 shows that SSRD continues to be effective in reducing the overall customer delay for the $M/M/1/K$ retrial model with a finite waiting line. When $C = 12$, the RRD improves from 10.87% to 15.50% as K increases from one to three. In other words, in terms of the RRD, SSRD seems to be more beneficial as K increases. In the future, we plan to carefully study the benefit of SSRD in finite-buffer queueing models.

8. Conclusions

For a retrial queueing model, we show that the average waiting time can be significantly reduced by manually creating multiple service grades while leaving the total service capacity unchanged. Unlike dynamic service differentiation rules, our policy is static (independent of the current state of the system), so it is convenient to implement in practice. We show that the benefits of our policy can be attributed to the combined effects of service rate differentiation and orbit (retrial) rate differentiation, not to either one alone. In contrast to the results of Xu et al. (2015), which suggest that the performance can be improved by creating more grades, we show that it is sufficient to differentiate customers into two groups. We present numerical experiments conducted to evaluate the effectiveness of SSRD.

Table 2. The System Performance in OrbitN_{orbit} with Truncation $M = 120$ When $\lambda_0 = 0.95$, $\theta_0 = \mu_0 = 1$

	C	1	2	4	6	8	10	12
$K = 1$	N_{SSRD}	36.01	35.59	34.48	33.64	33.02	32.54	32.16
		rel. red.	1.16%	4.25%	6.58%	8.30%	9.64%	10.69%
	w_{SSRD}	37.99	37.46	36.29	35.41	34.75	34.25	33.86
		rel. red.	1.40%	4.48%	6.79%	8.53%	9.85%	10.87%
	R_{SSRD}	37.99	37.68	37.22	37.04	37.06	37.20	37.43
		rel. red.	0.82%	2.03%	2.52%	2.45%	2.08%	1.47%
$K = 2$	N_{SSRD}	18.49	18.17	17.33	16.72	16.29	15.98	15.75
		rel. red.	1.73%	6.27%	9.57%	11.90%	13.57%	14.82%
	w_{SSRD}	20.34	20.01	19.13	18.49	18.04	17.71	17.47
		rel. red.	1.62%	5.95%	9.10%	11.30%	12.93%	14.11%
	R_{SSRD}	19.46	19.33	19.10	19.12	19.30	19.57	19.91
		rel. red.	0.67%	1.85%	1.75%	0.82%	−0.57%	−2.31%
$K = 3$	N_{SSRD}	16.60	16.10	15.32	14.69	14.28	13.99	13.76
		rel. red.	3.01%	7.71%	11.51%	13.98%	15.72%	17.11%
	w_{SSRD}	19.23	18.70	17.88	17.22	16.79	16.49	16.25
		rel. red.	2.76%	7.02%	10.05%	12.69%	14.25%	15.50%
	R_{SSRD}	17.47	17.14	16.95	16.92	17.10	17.39	17.72
		rel. red.	1.89%	2.98%	3.15%	2.12%	0.46%	−1.43%

8.1. Future Directions

We have shown that the dominance condition (10) is independent of the structure of the service time distribution and depends only on the mean service time. Hence, we conjecture that condition (10) continues to hold for nonexponential orbit times. Another future direction of study is to investigate the benefits of service differentiation in multiserver queues, which have been proven to be more practical for modeling realistic service systems. See Section EC.4 for the preliminary numerical results of these extensions. The simulations presented there show that SSRD continues to help improve the system performance; see Table EC.2 for the results for an $M/H_2/1$ model with *two-phase hyperexponential* (H_2) service times (a mixture of two exponential distributions) and H_2 orbit times, and see Table EC.3 for the results for a two-server $M/M/2$ retrial queueing system.

8.2. Models with Relative Priorities

The idea of SSRD may also be applied to improve performance in other types of queueing models. One example is a queueing model with relative priority (Haviv and Van Der Wal 1997). Once service for a customer has been completed, the next customer to enter service is selected from among all waiting customers with probabilities that are proportional to their relative priority parameters. Suppose that there are m customer grades and that the customers of grade i arrive following an independent Poisson arrival process of rate λ_i, $1 \le i \le m$. Let their relative priority parameters be θ_i, $1 \le i \le m$. If there are n_j customers of grade j upon a service completion, then the probability that a customer of grade i is the next to enter service is

$$\frac{n_i\theta_i}{\sum_{j=1}^{m} n_j\theta_j}, \qquad 1 \le j \le m. \tag{34}$$

We remark that, in a relative priority queue, the next customer can be selected for service immediately upon a service completion although, in a retrial model, a waiting customer may have to wait until the customer's orbit clock expires even if the server is idle. In particular, when the retrial rate $\theta_0 \to \infty$, the performance of a retrial queue approaches that of a relative priority model. Then, the expected waiting time for a customer of grade i becomes

$$w_i^{rp} = \frac{\lambda_0\beta_2}{2}x_i, \tag{35}$$

where $\lambda_0 = \sum_{i=1}^{m}\lambda_i$, β_2, and x_i have been defined in Section 3. See Haviv and Van Der Wal (2007) for a more detailed discussion of models with relative priority.

Acknowledgments

The authors thank the editors and anonymous referees for providing many constructive comments, which have greatly improved the quality of the paper. They also thank NCSU PhD student Kyle Hovey for helping to proofread the paper.

References

Adusumilli KM, Hasenbein JJ (2010) Dynamic admission and service rate control of a queue. *Queueing Systems* 66(2):131–154.

Aissani A, Phung-Duc T (2015) Optimal analysis for *M/G/1* retrial queue with two-way communication. Gribaudo M, Manini D, Remke A, eds. *Analytical and Stochastic Modelling Techniques and Applications ASMTA 2015*, Lecture Notes in Computer Science, vol. 9081 (Springer, Cham, Switzerland), 1–14.

Anand KS, Pac MF, Veeraraghavan S (2011) Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Sci.* 57(1):40–56.

Artalejo JR (1997) Analysis of an *M/G/1* queue with constant repeated attempts and server vacations. *Comput. Oper. Res.* 24(6):493–504.

Artalejo JR, Falin JI (1994) Stochastic decomposition for retrial queues. *TOP* 2(2):329–342.

Artalejo JR, Gómez-Corral A (2008) *Retrial Queueing Systems: A Computational Approach* (Springer, Berlin).

Artalejo JR, Lopez-Herrero MJ (2007) On the distribution of the number of retrials. *Appl. Math. Model.* 31(3):478–489.

Artalejo JR, Krishnamoorthy A, Lopez-Herrero MJ (2006) Numerical analysis of (s, s) inventory systems with repeated attempts. *Ann. Oper. Res.* 141(1):67–83.

Ata B, Shneorson S (2006) Dynamic control of an *M/M/1* service system with adjustable arrival and service rates. *Management Sci.* 52(11):1778–1791.

Avrachenkov K, Morozov E, Nekrasova R (2015). Optimal and equilibrium retrial rates in single-server multi-orbit retrial systems. Jonsson M, Vinel A, Bellalta B, Tirkkonen O, eds. *Multiple Access Communications, MACOM 2015*, Lecture Notes in Computer Science, vol. 9305 (Springer, Cham, Switzerland), 135–146.

Choi BD, Shin YW, Ahn WC (1992) Retrial queues with collision arising from unslotted CSMA/CD protocol. *Queueing Systems* 11(4):335–356.

Debo LG, Toktay LB, Van Wassenhove LN (2008) Queuing for expert services. *Management Sci.* 54(8):1497–1512.

Ding S, Remerova M, van der Mei RD, Zwart B (2015) Fluid approximation of a call center model with redials and reconnects. *Performance Evaluation* 92(2):24–39.

Elcan A (1994) Optimal customer return rate for an *M/M/1* queueing system with retrials. *Probab. Engrg. Inform. Sci.* 8(4):521–539.

Elcan A (1999) Asymptotic bounds for an optimal state-dependent retrial rate of the *M/M/1* queue with returning customers. *Math. Comput. Model.* 30(3):129–140.

Falin G, Templeton J (1997) *Retrial Queues* (CRC Press, Boca Raton, FL).

Fayolle G, Gelenbe E, Labetoulle J (1977) Stability and optimal control of the packet switching broadcast channel. *J. ACM* 24(3):375–386.

George JM, Harrison JM (2001) Dynamic control of a queue with adjustable service rate. *Oper. Res.* 49(5):720–731.

Gharbi N, Dutheillet C, Ioualalen M (2009) Colored stochastic petri nets for modelling and analysis of multiclass retrial systems. *Math. Comput. Model.* 49(7–8):1436–1448.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Sci.* 53(6):962–970.

Harchol-Balter M (2013) *Performance Modeling and Design of Computer Systems: Queueing Theory in Action* (Cambridge University Press, New York).

Hassin R, Haviv M (1996) On optimal and equilibrium retrial rates in a queueing system. *Probab. Engrg. Inform. Sci.* 10(2):223–228.

Haviv M, Van Der Wal J (1997) Equilibrium strategies for processor sharing and random queues with relative priorities. *Probab. Engrg. Inform. Sci.* 11(4):403–412.

Haviv M, Van Der Wal J (2007) Waiting times in queues with relative priorities. *Oper. Res. Lett.* 35(5):591–594.

Hopp WJ, Iravani S, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Hyytiä E, Aalto S, Penttinen A (2012) Minimizing slowdown in heterogeneous size-aware dispatching systems. *ACM SIGMETRICS Performance Evaluation Review*, vol. 40 (ACM, London), 29–40.

Kella O, Yechiali U (1988) Priorities in *M/G/1* queue with server vacations. *Naval Res. Logist.* 35(1):23–34.

Li S, Ekici E, Shroff N (2015a) Throughput-optimal queue length based CSMA/CA algorithm for cognitive radio networks. *IEEE Trans. Mobile Comput.* 14(5):1098–1108.

Li S, Geng N, Xie X (2015b) Radiation queue: Meeting patient waiting time targets. *IEEE Trans. Robotics Automation Magazine* 22(2):51–63.

Liang HM, Kulkarni VG (1999) Optimal routing control in retrial queues. Shanthikumar JG, Sumita U, eds. *Applied Probability and Stochastic Processes*, International Series in Operations Research & Management Science, vol. 19 (Springer, Boston), 203–218.

Liu Y, Whitt Y (2014) Stabilizing performance in networks of queues with time-varying arrival rates. *Probab. Engrg. Inform. Sci.* 28(4): 419–449.

Liu Y, Whitt Y (2017) Stabilizing performance in a service system with time-varying arrivals and customer feedback. *Eur. J. Oper. Res.* 256(2):473–486.

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* 52(6):836–855.

Mandelbaum A, Massey WA, Reiman M, Stolyar A (1999) Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. *Proc. of the Thirty-Seventh Annual Allerton Conf. on Comm., Control and Comput., Allerton, IL*, 1095–1104.

Mandelbaum A, Massey WA, Reiman M, Rider B, Stolyar A (2002) Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecomm. Systems* 21(2–4), 149–171.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the *M/M/1* queue. *Oper. Res.* 38(5):870–883.

Ren ZJ, Zhou Y-P (2008) Call center outsourcing: Coordinating staffing level and service quality. *Management Sci.* 54(2):369–383.

Schrage LE, Miller LW (1966) The queue *M/G/1* with the shortest remaining processing time discipline. *Oper. Res.* 14(4):670–684.

Smith WE (1956) Various optimizers for single-stage production. *Naval Res. Logist. Quart.* 3(1–2):59–66.

Tobagi FA, Hunt VB (1980) Performance analysis of carrier sense multiple access with collision detection. *Comput. Networks* 4(5): 245–259.

Tran-Gia P, Mandjes M (1997) Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Trans. Selected Areas Comm.* 15(8):1406–1414.

Wang J, Cao J, Li Q (2001) Reliability analysis of the retrial queue with server breakdowns and repairs. *Queueing Systems* 38(4):363–380.

Wang J, Wang F, Li WW (2017) Strategic spectrum occupancy for secondary users in cognitive radio networks with retrials. *Naval Res. Logist.* 64(7):599–609.

Xu Y, Scheller-Wolf A, Sycara K (2015) The benefit of introducing variability in single-server queues with application to quality-based service domains. *Oper. Res.* 63(1):233–246.

Yom-Tov G, Mandelbaum A (2014) Erlang r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing Service Oper. Management* 16(2):283–299.

Jinting Wang is a professor in the school of management science and engineering at the Central University of Finance and Economics, Beijing, China. His research interests include stochastic modeling, queueing theory, reliability, inventory management, applied probability, and their applications in internet of things, wireless communications, call centers, healthcare, and operations management.

Zhongbin Wang is a postdoctoral fellow at the business school, Nankai University. His research interests include stochastic modeling, queueing theory, and game theory with applications to service and communication systems.

Yunan Liu is an associate professor in the Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests include queueing theory, stochastic modeling, applied probability, simulations, online learning, and their applications in call centers, healthcare, transportation, blockchain, and manufacturing systems.