E-Companion

Supplementing the main paper, in this appendix we give additional results. In §EC.1, we give proofs of all results in the main paper. In §EC.2, we consider the SSRD with preemptive service. In §EC.3, we give additional discussions on how our results compare to Xu et al. (2015). In §EC.4 we provide additional simulation results.

EC.1. Proofs

Proof of Theorem 1

According to the waiting time formulas (6) and (9), $w_{SSRD} < w_0$ is equivalent to

$$\left(\frac{p_1}{\mu_1^2}+\frac{p_2}{\mu_2^2}\right)\mu_0^2 < \frac{\theta_1+\theta_2-(\rho_1\theta_1+\rho_2\theta_2)}{\theta_1+\theta_2-\rho(p_1\theta_1+p_2\theta_2)}. \tag{EC.1}$$

Because $p_1/\mu_1+p_2/\mu_2=1/\mu_0$, the left-hand side of (EC.1)

$$\left(\frac{p_1}{\mu_1^2}+\frac{p_2}{\mu_2^2}\right)\mu_0^2 = 1+p_1p_2\left(1/\mu_1-1/\mu_2\right)^2\cdot\mu_0^2, \tag{EC.2}$$

and the right-hand side of (EC.2)

$$\frac{\theta_1+\theta_2-(\rho_1\theta_1+\rho_2\theta_2)}{\theta_1+\theta_2-\rho(p_1\theta_1+p_2\theta_2)} = 1+\frac{\lambda_0(\theta_2-\theta_1)p_1p_2(1/\mu_1-1/\mu_2)}{\theta_1+\theta_2-\rho(p_1\theta_1+p_2\theta_2)}. \tag{EC.3}$$

We require that the second term of (EC.2) is positive, which implies that $\mu_2>\mu_1$ when $\theta_2>\theta_1$. Combining (EC.1), (EC.2) and (EC.3) yields that

$$\frac{\lambda_0(\theta_2-\theta_1)}{\theta_1+\theta_2-\rho(p_1\theta_1+p_2\theta_2)} > (1/\mu_1-1/\mu_2)\cdot\mu_0^2,$$

or equivalently

$$\frac{\rho}{(p_2-p_1)(1-\rho)+(2-\rho)/\theta_0\left(\frac{1}{\theta_1}-\frac{1}{\theta_2}\right)} > \mu_0\left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right). \tag{EC.4}$$

Let $C_\theta\equiv\theta_2/\theta_1$ and $C_\mu\equiv\mu_2/\mu_1$, we have $\theta_1=\theta_0(p_1+p_2/C_\theta)$, $\mu_1=\mu_0(p_1+p_2/C_\mu)$. Plugging C_θ and C_μ into (EC.4), we can further algebraically simplify (EC.4) to

$$\frac{C_\theta+1-\rho}{(1-\rho)C_\theta+1} > C_\mu. \tag{EC.5}$$

It is noted that $\frac{C_\theta+1-\rho}{(1-\rho)C_\theta+1}$ is increasing in C_θ, which is upper bounded by $1/(1-\rho)$. Thus we have $C_\mu<1/(1-\rho)$. By noticing that $(1-\rho)C_\mu<1$, then (EC.5) can be transformed to

$$C_\theta > \frac{C_\mu+\rho-1}{1-(1-\rho)C_\mu} \quad \text{and} \quad C_\mu < \frac{1}{1-\rho},$$

which completes this proof. ■

Proof of Lemma 1

To prove part (i), we decompose $-\mathbf{A}$ into a summation of an identity matrix $\mathbf{I}$ and an matrix $\mathbf{B}$ as follow

$$-\mathbf{A}=\mathbf{I}+\begin{pmatrix} -\sum_{i=1}^{m} a_{1,i}-a_{1,1} & -a_{1,2} & \cdots & -a_{1,m} \\ -a_{2,1} & -\sum_{i=1}^{m} a_{2,i}-a_{2,2} & \cdots & -a_{2,m} \\ \vdots & \cdots & \ddots & \vdots \\ -a_{m,1} & a_{m,2} & \cdots & -\sum_{i=1}^{m} a_{m,i}-a_{m,m} \end{pmatrix}=\mathbf{I}+\mathbf{B},$$

where $\mathbf{I}$ is an m-dimensional identity matrix. Based on the definition of $a_{i,j}$, we have $-\sum_{i=1}^{m} a_{k,i}-a_{k,k}\in(-1,0)$ for $i,k=1,2,...,m$ and $-a_{i,j}\in(-1,0)$ for $i\neq j$, thus all elements of $\mathbf{B}$ are in the interval $(-1,0]$, then the inverse of $-\mathbf{A}$ can be expressed as

$$(-\mathbf{A})^{-1}=(\mathbf{I}+\mathbf{B})^{-1}=\mathbf{I}+\sum_{i=1}^{\infty}(-1)^{i}\mathbf{B}^{i}.$$

Because $B<0$, we have $(-1)^{i}\mathbf{B}^{i}>0$ for $i>1$. Therefore, $(-\mathbf{A})^{-1}=(\mathbf{I}+\mathbf{B})^{-1}>\mathbf{I}-\mathbf{B}$. Hence, we can obtain that $\mathbf{A}^{-1}=-(\mathbf{I}+\mathbf{B})^{-1}<-(\mathbf{I}-\mathbf{B})\leq-\mathbf{I}<0$. Notice that $\mathbf{x}=-\mathbf{A}^{-1}\mathbf{e}$, where $-x_i$ equals the summation of the elements of the i^{th} row of matrix $\mathbf{A}^{-1}$, in which $\mathbf{A}^{-1}\leq-\mathbf{I}$, thus the summation of the elements in each row of $\mathbf{A}^{-1}$ is smaller than -1, i.e., $-x_i<-1$ $(x_i>1)$ for $i=1,2,...,m$, which completes this proof.

(ii) Because $\mathbf{x}=-\mathbf{A}^{-1}\mathbf{e}$, the solutions $x_1,\ldots,x_m$ satisfy

$$\sum_{k=1}^{m}\frac{C_k\rho_k}{C_k+C_i}(x_i+x_k)=x_i-1 \quad \textit{and} \quad \sum_{k=1}^{m}\frac{C_k\rho_k}{C_k+C_j}(x_j+x_k)=x_j-1.$$

Without loss of generality, we assume $i<j$, so that

$$\begin{aligned} x_i-x_j &=\sum_{k=1}^{m}\frac{C_k\rho_k x_k}{C_k+C_i}-\frac{C_k\rho_k x_k}{C_k+C_j}+\sum_{k=1}^{m}\frac{C_k\rho_k x_i}{C_k+C_i}-\frac{C_k\rho_k x_2}{C_k+C_j} \\ &>\sum_{k=1}^{m}\frac{C_k\rho_k x_k}{C_k+C_i}-\frac{C_k\rho_k x_k}{C_k+C_j}+\sum_{k=1}^{m}\frac{C_k\rho_k x_i}{C_k+C_i}-\frac{C_k\rho_k x_j}{C_k+C_i}=A_0+B_0(x_i-x_j). \end{aligned} \tag{EC.6}$$

Hence we can obtain that $x_i-x_j>A_0/(1-B_0)$, where $A_0=\sum_{k=1}^{m}C_k\rho_k x_k/(C_k+C_i)-C_k\rho_k x_k/(C_k+C_j)$, $B_0=\sum_{k=1}^{m}C_k\rho_k/(C_k+C_i)$. Because $B_0<\sum_{k=1}^{m}\rho_k=\rho<1$ and $A_0>0$, it is obvious to see that $x_i-x_j>0$.

(iii) We assume $\rho_i>0$ for all $i=1,2,...,m$ (If $\rho_k=0$ for some k, the m-grade case degenerates to the $(m-1)$-grade case). Because $\mathbf{A}\mathbf{x}=\mathbf{e}$, multiplying C_i to the i^{th} row and dividing by C_j for the j^{th} column of $\mathbf{A}$ yields

$$\left(\begin{pmatrix} \rho_1 & \cdots & \rho_m \\ \vdots & \ddots & \vdots \\ \rho_1 & \cdots & \rho_m \end{pmatrix}+\begin{pmatrix} \sum_{i=1}^{m} a_{1,i}-1-a_{1,1} & -a_{1,2} & \cdots & -a_{1,m} \\ -a_{2,1} & \sum_{i=1}^{m} a_{2,i}-1-a_{2,2} & \cdots & -a_{2,m} \\ \vdots & \cdots & \ddots & \vdots \\ -a_{m,1} & -a_{m,2} & \cdots & \sum_{i=1}^{m} a_{m,i}-1-a_{m,m} \end{pmatrix}\right)\cdot\begin{pmatrix} x_1 \\ C_2x_2 \\ \vdots \\ C_mx_m \end{pmatrix}=\begin{pmatrix} -1 \\ -C_2 \\ \vdots \\ -C_m \end{pmatrix}. \tag{EC.7}$$

Omitting the last row (after some algebraic steps), we have

$$\begin{pmatrix} -b_{1,1} & b_{1,2} & \cdots & -b_{1,m} \\ -b_{2,1} & -b_{2,2} & \cdots & -b_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ -b_{m-1,1} & -b_{m-1,2} & \cdots & b_{m-1,m} \end{pmatrix} \begin{pmatrix} x_1 \\ C_2 x_2 \\ \vdots \\ C_m x_m \end{pmatrix} = \begin{pmatrix} C_2 - 1 \\ C_3 - C_2 \\ \vdots \\ C_m - C_{m-1} \end{pmatrix}, \tag{EC.8}$$

where $b_{i,j} \geq 0$ and

$$\begin{aligned} b_{k,i} &= \left(\frac{1}{C_k + C_i} - \frac{1}{C_{k+1} + C_i} \right) C_i \rho_i, \quad k \neq i, i-1; \\ b_{k,k} &= 1 - \left(\sum_{i=1}^m \frac{C_i \rho_i}{C_k + C_i} \right) + \left(\frac{1}{C_k + C_k} - \frac{1}{C_{k+1} + C_k} \right) C_k \rho_k, \quad k = 1, \ldots, m-1; \\ b_{k,k+1} &= \sum_{j \neq k+1} b_{k,j}, \quad k = 1, \ldots, m-1. \end{aligned} \tag{EC.9}$$

Because $C_{k+1} > C_k$ and $b_{k,k+1} = \sum_{j \neq k+1} b_{k,j}$ for $k = 1, \ldots, m-1$, (EC.8) implies that

$$\begin{aligned} &\sum_{j=1}^m b_{1,j} C_2 x_2 > \sum_{j=1}^m b_{1,j} C_j x_j, \\ &\sum_{j=1}^m b_{k,j} C_{k+1} x_{k+1} > \sum_{j=1}^m b_{k,j} C_j x_j, \quad k = 2, ..., m-2; \\ &\sum_{j=1}^m b_{m-1,j} C_m x_m > \sum_{j=1}^m b_{m-1,j} C_j x_j. \end{aligned}$$

It is easy to find that $C_k x_k$ $(k \geq 2)$ is not the smallest one among $x_1, C_2 x_2, \ldots, C_m x_m$, otherwise we have $\sum_{j=1}^m b_{k-1,j} C_k x_k \leq \sum_{j=1}^m b_{k-1,j} C_j x_j$, which contradicts to the inequalities above. Therefore, we must have $x_1 < C_i x_i$ $(i \geq 2)$. Next, we will prove $C_2 x_2 < C_i x_i$ for $i = 3, \ldots, m$ in a similar way. In (EC.8), dividing by $C_{i+1} - C_i$ in i^{th} row for $i = 1, \ldots, m-1$ and subtracting $t_i = b_{i,1}(C_2 - 1)/[b_{1,1}(C_{i+1} - C_i)]$ times of the first row in i^{th} row for $i = 2, \ldots, m-1$ leads to

$$\begin{pmatrix} -b'_{1,1} & b'_{1,2} & \cdots & -b'_{1,m} \\ 0 & -b'_{2,2} - t_2 b'_{1,2} & \cdots & -b'_{2,m} + t_2 b'_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & -b'_{m-1,2} - t_{m-1} b'_{1,2} & \cdots & b'_{m-1,m} + t_{m-1} b'_{1,m} \end{pmatrix} \begin{pmatrix} x_1 \\ C_2 x_2 \\ \vdots \\ C_m x_m \end{pmatrix} = \begin{pmatrix} 1 \\ 1 - t_2 \\ \vdots \\ 1 - t_{m-1} \end{pmatrix}, \tag{EC.10}$$

where $b'_{i,j} = b_{i,j}/(C_{i+1} - C_i)$. In order to find the relationships among $C_2 x_2, C_3 x_3 \ldots, C_m x_m$, we rewrite the above equations as

$$\begin{pmatrix} -C_{2,2} & C_{2,3} & \cdots & -C_{2,m} \\ -C_{3,2} & -C_{3,3} & \cdots & -C_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{m-1,2} & -C_{m-1,3} & \cdots & C_{m-1,m} \end{pmatrix} \begin{pmatrix} C_2 x_2 \\ C_3 x_3 \\ \vdots \\ C_m x_m \end{pmatrix} = \begin{pmatrix} a_2 \\ a_3 \\ \vdots \\ a_{m-1} \end{pmatrix}, \tag{EC.11}$$

where

$$\begin{aligned}
&C_{k,2}=\frac{b_{1,1}b_{k,2}+b_{1,2}b_{k,1}}{(C_{k+1}-C_k)b_{1,1}}>0, \quad k=2,\dots,m-1;\\
&C_{k,k+1}=\frac{b_{1,1}b_{k,k+1}+b_{1,k+1}b_{k,1}}{(C_{k+1}-C_k)b_{1,1}}>0, \quad k=2,\dots,m-1;\\
&C_{k,j}=\frac{b_{1,1}b_{kj}-b_{1,j}b_{k,1}}{(C_{k+1}-C_k)b_{1,1}}, \quad j\neq 2,k+1;\\
&a_k=1-t_k, \quad k=2,\dots,m-1.
\end{aligned} \tag{EC.12}$$

Based on the definition of $C_{i,j}$, it is easy to verify that $C_{k,k+1}=\sum_{j\neq k+1}C_{k,j}$ for $k=2,...,m-1$. The structure of (EC.11) is similar to (EC.8). If $a_k>0$ and $C_{k,j}>0$, we have

$$\begin{aligned}
&\sum_{j=2}^{m}C_{2,j}C_3x_3>\sum_{j=2}^{m}b_{2,j}C_jx_j,\\
&\sum_{j=2}^{m}C_{k,j}C_{k+1}x_{k+1}>\sum_{j=2}^{m}b_{k,j}C_jx_j, \quad k=3,...,m-2;\\
&\sum_{j=2}^{m}C_{m-1,j}C_mx_m>\sum_{j=2}^{m}b_{m-1,j}C_jx_j.
\end{aligned}$$

Hence, we can deduce $C_2x_2<C_ix_i$ for $i=3,...,m$. Therefore, it is sufficient to complete the proof by showing that $a_k>0$ and $C_{k,j}>0$.

Because $(C_2-1)/[(C_{k+1}+1)(C_k+1)]<1/(C_2+1)$, we have

$$b_{1,1}-\frac{b_{k,1}(C_2-1)}{C_{k+1}-C_k}>1-\left(\rho_1+\sum_{i=2}^{m}\frac{C_i}{1+C_i}\rho_i\right)>1-\rho>0, \tag{EC.13}$$

which implies that $a_k>0$ for $k=2,...,m-1$.

Note that we have $C_{k,2}>0$ and $C_{k,k+1}>0$ from (EC.12), it remains to show that $C_{k,j}>0$ for $2\le k\le m-1$ and $j\notin\{2,k+1\}$, we consider the following cases:

(1) When $j=k$, we need to prove $b_{1,1}b_{k,k}>b_{1,k}b_{k,1}$, that is

$$\begin{aligned}
&\left(\frac{1}{1+C_k}-\frac{1}{C_2+C_k}\right)C_k\rho_k\left(\frac{1}{1+C_k}-\frac{1}{1+C_{k+1}}\right)\rho_1\\
&<\left[1-\sum_{i=1}^{m}\frac{C_i}{1+C_i}\rho_i+\left(\frac{1}{2}-\frac{1}{1+C_2}\right)\rho_1\right]\left[1-\sum_{i=1}^{m}\frac{C_i}{C_k+C_i}\rho_i+\left(\frac{1}{2}-\frac{C_k}{C_k+C_{k+1}}\right)\rho_k\right].
\end{aligned}$$

It is straightforward to verify that

$$\begin{aligned}
&1-\sum_{i=1}^{m}\frac{C_i}{1+C_i}\rho_i+\frac{\rho_1}{2}+\frac{C_k\rho_k}{C_k+1}>1-\sum_{i=1}^{m}\rho_i+\rho_1+\rho_k,\\
&1-\sum_{i=1}^{m}\frac{C_i}{C_k+C_i}\rho_i+\frac{\rho_k}{2}+\frac{\rho_1}{C_k+1}>1-\sum_{i=1}^{m}\rho_i+\rho_1+\rho_k.
\end{aligned}$$

By letting $Y \equiv 1-\rho+\rho_1+\rho_k$, we have

$$\left[1-\sum_{i=1}^m \frac{C_i}{1+C_i}\rho_i+\left(\frac{1}{2}-\frac{1}{1+C_2}\right)\rho_1\right]\left[1-\sum_{i=1}^m \frac{C_i}{C_k+C_i}\rho_i+\left(\frac{1}{2}-\frac{C_k}{C_k+C_{k+1}}\right)\rho_k\right]$$
$$>\left[Y-\left(\frac{\rho_1}{C_2+1}+\frac{C_k\rho_k}{1+C_k}\right)\right]\left[Y-\left(\frac{\rho_1}{C_k+1}+\frac{C_k\rho_k}{C_k+C_{k+1}}\right)\right]>\left[1-\left(\frac{x}{C_2+1}+\frac{C_k y}{1+C_k}\right)\right]^2 Y^2$$

and

$$\left(\frac{1}{1+C_k}-\frac{1}{C_2+C_k}\right)C_k\rho_k\left(\frac{1}{1+C_k}-\frac{1}{1+C_{k+1}}\right)\rho_1<\frac{C_k}{1+C_k}y\frac{x}{1+C_k}Y^2,$$

where $x=\rho_1/Y, y=\rho_k/Y$.

Therefore, it is sufficient to show that $[1-(2/x+\bar{a}y)]^2>\bar{a}(1-\bar{a})xy$, where $\bar{a}\equiv C_k/(C_k+1), \bar{a}\in(1/2,1)$, i.e. $2/x+\bar{a}y+\sqrt{\bar{a}(1-\bar{a})xy}<1$. Define $\phi(x)\equiv 2/x+\bar{a}(1-x)+\sqrt{\bar{a}(1-\bar{a})x(1-x)}>2/x+\bar{a}y+\sqrt{\bar{a}(1-\bar{a})xy}$ so that $\phi'(x)=[1/2-\bar{a}+(\sqrt{\bar{a}(1-\bar{a})}(1-2x))/(2\sqrt{x(1-x)})]$. Setting $\bar{A}\equiv(2\bar{a}-1)^2/[\bar{a}(1-\bar{a})]$, we find that (i) when $0<x<1/2-\sqrt{\bar{A}/(4\bar{A}+16)}$, $\phi'(x)>0$ and (ii) when $1/2-\sqrt{\bar{A}/(4\bar{A}+16)}<x<1$, $\phi'(x)<0$, then we derive that $2/x+\bar{a}y+\sqrt{\bar{a}(1-\bar{a})xy}<\phi(x)\le \max\phi(x)=f(1/2-\sqrt{\bar{A}/(4\bar{A}+16)})=(\bar{a}+1)/2<1$.

(2) When $j\neq k$, we need to prove $b_{1,1}b_{k,j}>b_{1,j}b_{k,1}$, that is

$$\frac{C_2-1}{(C_k+1)(C_{k+1}+1)}\frac{(C_{k+1}+C_j)(C_k+C_j)}{(1+C_j)(C_2+C_j)}\rho_1<1-\sum_{i=1}^m\frac{C_i}{1+C_i}\rho_i+\left(\frac{1}{2}-\frac{1}{1+C_2}\right)\rho_1.$$

Define $\Psi(x)=(C_{k+1}+x)(C_k+x)/[(C_{k+1}+1)(C_k+1)]-(1+x)(C_2+x)/(C_2+1)$. We have

$$\Psi'(x)=2x\left(\frac{1}{(C_{k+1}+1)(C_k+1)}-\frac{1}{C_2+1}\right)+\frac{C_k+C_{k+1}}{(C_{k+1}+1)(C_k+1)}-1<0$$

for all $x\ge 1$. Then $\Psi(x)$ is decreasing in $x\ge 1$. Notice that $\Psi(1)=-1$, then $\Psi(x)<0$ for $x\ge 1$, which gives that $\Psi(C_j)<0$ for $j\neq k$. Then we can get that

$$\frac{\Psi(C_j)}{(1+C_j)(C_2+C_j)}<0 \quad\Leftrightarrow\quad \frac{1}{(C_k+1)(C_{k+1}+1)}\frac{(C_{k+1}+C_j)(C_k+C_j)}{(1+C_j)(C_2+C_j)}<\frac{1}{C_2+1},$$
$$\Leftrightarrow \frac{C_2}{(C_k+1)(C_{k+1}+1)}\frac{(C_{k+1}+C_j)(C_k+C_j)}{(1+C_j)(C_2+C_j)}<\frac{C_2}{C_2+1},$$

which implies that

$$\frac{C_2-1}{(C_k+1)(C_{k+1}+1)}\frac{(C_{k+1}+C_j)(C_k+C_j)}{(1+C_j)(C_2+C_j)}+\frac{1}{C_2+1}<1.$$

Therefore, we have

$$1-\sum_{i=1}^m\frac{C_i}{1+C_i}\rho_i+\left(\frac{1}{2}-\frac{1}{1+C_2}\right)\rho_1-\frac{C_2-1}{(C_k+1)(C_{k+1}+1)}\frac{(C_{k+1}+C_j)(C_k+C_j)}{(1+C_j)(C_2+C_j)}\rho_1$$
$$>1-\rho_1-\sum_{i=2}^m\frac{C_i}{1+C_i}\rho_i>1-\rho>0,$$

which completes our proof. ■

Proof of Theorem 2 Our proof has two steps. First, we show the sub-optimality of m-grade case; second, derive the optimal SSRD parameters for the 2-grade case.

Step 1: Sub-optimality of cases $m \geq 3$. The optimization problem (14) can be rewritten as

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} \rho_i \sqrt{x_i(\rho_1,\dots,\rho_m)} \\ s.t. \quad & \sum_{i=1}^{m} \rho_i = \rho < 1 \\ & \rho_i \geq 0, \quad i=1,2,\dots,m, \end{aligned} \tag{EC.14}$$

where $x_i(\rho_1,\dots,\rho_m)$ is a function of $(\rho_1,\dots,\rho_m)$ and it is the solution of the linear equation $\mathbf{A}\mathbf{x} = -\mathbf{e}$. We apply the first-order Kuhn-Tucker condition to obtain the optimal work load $\rho_1^*,\dots,\rho_m^*$. Let $\lambda \geq 0$, $\mu_i \geq 0$ for $i=1,\dots,m$ be the Lagrange multipliers (Luenberger and Ye 2008). The corresponding Lagrangian of (EC.14) is

$$L(\rho_1,\dots,\rho_m,\lambda,\mu_1,\dots,\mu_m) = \sum_{i=1}^{m} \rho_i \sqrt{x_i(\rho_1,\dots,\rho_m)} - \beta(\sum_{i=1}^{m}\rho_i - \rho) + \sum_{i=1}^{m} \alpha_i \rho_i. \tag{EC.15}$$

The Kuhn-Tucker condition implies that if the minimizers $\rho_1^*,\dots,\rho_m^*$ exist, there exist $\beta \geq 0$, $\alpha_i \geq 0, i=1,\dots,m$ such that

$$\frac{\partial L}{\partial \rho_i} = 0, \quad \alpha_i\rho_i = 0, \quad \rho_i,\alpha_i \geq 0, \quad i=1,\dots,n, \tag{EC.16}$$

which is equivalent to

$$\sqrt{x_i} + \frac{1}{2}\left(\frac{x_{1i}\rho_1}{\sqrt{x_1}} + \frac{x_{2i}\rho_2}{\sqrt{x_2}} + \cdots + \frac{x_{mi}\rho_m}{\sqrt{x_m}}\right) - \beta + \alpha_i = 0, \quad \alpha_i\rho_i = 0, \quad \rho_i,\alpha_i \geq 0, \quad i=1,\dots,m, \tag{EC.17}$$

where $x_{i,j} = \partial x_i / \partial \rho_j$, which solves the linear equation

$$\mathbf{A}\mathbf{b}_i = -\mathbf{B}_i\mathbf{x}, \tag{EC.18}$$

with $\mathbf{b}_i = (x_{i,1},\dots,x_{i,m})^T$, $\mathbf{e} = (1,\dots,1)^T$,

$$\mathbf{B}_1 = \begin{pmatrix} 2c_{1,1} & 0 & \cdots & 0 \\ c_{2,1} & c_{2,1} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & 0 & 0 & c_{m,1} \end{pmatrix}, \quad \dots, \quad \mathbf{B}_m = \begin{pmatrix} c_{1,m} & 0 & \cdots & c_{1,m} \\ 0 & c_{2,m} & 0 & c_{2,m} \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 2c_{m,m} \end{pmatrix}, \quad c_{i,j} = \frac{C_j}{C_i + C_j}. \tag{EC.19}$$

For all $\rho_i > 0$, we have $\mathbf{b}_i = \mathbf{A}^{-1}\mathbf{B}_i\mathbf{A}^{-1}\mathbf{e}$.

We first consider the case $m=3$ and later extend to the case of general $m\geq 3$ by induction. If the optimal $\rho_1^*,\rho_2^*,\rho_3^*$ are all strictly greater than 0, we find that $\alpha_i=0, i=1,2,3$ from Kunh-Tucker conditions. Then $\rho_1^*,\rho_2^*,\rho_3^*$ satisfy the following equations

$$\begin{aligned} f_1(\rho_1,\rho_2,\rho_3) &= 2(\sqrt{x_1}-\sqrt{x_2})-\mathbf{a}(\mathbf{b}_1-\mathbf{b}_2)=0,\\ f_2(\rho_1,\rho_2,\rho_3) &= 2(\sqrt{x_1}-\sqrt{x_3})-\mathbf{a}(\mathbf{b}_1-\mathbf{b}_3)=0, \end{aligned} \tag{EC.20}$$

where $\mathbf{a}=(\rho_1/\sqrt{x_1},\rho_2/\sqrt{x_2},\rho_3/\sqrt{x_3})$. Note that

$$\mathbf{b}_1-\mathbf{b}_2=\mathbf{A}^{-1}(\mathbf{B}_1-\mathbf{B}_2)\mathbf{A}^{-1}\mathbf{e}=\mathbf{A}^{-1}\begin{pmatrix} c_{2,1} & c_{2,1}-1 & 0\\ c_{2,1} & c_{2,1}-1 & 0\\ c_{3,1} & -c_{3,2} & c_{3,1}-c_{3,2}\end{pmatrix}\mathbf{A}^{-1}\begin{pmatrix}1\\1\\1\end{pmatrix}. \tag{EC.21}$$

Define the matrix

$$\mathbf{A}^{-1}=\begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3}\\ d_{2,1} & d_{2,2} & d_{2,3}\\ d_{3,1} & d_{3,2} & d_{3,3}\end{pmatrix},$$

where $d_{ij}\leq 0$ from (i) of Lemma 1. Because $\mathbf{A}\mathbf{x}=-\mathbf{e}$, we have

$$\begin{aligned} &\mathbf{A}^{-1}\begin{pmatrix} c_{2,1} & c_{2,1}-1 & 0\\ c_{2,1} & c_{2,1}-1 & 0\\ c_{3,1} & -c_{3,2} & c_{3,1}-c_{3,2}\end{pmatrix}\mathbf{A}^{-1}\begin{pmatrix}1\\1\\1\end{pmatrix}=-\mathbf{A}^{-1}\begin{pmatrix} c_{2,1} & c_{2,1}-1 & 0\\ c_{2,1} & c_{2,1}-1 & 0\\ c_{3,1} & -c_{3,2} & c_{3,1}-c_{3,2}\end{pmatrix}\mathbf{x}\\ &=-\begin{pmatrix} x_1[(d_{1,1}+d_{1,2})c_{2,1}+d_{1,3}c_{3,1}]-x_2[(d_{1,1}+d_{1,2})c_{1,2}+d_{1,3}c_{3,2}]+x_3(c_{3,1}-a_{3,2})d_{1,3}\\ x_1[(d_{2,1}+d_{2,2})c_{2,1}+d_{2,3}c_{3,1}]-x_2[(d_{2,1}+d_{2,2})c_{1,2}+d_{2,3}c_{3,2}]+x_3(c_{3,1}-a_{3,2})d_{2,3}\\ x_1[(d_{3,1}+d_{3,2})c_{2,1}+d_{3,3}c_{3,1}]-x_2[(d_{3,1}+d_{3,2})c_{1,2}+d_{3,3}c_{3,2}]+x_3(c_{3,1}-a_{3,2})d_{3,3}\end{pmatrix}. \end{aligned}$$

From (iii) of Lemma 1, we have $x_1c_{2,1}-x_2c_{1,2}=x_1/(C_2+1)-C_2x_2/(C_2+1)<0$, and

$$\begin{aligned} x_1c_{3,1}-x_2c_{3,2}+x_3c_{3,1}-x_3c_{3,2} &= \left(\frac{x_1}{C_3+1}-\frac{C_2x_2}{C_3+C_2}\right)-\left(\frac{C_3x_3}{C_3+1}-\frac{C_3x_3}{C_3+C_2}\right)\\ &< \left(\frac{C_2x_2}{C_3+1}-\frac{C_2x_2}{C_3+C_2}\right)-\left(\frac{C_3x_3}{C_3+1}-\frac{C_3x_3}{C_3+C_2}\right)\\ &= (C_2x_2-C_3x_3)\left(\frac{1}{C_3+1}-\frac{1}{C_3+C_2}\right)<0, \end{aligned}$$

where the inequalities hold because from $C_2x_2<C_3x_3$ (part (iii) of Lemma 1). Since $d_{i,j}\leq 0$, we have $(x_1c_{2,1}-x_2c_{1,2})(d_{i,1}+d_{i,2})+[x_1c_{3,1}-x_2c_{3,2}+x_3c_{3,1}-x_3c_{3,2}]d_{i,3}>0$ for $i=1,2,3$. Hence,

$$\mathbf{A}^{-1}\begin{pmatrix} c_{2,1} & c_{2,1}-1 & 0\\ c_{2,1} & c_{2,1}-1 & 0\\ c_{3,1} & -c_{3,2} & c_{3,1}-c_{3,2}\end{pmatrix}\mathbf{A}^{-1}\begin{pmatrix}1\\1\\1\end{pmatrix}<0. \tag{EC.22}$$

Therefore, we have $\mathbf{b}_1-\mathbf{b}_2<0$, leading to $\mathbf{a}(\mathbf{b}_1-\mathbf{b}_2)<0$ (note that $\mathbf{a}>0$). According to (ii) of Lemma 1, we have $x_1>x_2$, which implies that

$$f_1(\rho_1,\rho_2,\rho_3)=2(\sqrt{x_1}-\sqrt{x_2})-\mathbf{a}(\mathbf{b}_1-\mathbf{b}_2)>0,$$

which cannot satisfy the optimal condition (EC.20). This shows that $(\rho_1^*, \rho_2^*, \rho_3^*)$ cannot be attained in the region $\{(\rho_1, \rho_2, \rho_3)|\rho_1 > 0, \rho_2 > 0, \rho_3 > 0\}$. Therefore, we must have $\rho_i = 0$ for some $i \in \{1, 2, 3\}$. If there are $\rho_i = \rho_j = 0$ for $i, j \in (1, 2, 3)$ and $i \neq j$, it degenerates to the homogenous service case. Because we have shown that SSRD policy outperforms the homogeneous policy, there should be only one $\rho_i = 0$ (the grade-2 case).

We next treat the general case $m \geq 3$. We assume that this structure holds for the i case, $i \leq m$, that is, if there are in total m service grades, the optimal SSRD allocation is to allocate the arriving customers with two classes. We consider the $m+1$-grade case. If $\rho_i^* > 0$ for $i = 1, \dots, m+1$, similar to (EC.20), we have

$$f_k(\rho_1, ..., \rho_{m+1}) = 2(\sqrt{x_1} - \sqrt{x_{k+1}}) - \mathbf{a}(\mathbf{b}_1 - \mathbf{b}_{k+1}) = 0, \qquad k = 1, \dots, m. \tag{EC.23}$$

When $k = 1$, similar to (EC.21), we have

$$\mathbf{b}_1 - \mathbf{b}_2 = \mathbf{A}^{-1}(\mathbf{B}_1 - \mathbf{B}_2)\mathbf{A}^{-1}\mathbf{e} = -A^{-1}\begin{pmatrix} c_{2,1} & c_{2,1}-1 & 0 & \cdots & 0 \\ c_{2,1} & c_{2,1}-1 & 0 & \cdots & 0 \\ c_{3,1} & -c_{3,2} & c_{3,1}-c_{3,2} & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ c_{m,1} & -c_{m,2} & 0 & \cdots & c_{m,1}-c_{m,2} \end{pmatrix}\mathbf{x},$$

where B_1 and B_2 are defined as (EC.19) analogically. Denote $d_{i,j}$ as the $(i,j)^{\text{th}}$ entries of $\mathbf{A}^{-1}$. From (i) of Lemma 1, we have $d_{i,j} \leq 0$. Because $x_1 c_{2,1} - x_2 c_{1,2} < 0$, the k^{th} element of the vector

$$\mathbf{A}^{-1}\begin{pmatrix} c_{2,1} & c_{2,1}-1 & 0 & \cdots & 0 \\ c_{2,1} & c_{2,1}-1 & 0 & \cdots & 0 \\ c_{3,1} & -c_{3,2} & c_{3,1}-c_{3,2} & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ c_{m,1} & -c_{m,2} & 0 & \cdots & c_{m,1}-c_{m,2} \end{pmatrix}\mathbf{x}$$

satisfies

$$\begin{aligned} & x_1[(d_{k,1} + d_{k,2})c_{2,1} + \sum_{i=3}^m d_{k,i}c_{i,1}] - x_2[(d_{k,1} + d_{k,2})c_{1,2} + \sum_{i=3}^m d_{k,i}c_{i,2}] + x_i \sum_{i=3}^m (c_{i,1} - c_{i,2})d_{k,i} \\ \geq & \sum_{i=3}^m d_{k,i}c_{i,1}x_1 - d_{k,i}c_{i,2}x_2 + d_{k,i}c_{i,1}x_i - d_{k,i}c_{i,2}x_i \\ = & \sum_{i=3}^m d_{k,i}\left[\frac{x_1}{C_1+1} - \frac{C_2x_2}{C_i+C_2} + \frac{x_i}{C_i+1} - \frac{C_2x_i}{C_i+C_2}\right] \\ = & \sum_{i=3}^m d_{k,i}\left[\frac{x_1}{C_1+1} - \frac{C_2x_2}{C_i+C_2} + \frac{C_ix_i}{C_i+C_2} - \frac{C_ix_i}{C_i+1}\right] \geq \sum_{i=3}^m d_{k,i}(C_2x_2 - C_ix_i)\left(\frac{1}{C_i+1} - \frac{1}{C_i+C_2}\right) > 0. \end{aligned}$$

Therefore, we have $\mathbf{b}_1 - \mathbf{b}_1 < 0$, which implies that $f_1(\rho_1, \dots, \rho_m) > 0$, because $\mathbf{a}(\mathbf{b}_1 - \mathbf{b}_2) < 0$ and $\sqrt{x_1} > \sqrt{x_2}$ ((i) of Lemma 1). This means that $\rho_i > 0$ for all i can not be optimal for $(m+1)$-grade case, so that we must have $\rho_i = 0$ for some i. Hence, the $m+1$ case degenerates to the k-grade case for some $k \leq m$.

Step 2: Treating the $m=2$ case. For the 2-grade case with $C=\theta_2/\theta_1$, the optimal allocation probability can be derived directly from Proposition 1, namely,

$$p_1^*(C,\boldsymbol{\rho})=\frac{\rho_1\sqrt{1+C-C\rho}}{\rho_1\sqrt{1+C-C\rho}+\rho\sqrt{1+C-\rho}},\quad p_2^*(C,\boldsymbol{\rho})=\frac{\rho\sqrt{1+C-\rho}}{\rho_1\sqrt{1+C-C\rho}+\rho\sqrt{1+C-\rho}}. \tag{EC.24}$$

We next develop the optimal policy for $\rho_1^*(C)$ and $\rho_2^*(C)$ by unconditioning on $\boldsymbol{\rho}$. Substituting (EC.24) into (14), our minimization problem becomes

$$\begin{aligned}\min_{\rho_1,\rho_2}\quad & \frac{(\rho_1\sqrt{1+C-\rho}+\rho_2\sqrt{1+C-C\rho})^2}{1+C-C\rho+(C-1)\rho_1}\\ s.t.\quad & \rho_1+\rho_2=\rho<1.\end{aligned} \tag{EC.25}$$

We can give the optimal SSRD results (as a function of C and ρ) below.
Let $E=1+C-C\rho$, $F=1+C-\rho$, $\rho_1=z$ and $\rho_2=\rho-z$. Define $g(z)\equiv(\sqrt{F}z+(\rho-z)\sqrt{E})^2/[E+(C-1)z]$. Then the first-order derivative of $g(z)$ with respect to z is

$$\frac{dg(z)}{dz}=\frac{\Big((\sqrt{F}-\sqrt{E})z+\sqrt{E}\rho\Big)\cdot\Big((C-1)(\sqrt{F}-\sqrt{E})z-\Big((C-1)\sqrt{E}\rho-2((\sqrt{F}-\sqrt{E})E)\Big)\Big)}{(E+(C-1)z)^2}. \tag{EC.26}$$

Because $F>E$ and $C>1$, it is easy to verify that $(C-1)(\sqrt{F}-\sqrt{E})x>0$ and $(C-1)\sqrt{E}\rho-2(\sqrt{F}-\sqrt{E})E>0$. So we conclude that

$$\frac{dg(z)}{dz}\begin{cases}<0, & \text{when}\quad 0<z<\frac{\sqrt{E}\rho}{\sqrt{F}-\sqrt{E}}-\frac{2E}{C-1};\\ >0, & \text{when}\quad \frac{\sqrt{E}\rho}{\sqrt{F}-\sqrt{E}}-\frac{2E}{C-1}<z<\rho.\end{cases}$$

Hence, $z^*=\sqrt{E}\rho/(\sqrt{F}-\sqrt{E})-2E/(C-1)$ is the unique minimizer in $(0,\rho)$. Substituting it into $p_1^*(C,\rho)$ and $p_2^*(C,\rho)$ in (EC.24) yields $(p_1^*(C,\rho),p_2^*(C_2,\rho))=(E/E+F,F/E+F)$. The corresponding optimal SSRD parameters $(\boldsymbol{\rho}^*,\mathbf{p}^*,\boldsymbol{\mu}^*,\boldsymbol{\theta}^*)$ and delay $w^*(C,\rho)$ can be obtained, accordingly, by replacing ρ_1 with z^*.

We have showed that, if there are $m\geq 2$ "candidate" grades, it is optimal to consider 2 grades. It remains to argue that we should choose grade 1 and grade m, not any other grade j, $1<j<m$. For a 2-grade SSRD policy with a fixed $\rho>0$, $w^*(C,\rho)$ is decreasing in C (see Proposition 2). Hence, we have $w^*(C_m,\rho)<w^*(C_{m-1},\rho)<\cdots<w^*(C_1,\rho)$. This concludes that it is optimal to to allocate all customers to the classes having the *maximum* retrial rate (case m) and the *minimal* retrial rate (case 1) (none to any other classes). ■

Proof of Proposition 2 Plugging the optimal SSRD parameters ρ_1^*,ρ_2^* and p_1^*,p_2^* into (9), we obtain that the expected waiting time as a function of C:

$$w^*(C,\rho)=\frac{2\rho^2(c_v^2+1)}{(1-\rho)\lambda_0}\left(\frac{\sqrt{(1+C-C\rho)(1+C-\rho)}}{(\sqrt{1+C-C\rho}+\sqrt{1+C-\rho})^2}\right)+\frac{\rho}{(1-\rho)\theta_0},$$

which has the derivative with respect to C

$$\frac{dw^*(C,\rho)}{dC} = -\frac{(c_v^2+1)(2-\rho)\rho^4(C-1)}{2\sqrt{1+C-\rho}\sqrt{1+C-C\rho}\left(\sqrt{1+C-\rho}+\sqrt{1+C-C\rho}\right)^4(1-\rho)\lambda_0}. \tag{EC.27}$$

Since $C>1$ and $\rho<1$, we can easily validate that $dw^*(C,\rho)/dC<0$, so that $w^*(C,\rho)$ is decreasing in $C\geq 1$. Letting $C\to\infty$ in (15)–(17) yields (19)–(20). ■

Proof of Proposition 3 Let $E=1+C-C\rho$ and $F=1+C-\rho$, we have

$$\lim_{C\to\infty} w^*(C,\rho) = \lim_{C\to\infty}\left(w_0^B\left(\frac{4\sqrt{EF}}{(\sqrt{E}+\sqrt{F})^2}\right)+w_0^I\right) = w_0^B\left(\frac{\sqrt{1-\rho}}{(\sqrt{1-\rho}+1)^2}\right)+w_0^I.$$

Next, for any $\rho\in(0,1)$, we have

$$\lim_{C\to\infty}\left(w^*(C,\rho)-w_0^I-w_0^B\frac{\sqrt{1-\rho}}{(\sqrt{1-\rho}+1)^2}\right)C = \frac{(2-\rho)\rho^2}{2\sqrt{1-\rho}(1+\sqrt{1-\rho})^4}<\infty.$$

Thus we have $w^*(C,\rho)=w_0^I+w_0^B\sqrt{1-\rho}(1/(\sqrt{1-\rho}+1)^2+O(1/C))$. That is, $w^*(C,\rho)$ converges to $w_0^B\left(\frac{\sqrt{1-\rho}}{(\sqrt{1-\rho}+1)^2}\right)+w_0^I$ when $C\to\infty$ in the order of $O(1/C)$. Substituting the optimal allocation into (21) yields

$$R_D(C,\rho) = \frac{w_0-w^*(C,\rho)}{w_0} = \frac{1-\left(\frac{4\sqrt{EF}}{(\sqrt{E}+\sqrt{F})^2}\right)}{1+2\mu_0/(\theta_0(c_v^2+1))}. \tag{EC.28}$$

It is sufficient to prove that $(E+F)/\sqrt{EF}$ is increasing in C and ρ, namely,

$$\begin{aligned}\frac{d(E+F)/\sqrt{EF}}{dC} &= \frac{(2-\rho)(C-1)\rho^2}{2(1+C-\rho)^{3/2}(1+C-C\rho)^{3/2}}>0,\\ \frac{d(E+F)/\sqrt{EF}}{d\rho} &= \frac{(C-1)^2(1+C)\rho}{2(1+C-\rho)^{3/2}(1+C-C\rho)^{3/2}}>0.\end{aligned}$$

Therefore, $R_D(C,\rho)$ is increasing in C and ρ. For any given μ_0 and θ_0, as $C\to\infty$ and $\rho\to 1$, we obtain the desired result. In addition, the asymptotic order of growth in the homogeneous case is $O(1/(1-\rho))$ because $w_0=(1/\theta_0+(c_v^2+1)/2\mu_0)\rho/(1-\rho)$. For the waiting time under SSRD, it is decreasing in ratio C. When $C\to\infty$, by substituting $C=\infty$ into E and F in (17), we have $w^*(\infty,\rho)=O(1/\sqrt{1-\rho})+O(1/(1-\rho))$. Therefore, $w_0-w^*(\infty,\rho)=O(1/(1-\rho))-O(1/\sqrt{1-\rho})$, which gives

$$\frac{w_0-w^*(C,\rho)}{w_0} = \frac{O(1/(1-\rho))-O(1/\sqrt{1-\rho})}{O(1/(1-\rho))} = \frac{1}{1+2\mu_0/(\theta_0(c_v^2+1))}. \quad ■$$

Proof of Proposition 4 For fixed $C=\theta_2/\theta_1>1$, θ_0 and ρ, the average number of retrials under SSRD (Theorem 2) and homogeneous service are given by

$$r^*(C,\rho)=\frac{(c_v^2+1)\rho^2\theta_0}{(1-\rho)\lambda_0}\left(\frac{\sqrt{EF}(CE+F)(C+1)}{C(E+F)(\sqrt{E}+\sqrt{F})^2}\right)+\frac{\rho}{1-\rho}\quad\text{and}\quad r_0=\frac{\rho\theta_0}{(1-\rho)\mu_0}+\frac{\rho}{1-\rho}.$$

Therefore, we have

$$R_T(C,\rho) = \frac{r_0 - r^*(C,\rho)}{r_0} = \frac{1 - \frac{2\sqrt{r_p}(C+r_p)(C+1)}{C(1+r_p)(1+\sqrt{r_p})^2}}{1+2\mu_0/(\theta_0(c_v^2+1))}. \tag{EC.29}$$

In order for our SSRD policy to outperform the homogeneous service policy, we need

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

or equivalently,

$$4(1+C-\rho)(1+C-C\rho)\left(1+C-\rho+C^2(1+C-C\rho)\right)^2 < C^2(1+C)^4(2-\rho)^4.$$

We define

$$h(C,\rho) \equiv C^2(1+C)^4(2-\rho)^4 - 4(1+C-\rho)(1+C-C\rho)\left(1+C-\rho+C^2(1+C-C\rho)\right)^2. \tag{EC.30}$$

By taking the first and second partial derivatives of $h(C,\rho)$ with respect to ρ, we have

$$\frac{\partial h(C,\rho)}{\partial \rho} = 4\left(C^2-1\right)^2\left(3\left(C^4+1\right)(1-\rho)^2 + C^2\left(14-22\rho+9\rho^2+\rho^3\right) - 2C\left(1+C^2\right)\left(-5+9\rho-6\rho^2+2\rho^3\right)\right),$$

$$\frac{\partial h^2(C,\rho)}{\partial \rho^2} = 4\left(C^2-1\right)^2\left(6\left(1+C^4\right)(\rho-1) + C^2\left(-22+18\rho+3\rho^2\right) - 2C\left(1+C^2\right)\left(9-12\rho+6\rho^2\right)\right).$$

For any $C>1$, we have $\rho<1$, $-22+18\rho+3\rho^2<0$ and $9-12\rho+6\rho^2 = 6(\rho-1)^2+3>0$, which imply $\partial h^2(C,\rho)/d\rho^2<0$. Because $\partial h(C,\rho)/d\rho|_{\rho=0}>0$ and $\partial h(C,\rho)/d\rho|_{\rho=1}>0$, $h(C,\rho)$ must be increasing in $\rho\in[0,1]$. It is also noted that $h(C,0) = -4(1+C)^4(C^2-1)^2<0$ and $h(C,1) = C^2(1+C)^2(C-1)^2>0$. Therefore, there exists a unique $\overline{\rho}_C$ such that $R_T(C,\rho)>0$ when $\rho>\overline{\rho}_C$ for any given $C>1$. The value of $\overline{\rho}_C$ can be found by solving $f(C,\rho)=0$. If $\lim_{\rho\to1,C\to\infty} C^{2/3}\cdot(1-\rho) = \vartheta\in[0,\infty]$, then $\lim_{\rho\to1,C\to\infty} r_p/C^{2/3} = \frac{(1-\rho)/C+1}{C^{2/3}(1-\rho)+C^{-1/3}} = 1/\vartheta$, which leads to

$$\begin{aligned}\lim_{\rho\to1,C\to\infty}\frac{2\sqrt{r_p}(C+r_p)(C+1)}{C(1+r_p)(1+\sqrt{r_p})^2} &= \lim_{\rho\to1,C\to\infty}\frac{2C^{1/3}(C+\vartheta C^{2/3})(C+1)}{C\sqrt{\vartheta}(1+C^{2/3}/\vartheta)(1+C^{1/3}/\sqrt{\vartheta})^2},\\ &= \lim_{\rho\to1,C\to\infty}\frac{2\sqrt{\vartheta}(C^{1/3}+1/\vartheta)(C+1)}{(1+1/\vartheta C^{2/3})C^{2/3}} = 2\sqrt{\vartheta},\end{aligned}$$

yielding $\lim_{\rho\to1,C\to\infty} R_T(C,\rho) = (1-2\sqrt{\vartheta})/(1+2\mu_0/[\theta_0(c_v^2+1)])$.

Next we will identify the maximal $R_T(C,\rho)$. First, we will show that $R_T(C,\rho)$ is increasing in $\rho\in(0,1)$ for any $C>1$. Then it is sufficient to prove that $u(r_p)\equiv\frac{2\sqrt{r_p}(C+r_p)(C+1)}{C(1+r_p)(1+\sqrt{r_p})^2}$ is decreasing in r_p, because $r_p = \frac{1-\rho+C}{1+C(1-\rho)}\geq 1$ is increasing in $\rho\in(0,1)$. The derivative of $u(r_p)$ is

$$\frac{du(r_p)}{dr_p} = \frac{C+1}{C}\cdot\left[\frac{\bar{\phi}(r_p)}{\sqrt{r_p}(1+\sqrt{r_p})^3(1+r_p)^2}\right],$$

where $\bar{\phi}(r_p) = r_p(3+\sqrt{r_p}+r_p-r_p^{3/2}) - C(r_p+\sqrt{r_p}+3r_p^{3/2}-1)$. Because $\bar{\phi}(1)\leq 0$ and

$$\begin{aligned}\frac{d\bar{\phi}(r_p)}{dr_p} &= \frac{6\sqrt{r_p}+3r_p+4r_p^{3/2}-5r_p^2 - C\left(1+2\sqrt{r_p}+9r_p\right)}{2\sqrt{r_p}}\\ &\leq \frac{6\sqrt{r_p}+3r_p - C\left(1+2\sqrt{r_p}+9r_p\right)}{2\sqrt{r_p}} \leq \frac{4\sqrt{r_p}-6r_p-1}{2\sqrt{r_p}} < 0,\end{aligned}$$

we conclude that $\bar{\phi}(r_p) \le 0$, and more important, $u(r_p)$ is decreasing in r_p. Taking $\rho \to 1$, we have

$$R_T(C,1) = \frac{r_0 - r^*(C,\rho)}{r_0} = \frac{1 - \frac{4\sqrt{C}}{(1+\sqrt{C})^2}}{1 + 2\mu_0/(\theta_0(c_v^2+1))}. \tag{EC.31}$$

Because $R_T(C,1)$ increases in $C>1$, it can be asymptotically maximized when $C \to \infty$ (the upper bound can be attained at $\vartheta = 0$). ■

Proof of Proposition 5 We consider the asymptotic value of $R_S(C,\rho)$ as $\rho \to 1, C \to \infty$. First, note that $\lim_{\rho\to 1, C\to\infty} 1-\rho = \lim_{\rho\to 1, C\to\infty} 1/C = \lim_{\rho\to 1, C\to\infty} (1-\rho)/C = \lim_{\rho\to 1, C\to\infty} \sqrt{r_p}/(1+r_p) = 0$ and $\lim_{\rho\to 1, C\to\infty} = C/[(1-\rho)C+1] = \infty$. With $\xi = \lim_{\rho\to 1, C\to\infty} (1-\rho)C$, we have

$$\begin{aligned}
\lim_{\rho\to 1, C\to\infty} \frac{\gamma_0 - \gamma^*(C,\rho)}{\gamma_0} &= \lim_{\rho\to 1, C\to\infty} 1 - \frac{1 + \frac{\rho}{1-\rho}\left[\frac{(C+r_p\sqrt{r_p})(1+\sqrt{r_p})}{(1+r_p)\theta_0(C+r_p)} + \frac{(c_v^2+1)\sqrt{r_p}}{\mu_0(1+r_p)}\right] E[S_0^{-1}]}{1 + \frac{\rho}{1-\rho}\left(\frac{c_v^2+1}{2\mu_0} + \frac{1}{\theta_0}\right) E[S_0^{-1}]} \\
&= \lim_{\rho\to 1, C\to\infty} 1 - \frac{\frac{(C+r_p\sqrt{r_p})(1+\sqrt{r_p})}{(1+r_p)\theta_0(C+r_p)} + \frac{(c_v^2+1)\sqrt{r_p}}{\mu_0(1+r_p)}}{\frac{c_v^2+1}{2\mu_0} + \frac{1}{\theta_0}} \\
&= \lim_{\rho\to 1, C\to\infty} \frac{\frac{c_v^2+1}{2\mu_0} + \frac{1}{\theta_0} - \frac{\left[1+\frac{1}{(1-\rho)C+1}\sqrt{\frac{C}{(1-\rho)C+1}}\right]\left(1+\sqrt{\frac{C}{(1-\rho)C+1}}\right)}{\left(1+\frac{C}{(1-\rho)C+1}\right)\left(1+\frac{1}{(1-\rho)C+1}\right)\theta_0}}{\frac{c_v^2+1}{2\mu_0} + \frac{1}{\theta_0}} \\
&= \lim_{\rho\to 1, C\to\infty} \frac{\frac{c_v^2+1}{2\mu_0} + \frac{1}{\theta_0} - \frac{1}{(\xi+2)\theta_0}}{\frac{c_v^2+1}{2\mu_0} + \frac{1}{\theta_0}} = \lim_{\rho\to 1, C\to\infty} \frac{c_v^2 + 1 + 2[(\xi+1)/(\xi+2)]\mu_0/\theta_0}{c_v^2+1+2\mu_0/\theta_0}.
\end{aligned}$$

Proof of Corollary 3

In Proposition 3, we have showed that RRD can be maximized as $C \to \infty$ and $\rho \to 1$. When $C = O(1/(1-\rho)^\alpha)$ with $\alpha \in (1, 3/2)$, we have $\lim_{C\to\infty, \rho\to 1} C^{2/3}\cdot(1-\rho) = 0$ and $\lim_{C\to\infty,\rho\to 1} C\cdot(1-\rho) = \infty$, so that the maximum RRT and RRS can be attained, respectively. ■

Proof of Proposition 6

We first note that $\partial F(x_C, y_C, C)/\partial x_C = \partial F(x_C, y_C, C)/\partial y_C = 0$. Then by taking the derivative of $dF(x_C, y_C, C)$ with respect to C, we have

$$\begin{aligned}
&\frac{dF(x_C,y_C,C)}{dC} \\
=\ &\frac{\partial F(x_C,y_C,C)}{\partial x_C}\frac{\partial x_C}{\partial C} + \frac{\partial F(x_C,y_C,C)}{\partial y_C}\frac{\partial y_C}{\partial C} + \frac{\partial F(x_C,y_C,C)}{\partial C} \\
=\ &\frac{\partial F(x_C,y_C,C)}{\partial C} \\
=\ &\frac{\partial f(w_1(x_C,y_C,C))}{\partial w_1(x_C,y_C,C)}\frac{\partial w_1(x_C,y_C,C)}{\partial C} x_C + \frac{\partial f(w_2(x_C,y_C,C))}{\partial w_2(x_C,y_C,C)}\frac{\partial w_2(x_C,y_C,C)}{\partial C}(1-x_C) \\
=\ &\frac{\partial f(w_1(x_C,y_C,C))}{\partial w_1(x_C,y_C,C)}\left(\frac{(2-\rho)(\rho-y_C)\left(\frac{y_C^2}{x_C} + \frac{(\rho-y_C)^2}{1-x_C}\right)}{(1-y_C+(1-\rho+y_C)C)^2\lambda_0(1-\rho)} + \frac{(1-x_C)\rho}{(x_C(C-1)+1)^2\theta_0(1-\rho)}\right) x_C
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\partial f(w_2(x_C,y_C,C))}{\partial w_2(x_C,y_C,C)}\left(\frac{-y_C(2-\rho)\left(\frac{y_C^2}{x_C}+\frac{(\rho-y_C)^2}{1-x_C}\right)}{(1-y_C+(1-\rho+y_C)C)^2\lambda_0(1-\rho)}-\frac{x_C\rho}{(x_C(C-1)+1)^2\theta_0(1-\rho)}\right)(1-x_C)\\
&=\left(\frac{\partial f(w_1(x_C,y_C,C))}{\partial w_1(x_C,y_C,C)}x_C(\rho-y_C)-\frac{\partial f(w_2(x_C,y_C,C))}{\partial w_2(x_C,y_C,C)}y_C(1-x_C)\right)\frac{(2-\rho)\left(\frac{y_C^2}{x_C}+\frac{(\rho-y_C)^2}{1-x_C}\right)}{(1-y_C+(1-\rho+y_C)C)^2\lambda_0(1-\rho)}\\
&+\left(\frac{\partial f(w_1(x_C,y_C,C))}{\partial w_1(x_C,y_C,C)}-\frac{\partial f(w_2(x_C,y_C,C))}{\partial w_2(x_C,y_C,C)}\right)\frac{x_C(1-x_C)\rho}{(x_C(C-1)+1)^2\theta_0(1-\rho)}.
\end{aligned}
$$

Let $G\equiv\partial f(w_1(x_C,y_C,C))/\partial w_1(x_C,y_C,C)-\partial f(w_2(x_C,y_C,C))/\partial w_2(x_C,y_C,C)$, we consider the following three cases:

Case 1. If f is linear (i.e., $G=0$), we have $dF(x_C,y_C,C)/dC<0$, because $x_C(\rho-y_C)-y_C(1-x_C)=\lambda x_C(1-x_C)(1/\mu_2-1/\mu_1)<0$, which implies that $C^*=\infty$.

Case 2. If f is concave (i.e., $G<0$), we have $dF(x_C,y_C,C)/dC<0$ (which similar to the analysis in case 1).

Case 3. If f is convex (i.e., $G>0$), we know that $\frac{\partial f(w_1(x_C,y_C,C))}{\partial w_1(x_C,y_C,C)}$ $(\frac{\partial f(w_2(x_C,y_C,C))}{\partial w_2(x_C,y_C,C)})$ is increasing (decreasing) in C because $w_1(x_C,y_C,C)$ $(w_2(x_C,y_C,C))$ is increasing (decreasing) in C. It is evident that $\lim_{C\to\infty}\frac{dF(x_C,y_C,C)}{dC}=0$ and $\lim_{C\to\infty}|C^2(\frac{dF(x_C,y_C,C)}{dC})|<\infty$. In particular, if $\lim_{C\to\infty}C^2(\frac{dF(x_C,y_C,C)}{dC})>0$, it implies that $F(x_C,y_C,C)$ will keep increasing when C is large, hence the optimal ratio that minimizes $F(x_C,y_C,C)$ can only be attained at a certain finite value $C^*\in(0,\infty)$.

The limits $C\to\infty$, $\lim_{C\to\infty}x_C=x_\infty$ and $\lim_{C\to\infty}y_C=y_\infty$ can be derived by solving the equations (32)-(33) with $C=\infty$. By substituting them into $w_1(\infty)=w_1(x_\infty,y_\infty,\infty)$ and $w_2(\infty)=w_2(x_\infty,y_\infty,\infty)$, we can obtain the resulting waiting times. Therefore, a sufficient condition that a finite C^* can be attained is

$$
\begin{aligned}
\lim_{C\to\infty}C^2\cdot\frac{dF(x_C,y_C,C)}{dC}>0\Leftrightarrow&\left(\frac{\partial f(w_1(\infty))}{\partial w_1(\infty)}x_\infty(\rho-y_\infty)-\frac{\partial f(w_2(\infty))}{\partial w_2(\infty)}y_\infty(1-x_\infty)\right)\frac{(2-\rho)\left(\frac{y_\infty^2}{x_\infty}+\frac{(\rho-y_\infty)^2}{1-x_\infty}\right)}{(1-\rho+y_\infty)^2\lambda_0(1-\rho)}\\
&+\left(\frac{\partial f(w_1(\infty))}{\partial w_1(\infty)}-\frac{\partial f(w_2(\infty))}{\partial w_2(\infty)}\right)\frac{(1-x_\infty)\rho}{x_\infty\theta_0(1-\rho)}>0.
\end{aligned}\tag{EC.32}
$$

Note that the condition (EC.32) always holds as long as $x_\infty(\rho-y_\infty)>y_\infty(1-x_\infty)$, or equivalently $x_\infty\rho>y_\infty$. From (33), we have

$$\left(\frac{df(w_1(x,y,C))}{dw_1(x,y,C)}(1-\rho)(1-x)+\frac{df(w_2(x,y,C))}{dw_2(x,y,C)}x\right)\frac{\partial w_1(x,y,C)}{\partial y}=0.$$

Because $\partial w_1(x,y,C)/\partial y=0$ (due to the fact that $[df(w_1(x,y,C))/dw_1(x,y,C)](1-\rho)(1-x)+[df(w_2(x,y,C))/dw_2(x,y,C)]x>0$), the equation above implies that $y_C=[\sqrt{(1+x_C)^2(1-\rho)^2+4x_C\rho(2-\rho)}-(1+x_C)(1-\rho)]/2$, which yields

$$
\begin{aligned}
x_\infty\rho>y_\infty&\Leftrightarrow 2(2-\rho)<\sqrt{(1+x_\infty)^2(1-\rho)^2+4x_\infty\rho(2-\rho)}-(1+x_\infty)(1-\rho)\\
&\Leftrightarrow(2(2-\rho)+(1+x_\infty)(1-\rho))^2<(1+x_\infty)^2(1-\rho)^2+4x_\infty\rho(2-\rho)\Leftrightarrow x_\infty>\frac{1-\rho}{1+\rho}.\quad\blacksquare
\end{aligned}
$$

Proof of Lemma 7 Note that i and j are the numbers of customers in the buffer and orbit, respectively, we have the following balance equations for $1 \le i \le K-1$ and $j \ge 0$:

$$(\lambda_0 + j\theta_0)p_{(0,j)} = \mu_0 p_{(1,j)}, \tag{EC.33}$$

$$(\lambda_0 + \mu_0 + j\theta_0)p_{(i,j)} = \mu_0 p_{(i+1,j)} + (j+1)\theta_0 p_{(i-1,j+1)} + \lambda_0 p_{i-1,j}, \tag{EC.34}$$

$$(\lambda_0 + \mu_0)p_{(K,j)} = (j+1)\theta_0 p_{(K-1,j+1)} + \lambda_0 p_{K-1,j} + \lambda_0 p_{K,j-1}. \tag{EC.35}$$

We will solve the above equations using the generating function below:

$$\Pi_i(z) = \sum_{j=0}^{\infty} z^j p_{(i,j)}, \quad 1 \le i \le K-1.$$

Multiplying equations (EC.33)-(EC.35) by z^j and summing up over all $j \ge 0$, we obtain the balance equations of the generating functions:

$$\lambda_0 \Pi_0(z) + z\theta_0 \Pi_0'(z) = \mu_0 \Pi_1(z), \tag{EC.36}$$

$$(\lambda_0 + \mu_0)\Pi_i(z) + z\theta_0 \Pi_i'(z) = \mu_0 \Pi_{i+1}(z) + \theta_0 \Pi_{i-1}'(z) + \lambda_0 \Pi_{i-1}(z), \tag{EC.37}$$

$$(\lambda_0 - \lambda_0 z + \mu_0)\Pi_K(z) = \theta_0 \Pi_{K-1}'(z) + \lambda_0 \Pi_{K-1}(z). \tag{EC.38}$$

Multiplying equations in (EC.37) by z^i for $1 \le i \le K-1$ and then sum them over for (EC.36)-(EC.38), we have

$$\begin{aligned}
&\lambda_0 \Pi_0(z) + (\lambda_0 + \mu_0)\sum_{i=1}^{K-1} \Pi_i(z) z^i + (\lambda_0 - \lambda_0 z + \mu_0)\Pi_K(z) z^K \\
&= \mu_0 \Pi_1(z) + \sum_{i=1}^{K-1} (\mu_0 \Pi_{i+1}(z) + \lambda_0 \Pi_{i-1}(z)) z^i + \lambda_0 \Pi_{K-1}(z) z^K \\
&\Leftrightarrow \lambda_0 (1-z)\Pi_0(z) + \sum_{i=1}^{K} (\lambda_0 z^i (1-z)\Pi_i(z) - \mu_0 z^{i-1}(1-z)\Pi_i(z)) = 0 \\
&\Leftrightarrow \lambda_0 \Pi_0(z) + \sum_{i=1}^{K} (\lambda_0 z - \mu_0) z^{i-1} \Pi_i(z) = 0.
\end{aligned} \tag{EC.39}$$

Setting $z = 1$ in (EC.39) yields $\sum_{i=1}^{K} \Pi_i = \frac{\lambda_0}{\mu_0}$, which is the probability that the server is busy. Furthermore, by taking the derivative with respect to z in (EC.39) and letting $z = 1$, we get

$$\begin{aligned}
&\lambda_0 \Pi_0'(1) + \sum_{i=1}^{K} ((i\lambda_0 - (i-1)\mu_0)\Pi_i + (\lambda_0 - \mu_0)\Pi_i'(1)) = 0 \\
\Leftrightarrow \quad &\mu_0 \Pi_0'(1) + \sum_{i=1}^{K} (i\lambda_0 - (i-1)\mu_0)\Pi_i = (\mu_0 - \lambda_0)\sum_{i=0}^{K} \Pi_i'(1) \\
\Leftrightarrow \quad &N_{orbit} = \frac{\mu_0(\Pi_1 - \rho_0(1-\rho_0))}{\theta_0(1-\rho_0)} + \sum_{i=1}^{K} \frac{(i\rho_0 - (i-1))\Pi_i}{1-\rho_0},
\end{aligned}$$

where $N_{orbit}=\sum_{i=0}^{K}\Pi_i'(1)$ is the mean number of customers in the orbit. ■

Proof of Proposition 8 When the capacity of waiting line is K, we let $I(t)$ be the number of customers in line and $N(t)$ be the state of the waiting line, where $N(t)\in\{0,1,\dots,2^K\}$. When the state of waiting line is $N(t)$, the total number of customers in the buffer can be uniquely determined by $I(t)=\lfloor \log_2^{N(t)+1}\rfloor$, see Figure EC.1. Define $N_{I(t)}(t)\equiv N(t)$ and $N_{i-1}(t)\equiv\lfloor (N_i(t)-1)/2\rfloor$ for $i=I(t),I(t)-1,...,1$. Then the i^{th} customer in the waiting line is a type-2 customer if and only if $N_i(t)$ is even. Therefore, the state of waiting line can be characterized by $N(t)$ uniquely. Then the system state under SSRD can be modeled by the *continuous-time Markov chain* (CTMC) $\{(I(t),N(t),Q_1(t),Q_2(t));t\geq 0\}$, where $Q_i(t)$ is the number of type-i customer in orbit i, $i=1,2$. Its infinitesimal generator of the CTMC is given as follows:

$$q_{(i,n,m_1,m_2),(i',n',m_1',m_2')}=\begin{cases}\mu_{\lceil \frac{n}{2^{i-1}\cdot 3-2}\rceil}, & \text{if } (i',n',m_1',m_2')=(i-1,n-2^{i+\lceil\frac{n}{2^{i-1}\cdot 3-2}\rceil-2},m_1',m_2');\\ m_1\theta_1, & \text{if } (i',n',m_1',m_2')=(i+1,2n+1,m_1'-1,m_2');\\ m_2\theta_2, & \text{if } (i',n',m_1',m_2')=(i+1,2n+2,m_1',m_2'-1);\\ \lambda_1, & \text{if } (i',n',m_1',m_2')=(i+1,2n+1,m_1',m_2');\\ \lambda_2, & \text{if } (i',n',m_1',m_2')=(i+1,2n+2,m_1',m_2').\end{cases}$$

Let $P(i,n)=\sum_{m_1}\sum_{m_2}p(i,n,m_1,m_2)z_1^{m_1}z_2^{m_2}$, through Kolmogorov equations for the stationary distributions, and take the summation for $m_1\geq 0$, $m_2\geq 0$, we have

$$\begin{aligned}&(\lambda_0+\mu_{\lceil\frac{2n+1}{2^{i-1}\cdot 3-2}\rceil})P(i,2n+1)+\frac{z_1\partial P(i,2n+1)\theta_1}{\partial z_1}+\frac{z_2\partial P(i,2n+1)\theta_2}{\partial z_2}\\ &=\mu_1P(i+1,2n+1+2^i)+\mu_2P(i+1,2n+1+2^{i+1})+\frac{\theta_1\partial P(i-1,n)}{\partial z_1}+\lambda_1P(i-1,n), \quad \text{(EC.40)}\\ &(\lambda_0+\mu_{\lceil\frac{2n+2}{2^{i-1}\cdot 3-2}\rceil})P(i,2n+2)+\frac{z_1\partial P(i,2n+2)\theta_1}{\partial z_1}+\frac{z_2\partial P(i,2n+2)\theta_2}{\partial z_2}\\ &=\mu_1P(i+1,2n+2+2^i)+\mu_2P(i+1,2n+2+2^{i+1})+\frac{\theta_2\partial P(i-1,n)}{\partial z_2}+\lambda_2P(i-1,n), \quad \text{(EC.41)}\\ &\lambda_0P(0,0)+\frac{\partial P(0,0)\theta_1}{\partial z_1}+\frac{\partial P(0,0)\theta_2}{\partial z_2}=\mu_1P(1,1)+\mu_2P_{1,2},\\ &(\lambda_0+\mu_{\lceil\frac{2n+1}{2^{i-1}\cdot 3-2}\rceil}-\lambda_1z_1-\lambda_2z_2)P(K,2n+1)=\frac{\partial P(K-1,n)\theta_1}{\partial z_1}+\lambda_1P(K-1,n), \quad \text{(EC.42)}\\ &(\lambda_0+\mu_{\lceil\frac{2n+2}{2^{i-1}\cdot 3-2}\rceil}-\lambda_1z_1-\lambda_2z_2)P(K,2n+2)=\frac{\partial P(K-1,n)\theta_2}{\partial z_2}+\lambda_2P(K-1,n), \quad \text{(EC.43)}\end{aligned}$$

where $i\geq 1$, $2^{i-1}\leq n\leq 2^i-2$.

Multiplying $z_1^i\cdot(z_2/z_1)^{\sum_{j=0}^{i}\lfloor\frac{n+1-2^{i-1}}{2^i}\rfloor}$, $z_1^i\cdot(z_2/z_1)^{\sum_{j=0}^{i}\lfloor\frac{n+3/2-2^{i-1}}{2^i}\rfloor}$, $z_1^K\cdot(z_2/z_1)^{\sum_{j=0}^{K}\lfloor\frac{n+1-2^{i-1}}{2^i}\rfloor}$ and $z_1^K\cdot(z_2/z_1)^{\sum_{j=0}^{K}\lfloor\frac{n+3/2-2^{i-1}}{2^i}\rfloor}$ on the both sides of (EC.40), (EC.41), (EC.42) and (EC.43), and then summing them up over all $i=1,2,...,K-1$ and $2^{i-1}-1\leq n\leq 2^i-2$, we can eliminate all terms of $\frac{\partial P(i,n)}{\partial z_1}$ and $\frac{\partial P(i,n)}{\partial z_1}$. Letting $z_1=1$ and $z_2=z$, we can eliminate the $(1-z)$ on the both sides of the equation above. By letting $z=1$ and using $P_{0,0}+\sum_{i=1}^{K}\sum_{n=2^i-1}^{2^{i-1}\cdot 3-2}P(i,n)+\sum_{i=1}^{K}\sum_{n=2^{i-1}\cdot 3-2}^{2^{i+1}-2}P(i,n)=1$, we have

$$P(0,0)\lambda_2+\sum_{i=1}^{K}\sum_{n=2^i-1}^{2^{i-1}\cdot 3-2}P(i,n)\lambda_2+\sum_{i=1}^{K}\sum_{n=2^{i-1}\cdot 3-1}^{2^{i+1}-2}P(i,n)(\lambda_2-\mu_2)=0$$

$$\Leftrightarrow \quad P(0,0)\lambda_2 + \sum_{i=1}^{K} \sum_{n=2^i-1}^{2^{i-1}\cdot 3-2} P(i,n)\lambda_2 + \sum_{i=1}^{K} \sum_{n=2^{i-1}\cdot 3-2}^{2^{i+1}-1} P(i,n)\lambda_2 = \sum_{i=1}^{K} \sum_{n=2^{i-1}\cdot 3-1}^{2^{i+1}-2} P(i,n)\mu_2$$

$$\Leftrightarrow \quad \lambda_2 = \mu_2 \sum_{i=1}^{K} \sum_{n=2^{i-1}\cdot 3-1}^{2^{i+1}-2} P(i,n).$$

Notice that $N(t) \in [2^{i-1}\cdot 3-1, 2^{i+1}-2]$ for $i=1,\ldots,K$ implies that the service area is occupied by type-2 customer, then the probability that the service area is occupied by type-2 customer is $\sum_{i=1}^{K}\sum_{n=2^{i-1}\cdot 3-1}^{2^{i+1}-2} P(i,n)$, which gives $\rho_2 = \sum_{i=1}^{K}\sum_{n=2^{i-1}\cdot 3-1}^{2^{i+1}-2} P(i,n) = \lambda_2/\mu_2$. Similarly, we can conclude that $\rho_1 = \lambda_1/\mu_1$, which completes this proof. ■

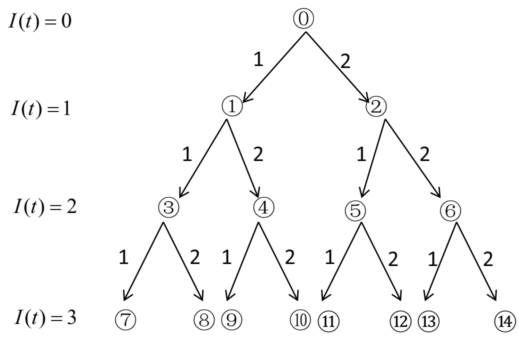


Figure EC.1 The system states in $M/M/1/K$ retrial model with SSRD

Algorithm 1

Step 1. Set the initial value of K, M, C and the λ, μ, θ, $\mathbf{p}$ under SSRD

Step 2. Define the transition matrix $\mathbf{Q}$

Step 3. Define an $\mathbf{e}^T$ in an additional column of $\mathbf{Q}$, and an additional 1 in a vector of 0's, $\mathbf{I}$.

Step 4. Calculate $\mathbf{\Pi} = \mathbf{I}\mathbf{Q}^{-1}$.

Step 5. Derive N_1, N_2 and the expected queue length L in the buffer through $\mathbf{\Pi}$.

EC.2. The Preemptive SSRD

In the main paper, we have studied the non-preemptive SSRD policy, where high priority customers may not always receive service before low priority customers, but they have a higher probability to receive service first. As a result, the performance (delay and number of trials) of high priority customers are somewhat influenced by low priority customers. In this section, we assume that type-1 customers may be preempted by type-2 customers. For tractability, we restrict our attention to 2 service groups. Artalejo et al. (2001) considered a retrial queueing system where retrial customers have preemptive priority over customers in the waiting line. Here we consider two retrial groups among which one group prempts the other.

We make the following model assumptions:

- An arrival seeing an idle server immediately enters service;
- If a type-2 customer, upon arrival or retrial, finds a type-1 customer in service, it immediately enters service by preempting that type-1 customer to the orbit queue;
- If a type-2 customer, upon arrival or retrial, finds the server is occupied by another type-2 customer, she will be blocked and enter the orbit queue;
- If a type-1 customer, upon arrival or retrial, finds the server is occupied by another customer (of type 1 or type 2), she will be blocked and enter the orbit queue.

We assume the retrial rates and service rates are θ_1,θ_2 and μ_1,μ_2 for the two classes. It is evident that performance of type-2 customers are not affected by type-1 customers, so that the expected waiting time for type-2 customers are given by (6). In particular,

$$w_2=\frac{\rho_2}{1-\rho_2}\left(\frac{1}{\theta_2}+\frac{1}{\mu_2}\right). \tag{EC.44}$$

The main difficulty is to compute the expected delay for type-1 customers. We will first obtain the stationary marginal distribution of the number of type-1 customers via the principle of maximum entropy; and we will next derive the expected number of customers using generating functions. Specifically, we consider a three dimensional CTMC $\{X(t);\, t\geq 0\}$=$\{(L(t),N_1(t),N_2(t))\ ;\ t\geq 0\}$, where $L(t)$ denotes the type of the customer in service (if any), and $N_i(t)$ is the number of type-i orbiting customers, $i=1,2$. The states $L(t)=0,1,2$ correspond to the case of an idle server, a type-1 customer in service, and a type-2 customer in service. For $m_1,m_2\geq 0$, we set up the following balance equations:

$$(\lambda+m_1\theta_1+m_2\theta_2)p_{(0,m_1,m_2)}=\mu_1 p_{(1,m_1,m_2)}+\mu_2 p_{(2,m_1,m_2)}, \tag{EC.45}$$

$$(\lambda+\mu_1+m_2\theta_2)p_{(1,m_1,m_2)}=\lambda_1 p_{(0,m_1,m_2)}+(m_1+1)\theta_1 p_{(0,m_1+1,m_2)}, \tag{EC.46}$$

$$\begin{aligned}(\lambda+\mu_2)p_{(2,m_1,m_2)}&=\lambda_2 p_{(0,m_1,m_2)}+(m_2+1)\theta_2 p_{(0,m_1,m_2+1)}+\lambda_2 p_{(1,m_1-1,m_2)}\\&\quad+\lambda_1 p_{(2,m_1-1,m_2)}+\lambda_2 p_{(2,m_1,m_2-1)},\end{aligned} \tag{EC.47}$$

where $p_{(i,-1,m_2)}=p_{(i,m_1,-1)}=0$ for $i=1,2$. We also define the following generating functions:

$$\Pi_0(z_1,z_2)=\sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty} z_1^{m_1}z_2^{m_2}p_{(0,m_1,m_2)},$$

$$\Pi_1(z_1,z_2)=\sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty} z_1^{m_1}z_2^{m_2}p_{(1,m_1,m_2)},$$

$$\Pi_2(z_1,z_2)=\sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty} z_1^{m_1}z_2^{m_2}p_{(2,m_1,m_2)}.$$

Multiplying equations (EC.45) and (EC.47) by $z_1^{m_1}$ and $z_2^{m_2}$, and summing up over all m_1 and m_2, we obtain the following balance equations for the generating functions:

$$z_1\theta_1\frac{\partial\Pi_0}{\partial z_1}+z_2\theta_2\frac{\partial\Pi_0}{\partial z_2}+\lambda\Pi_0=\mu_1\Pi_1+\mu_2\Pi_2, \tag{EC.48}$$

$$\Pi_1(\lambda+\mu_1)+z_2\theta_2\frac{\partial\Pi_1}{\partial z_2}=\lambda_1\Pi_0+\theta_1\frac{\partial\Pi_0}{\partial z_1}+z_1\lambda_1\Pi_1, \tag{EC.49}$$

$$\Pi_2(\lambda+\mu_2)=\lambda_2\Pi_0+\theta_2\frac{\partial\Pi_0}{\partial z_2}+z_1\theta_2\frac{\partial\Pi_1}{\partial z_2}+\lambda_2z_1\Pi_1+\lambda_1z_1\Pi_2+\lambda_2z_2\Pi_2. \tag{EC.50}$$

Comparing the two workloads ρ_1 and ρ_2 yields the following result.

PROPOSITION EC.1. *Considering the preemptive $M/M/1$ retrial queues having two customer classes. The workloads are $\rho_1=\lambda_1/\mu_1$ and $\rho_2=\lambda_2/\mu_2$.*

Proof. First, we have $\rho_i=\Pi_i$ for $i=1,2$. In order to find Π_1 and Π_2, we multiply (EC.49) and (EC.50) by z_1 and z_2 respectively and substract them from (EC.48), which yields

$$(\lambda-z_1\lambda_1-z_2\lambda_2)\Pi_0+(z_1(\lambda+\mu_1)-z_1^2\lambda_1-\lambda_2z_1z_2-\mu_1)\Pi_1+((\lambda+\mu_2)z_2-(\lambda_1z_1+\lambda_2z_2)z_2-\mu_2)\Pi_2=0. \tag{EC.51}$$

Letting $z_1=1$ and $z_2=1$, and removing $(1-z_2)$ and $(1-z_1)$ on the both sides of (EC.51) yield

$$\lambda_2\Pi_0+\lambda_2\Pi_1+(\lambda_2z_2-\mu_2)\Pi_2=0, \tag{EC.52}$$

$$\lambda_1\Pi_0-(\mu_1-z_1\lambda_1)\Pi_1+\lambda_1\Pi_2=0. \tag{EC.53}$$

Setting $\lambda_2=\lambda_1=1$ in (EC.52) and (EC.53), we have $\Pi_1=\lambda_1/\mu_1$ and $\Pi_2=\lambda_2/\mu_2$. ∎

Proposition EC.1 shows that the steady-state workloads of the two classes remain unchanged when the service policy becomes preemptive. Hence the *fixed-capacity* constraints in (2) continue to hold under the preemptive rule. When $L(t)=j$, we denote the expected number of type-i customers in orbit i as $L_{j,i}=\partial\Pi_j/\partial z_i|_{z_i=1}$ for $j=0,1,2$ and $i=1,2$. Therefore, the mean number of type-1 customers satisfies $N_1=L_{0,1}+L_{1,1}+L_{2,1}$. We next explain how to compute N_1.

PROPOSITION EC.2. *Considering the preemptive $M/M/1$ retrial queues having two customer classes, the expected number of type-1 orbiting customers is*

$$N_1=A\cdot L_{2,1}+B, \tag{EC.54}$$

where

$$A=\frac{\mu_1\theta_1-\mu_2\theta_2}{-\lambda_1\theta_1+\mu_1\theta_1-\lambda_2\theta_2},$$
$$B=\frac{\lambda_1^2(\lambda_2-\mu_2)\mu_2(\mu_1+\theta_1)+\lambda_1\lambda_2\left(-\mu_2(\mu_1+\mu_2)(\mu_1+\theta_2)+\lambda_2\left(\mu_1^2+\mu_2\theta_2\right)\right)}{\mu_1\mu_2(\lambda_2-\mu_2)(-\lambda_1\theta_1+\mu_1\theta_1-\lambda_2\theta_2)}.$$

Proof. Because type-2 customers are not affected by type-1 customers and $N_2 = L_{0,2} + L_{1,2} + L_{2,2}$, we have $L_{0,2} + L_{1,2} = \lambda_2^2/(\theta_2\mu_2)$ and $L_{2,2} = (\theta_2 + \lambda_2)\lambda_2^2/(\mu_2\theta_2(\mu_2 - \lambda_2))$. From (EC.49) and (EC.50), we have

$$L_{0,2} = \frac{\lambda(1-\Pi_0) - \theta_1 L_{0,1}}{\theta_2}, \qquad L_{1,2} = \frac{\theta_1 L_{0,1} + \lambda_1\Pi_1 + \lambda_1\Pi_0 - \Pi_1(\lambda + \mu_1)}{\theta_2}. \tag{EC.55}$$

That is, both $L_{0,2}$ and $L_{1,2}$ are functions of $L_{0,1}$. Differentiating (EC.53) on both sides with respect to z_1 and and setting $z_1 = 1$ yield

$$\lambda_1 L_{0,1} - (\mu_1 - \lambda_1)L_{1,1} + \lambda_1\Pi_1 + \lambda_1 L_{2,1} = 0. \tag{EC.56}$$

Letting $z_1 = z_2 = z$ in (EC.51), we have

$$\begin{aligned} & \lambda(1-z)\Pi_0 + (z(\lambda+\mu_1) - z^2\lambda - \mu_1)\Pi_1 + ((\mu_2+\lambda)z - \lambda z^2 - \mu_2)\Pi_2 = 0 \\ \Leftrightarrow\ & \lambda\Pi_0 + (\lambda z - \mu_1)\Pi_1 + (\lambda z - \mu_2)\Pi_2 = 0 \\ \Leftrightarrow\ & \lambda(L_{01} + L_{02}) + \lambda\Pi_1 + (\lambda - \mu_1)(L_{11} + L_{12}) + (\lambda - \mu_2)(L_{21} + L_{22}) + \lambda\Pi_2 = 0. \end{aligned} \tag{EC.57}$$

Plugging (EC.55) into (EC.57) and combining (EC.56) and (EC.57), we can express both $L_{0,1}$ and $L_{1,1}$ as functions of $L_{2,1}$. Equation (EC.54) is obtained using the relation $N_1 = L_{0,1} + L_{1,1} + L_{2,1}$. ■

It now remains to compute L_{21}. In the rest of this section, we develop a procedure to compute the stationary distribution of $p_{(2,m_1,m_2)}$ for $m_1, m_2 \geq 0$. We define the marginal distribution as

$$p_{(0,\cdot,m_2)} = \sum_{m_1\geq 0} p_{(0,m_1,m_2)}, \quad p_{(1,\cdot,m_2)} = \sum_{m_1\geq 0} p_{(1,m_1,m_2)}, \quad p_{(2,\cdot,m_2)} = \sum_{m_1\geq 0} p_{(2,m_1,m_2)}.$$

The exact marginal distribution for type-2 customers is given as follow (Artalejo et al. (2001)):

$$p_{(2,\cdot,m_2)} = \sum_{m_1\geq 0} p_{(2,m_1,m_2)} = \frac{\rho_2^{m_2+1}}{m_2!\theta_2^{m_2}}(1-\rho_2)^{1+\frac{\lambda_2}{\theta_2}} \prod_{n=1}^{m_2}(\lambda_2 + n\theta_2), \tag{EC.58}$$

$$p_{(0,\cdot,m_2)} + p_{(1,\cdot,m_2)} = \sum_{m_1\geq 0} p_{(2,m_1,m_2)} = \frac{\rho_2^{m_2}}{m_2!\theta_2^{m_2}}(1-\rho_2)^{1+\frac{\lambda_2}{\theta_2}} \prod_{n=0}^{m_2-1}(\lambda_2 + n\theta_2). \tag{EC.59}$$

We truncate the type-1 and type-2 orbit queues by K and M, respectively. For a given large number K and a certain prespecified error parameter $\epsilon > 0$, the minimal M can be determined as follow

$$M = \min\{M | \sum_{m_1=0}^{M} p_{(0,\cdot,m_2)} + p_{(1,\cdot,m_2)} + p_{(2,\cdot,m_2)} > 1 - \epsilon\}. \tag{EC.60}$$

We write (EC.45)–(EC.47) in the vector notations:

$$\mathbf{p}_{0,m_2}\mathbf{A}_{m_2} = \mu_1\mathbf{p}_{1,m_2} + \mu_2\mathbf{p}_{2,m_2}, \tag{EC.61}$$

$$\mathbf{p}_{1,m_2}\mathbf{B}_{m_2} = \mathbf{p}_{0,m_2}C_{m_2}, \tag{EC.62}$$

$$\mathbf{p}_{2,m_2}\mathbf{D}_{m_2} = \lambda_2\mathbf{p}_{0,m_2} + \mathbf{p}_{0,m_2+1}\mathbf{E}_{m_2} + \mathbf{p}_{1,m_2}\mathbf{F}_{m_2} + \lambda_2\mathbf{p}_{2,m_2-1} + \mathbf{p}_{1,m_2+1}\mathbf{G}_{m_2}, \tag{EC.63}$$

where $m_2 = 0, 1, ..., M$ and $\mathbf{p}_{i,m_2} = (p_{(i,0,m_2)}, p_{(i,1,m_2)}, ..., p_{(i,K,m_2)})$, $i = 0, 1, 2$, $\mathbf{A}_{m_2}^{(i,j)} = \lambda + (i-1)\theta_1 + m_2\theta_2$ for $i = j$ and $1 \le i \le K+1$, $\mathbf{E}_{m_2}^{(i,j)} = (m_2+1)\theta_2$ for $i = j$ and $1 \le i \le K+1$, $\mathbf{F}_{m_2}^{(i,j)} = \lambda_2$ for $j = i+1$ and $1 \le i \le K$, $\mathbf{G}_{m_2}^{(i,j)} = (m_2+1)\theta_2$ for $j = i+1$ and $1 \le i \le K$,

$$\mathbf{B}_{m_2}^{(i,j)} = \begin{cases} \lambda + \mu_1 + m_2\theta_2, & \text{for } i = j \text{ and } 1 \le i \le K+1; \\ -\lambda_1, & \text{for } j = i+1 \text{ and } 1 \le i \le K; \\ 0, & \text{else.} \end{cases}$$

$$\mathbf{C}_{m_2}^{(i,j)} = \begin{cases} \lambda_1, & \text{for } i = j \text{ and } 1 \le i \le K+1; \\ (i-1)\theta_1, & \text{for } i = j+1 \text{ and } 2 \le i \le K+1; \\ 0, & \text{else.} \end{cases}$$

$$\mathbf{D}_{m_2}^{(i,j)} = \begin{cases} \lambda_2 + \mu_2, & \text{for } i = j \text{ and } 1 \le i \le K+1; \\ -\lambda_1, & \text{for } j = i+1 \text{ and } 1 \le i \le K; \\ 0, & \text{else.} \end{cases}$$

From (EC.61) and (EC.62), we have

$$\mathbf{p}_{0,m_2} = \mu_2 \mathbf{p}_{2,m_2} (\mathbf{A}_{m_2} - \mathbf{C}_{m_2} \mathbf{B}_{m_2^{-1}} \mu_1)^{-1}, \tag{EC.64}$$

$$\mathbf{p}_{1,m_2} = \mu_2 \mathbf{p}_{2,m_2} (\mathbf{A}_{m_2} - \mathbf{C}_{m_2} \mathbf{B}_{m_2^{-1}} \mu_1)^{-1} \mathbf{C}_{m_2} \mathbf{B}_{m_2}^{-1}. \tag{EC.65}$$

Substituting (EC.64) and (EC.65) into (EC.63), we can obtain

$$\mathbf{p}_{2,m_2} \Theta_{m_2} = \mathbf{p}_{2,m_2+1} \Delta_{m_2+1} + \lambda_2 \mathbf{p}_{2,m_2-1}, \tag{EC.66}$$

for $0 \le m_2 \le M$, where $\Theta_{m_2} = \mathbf{D}_{m_2} - \lambda_2(\mu_2(\mathbf{A}_{m_2} - \mathbf{C}_{m_2}\mathbf{B}_{m_2}^{-1}\mu_1)^{-1}) - (\mu_2(\mathbf{A}_{m_2} - \mathbf{C}_{m_2}\mathbf{B}_{m_2}^{-1}\mu_1)^{-1})\mathbf{C}_{m_2}\mathbf{B}_{m_2}^{-1}\mathbf{F}_{m_2}$ and $\Delta_{m_2+1} = (\mu_2(\mathbf{A}_{m_2+1} - C_{m_2+1}\mathbf{B}_{m_2+1}^{-1}\mu_1)^{-1})(\mathbf{E}_{m_2} + \mathbf{C}_{m_2+1}\mathbf{B}_{m_2+1}^{-1}\mathbf{G}_{m_2})$.

In summary, we can compute the distribution $\mathbf{p}_{2,m_2}$ for $0 \le m_2 \le M$ by follow Algorithm 2 below.

Algorithm 2

Step 1. Calculate $p_{(2,\cdot,m_2+1)}$ and $p_{(0,\cdot,m_2+1)} + p_{(1,\cdot,m_2+1)}$ for $0 \le m_2 \le M$ from (EC.58) and (EC.59), where M is determined by (EC.60) for any given small ϵ.

Step 2. Let $\mathbf{p}_{2,M+1}^* = \frac{p_{2,\cdot,M+1}}{M+1} e^T$.

Step 3. Take $\overline{\Theta}_{M-1} = \frac{\Theta_M}{\lambda_2}$, $\overline{\Theta}_{M-2} = \overline{\Theta}_{M-2}\frac{\Theta_{M-1}}{\lambda_2} - \frac{\Delta_M}{\lambda_2}$.

Step 4. Calculate $\overline{\Theta}_{m_2} = \frac{\overline{\Theta}_{m_2+1}\Theta_{m_2+1} - \overline{\Theta}_{m_2+2}\Delta_{m_2+2}}{\lambda_2}$ and $\overline{\Delta}_{m_2} = \frac{\overline{\Delta}_{m_2+1}\Theta_{m_2+1} - \overline{\Delta}_{m_2+2}\Delta_{m_2+2}}{\lambda_2}$ for $0 \le m_2 \le M-3$

Step 5. $\mathbf{p}_{2,M}^* = (\overline{\Delta}_1\Delta_1 - \overline{\Delta}_0\Theta_0)(\overline{\Theta}_0\Theta_0 - \overline{\Theta}_1\Delta_1)^{-1}$.

Step 6. $\mathbf{p}_{2,m_2}^* = \mathbf{p}_{2,M}^*\overline{\Theta}_i + \overline{\Delta}_i$ for $0 \le m_2 \le M-1$.

Step 7. $L_{21} = \sum_{m_2=0}^{M} \mathbf{p}_{2,m_2}^* \cdot \beta$, where $\beta = (0, 1, 2, ..., K)^T$.

The initial value in Step 2 of Algorithm 2 can be estimated via the principle of the maximum entropy. We consider an example to illustrate this algorithm. For $\theta_1 = 0.6111$, $\theta_2 = 1.8333$,

$\lambda_1 = 0.21083,\ \lambda_2 = 0.2917,\ \mu_1 = 0.9097,\ \mu_2 = 1.0763,\ p_1 = 0.4167,\ p_2 = 0.5833$, the approximate distribution and exact distribution are given in Table EC.1, with the error parameter $\epsilon = 0.005$ $(M = 5, K = 35)$.

Table EC.1 The comparison between approximated distribution and stationary distribution when $\rho = 0.5$

	$p_{2,\cdot,0}$	$p_{2,\cdot,1}$	$p_{2,\cdot,2}$	$p_{2,\cdot,3}$	$p_{2,\cdot,4}$	$p_{2,\cdot,5}$
Approximate distribution	0.1851	0.0581	0.0170	0.0049	0.0014	3.8357E-4
Exact distribution	0.1879	0.0590	0.0173	0.0049	0.0014	3.8803E-4

Table EC.1 shows that the desired accuracy can be achieved when $M = 5$, which also implies that $L_{2,1} = 0.4078$ and $N_1 = 1.0022$. It is noted that the approximated distribution are less than the stationary distributions, i.e., $\mathbf{p}^*_{2,m} \le \mathbf{p}_{2,m}$, due to the finite truncation of the orbit queue. To normalize the probabilities so that they add up to 1, we may add an additional step between Step 6 and Step 7, namely, $\mathbf{p}^*_{2,m} = (p_{2,;m}/\mathbf{p}^*_{2,m}\mathbf{e})\mathbf{p}^*_{2,m}$.

However, when ρ is large, the value of the truncation K and M should be more carefully selected. For example, when $\rho = 0.9$, we set $K = 200$ and $M = 15$ to keep the error within the tolerance, in which $L_{21} = 11.7592$, $N_1 = 20.7155$. Therefore, by carefully selecting the truncated values K and M, desired accuracy can be achieved.

Plugging $L_{2,1}$ into (EC.54), we obtain the expected number of type-1 customers N_1. Next, the mean delay of type-1 customers can be determined using Little's law $w_1 = N_1/\lambda_1$. Following (EC.44), the expected total orbiting time and total number of trials for all customers can be derived as $w^P_{SSRD} = w_1p_1 + w_2p_2$ and $r^P_{SSRD} = w_1\theta_1p_1 + w_2\theta_2p_2$. In Figure EC.2, we plot delays and number of trials as a function of C, with $\rho = 0.7, 0.9$. Because C ranges from 0.01 to 100, the case $C < 1$ $(C > 1)$ represents the case where class 1 (class 2) has a higher priority.

Remark EC.1. Considering the preemptive $M/M/1$ retrial queues with two customer types.

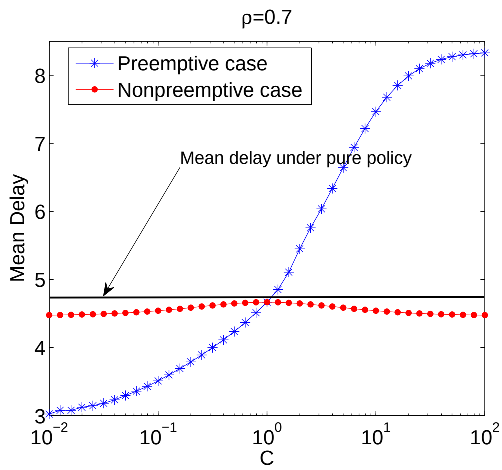
- If type-2 customers receive a higher priority, then

$$w^P_{SSRD} > w_0 > w_{SSRD}, \qquad r^P_{SSRD} < r_0 < r_{SSRD}.$$

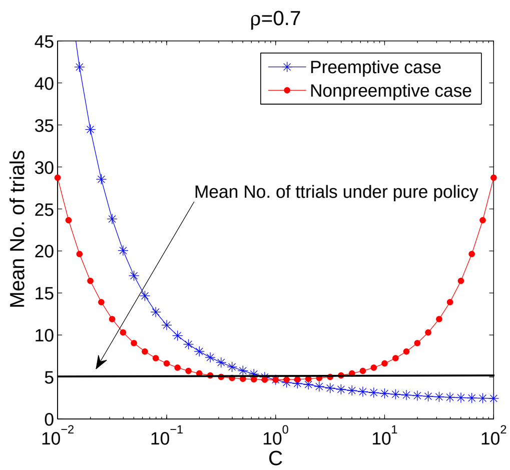
- If type-1 customers receive a higher priority, then

$$w^P_{SSRD} < w_{SSRD} < w_0, \qquad r^P_{SSRD} > r_{SSRD} > r_0.$$

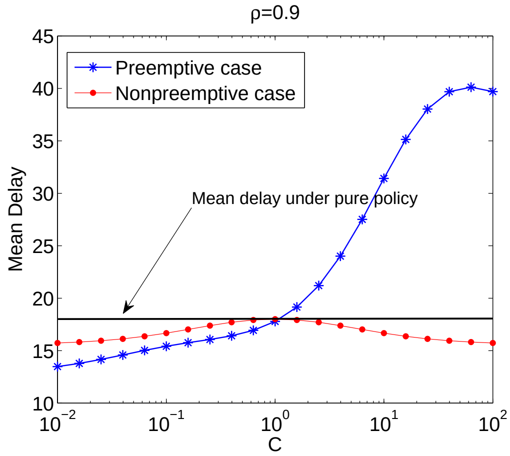
Specifically, a preemptive priority (with type-2 customers receiving a higher priority) can reduce the number of trials but increases the overall delay. On the other hand, a preemptive priority (with type-1 customers receiving a higher priority) will increase the number of trials but reduce the overall delay. In summery, the *preemptive* differentiation policy cannot reduce both the delay and number of trials simultaneously as in our *non-preemptive* SSRD policy.



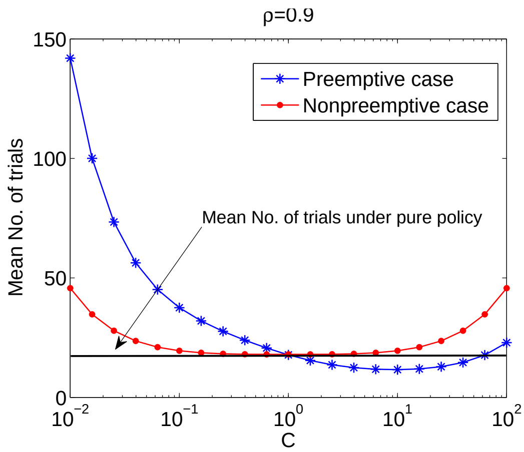
(a)



(b)



(c)



(d)

Figure EC.2 The comparison of preemptive and non-preemptive case under SSRD

EC.3. Comparison to Xu et al. (2015)

EC.3.1. Monotonicity of variability

To support the discussion in Part (c) of Remark 5, we compare the variance of delay in the $M/G/1$ model in Xu et al. (2015) and in our $M/G/1$ retrial model under SSRD.

First, following §4.1 of Kella and Yechiali (1988), we obtain the variance of delay in Xu et al. (2015) by $Var[W] = \sum_{k=1}^{m} p_k(E[W_k^2] - E^2[W_k])$, where

$$E[W_k] = \frac{\sum_{i=1}^{m} \lambda_i E[S_i^2]}{2(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)},$$

$$E[W_k^2] = \left[\left(\frac{\sum_{i=1}^{k}\lambda_i E[S_i^2]}{1-\sum_{i=1}^{k}\rho_i} + \frac{\sum_{i=1}^{k-1}\lambda_i E[S_i^2]}{1-\sum_{i=1}^{k-1}\rho_i}\right) E[W_k] + \frac{\sum_{i=1}^{m}\lambda_i E[S_i^3]}{3(1-\sum_{i=1}^{k}\rho_i)(1-\sum_{i=1}^{k-1}\rho_i)}\right] \frac{1}{(1-\sum_{i=1}^{k-1}\rho_i)}.$$

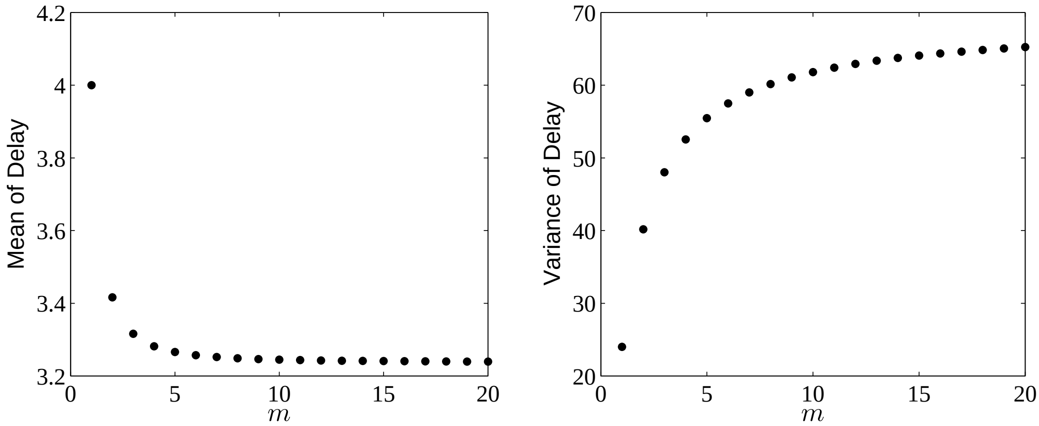


Figure EC.3 Mean and variance of delay of the $M/M/1$ model in Xu et al. (2015) as a function of the service grade m, with $\mu_0 = 1$ and $\lambda_0 = 0.9$.

Figure EC.3 illustrates the mean and variance of delay when $\mu_0 = 1$ and $\lambda_0 = 0.9$. We observe that the mean and variance of delay are decreasing and increasing in the number of service grades m, respectively, under the optimal differentiation policy (Corollary 3 and (24) of Xu et al. (2015)). That is, the delay can be further reduced when the variance increases (which occurs when the service grades m increases).

Next we study the monotonicity of the variance of delay for our $M/G/1$ retrial queue under SSRD. Consider $m \geq 2$ service grades, with the maximum ODR $C = \theta_m/\theta_1$. Let $C_i = 1 + (i-1)(C-1)/(m-1)$ and $\rho_i = \rho/m$ for $i = 1, 2, \ldots, m$. According to the optimal allocation (13) and constraint (2), we have $p_i = (1/\sqrt{x_i})/(\sum_{j=1}^m 1/\sqrt{x_j})$, $\mu_i = \lambda_i/\rho_i = m\lambda_0 p_i/\rho$ and $\theta_i = C_i\theta_0 \sum_{j=1}^m p_i/C_i$. For $C = 5$, we examine the mean and variance of delay as functions of the number of service grades m. Figure EC.4 shows that the mean (variance) of delay significantly decreases (increases) as $m = 1$ increases from 1 to 2. However, the variance (mean) of delay becomes decreasing (increasing) in m when $m \geq 2$. Indeed, the minimum mean delay is achieved at $m = 2$ (which is consistent with our main result in Theorem 2), which yields the maximum variance of delay. Similar to results in Xu et al. (2015), the reduction of delay benefits from the increased variance, which now decreases as m increases when $m \geq 2$.

Finally, we use simulations to demonstrate the growth of the variance of delay in C. Figure EC.5 shows that, under the optimal SSRD given by Theorem 2 with $m = 2$, the variance of delay is increasing in the ORD C, then the mean of delay is decreasing in C, which is consistent with Proposition 2.

EC.3.2. Limiting distribution of the random service rate in Xu et al. (2015)

We hereby provide support to Part (b) of Remark 5. It has been shown in Xu et al. (2015) that creating service variability can reduce the mean waiting time in an $M/G/1$ queue. Especially, the

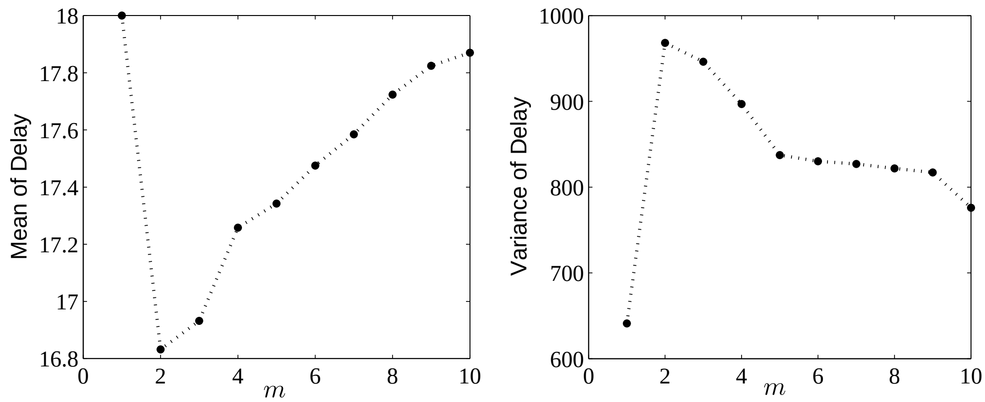


Figure EC.4 Mean and variance of delay of the $M/M/1$ retrial model under SSRD as a function of the service grade m, with $\theta_0 = \mu_0 = 1$ and $\lambda_0 = 0.9$

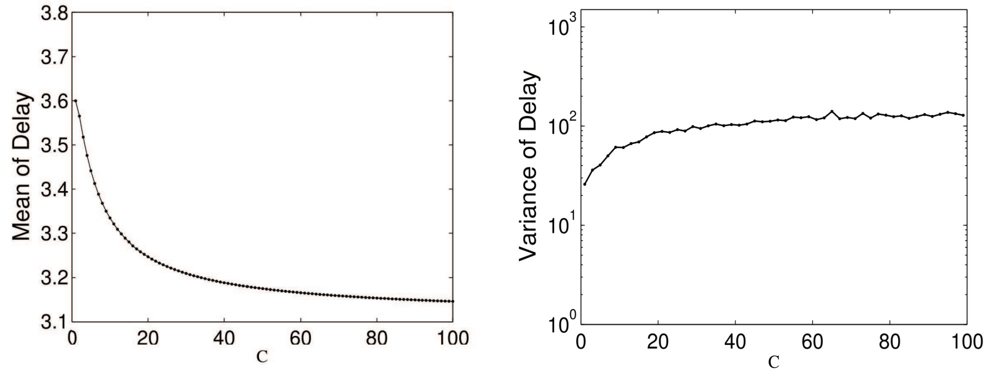


Figure EC.5 Mean and variance of delay of the $M/M/1$ retrial model under SSRD as a function of C, with $\theta_0 = \mu_0 = 5$ and $\lambda_0 = 0.9$

optimal performance can be achieved when the number of service grade $m \to \infty$. We discovered that the optimal case ($m \to \infty$) yields a nice continuous distribution for the random service rate.

Proposition EC.3 **(Limiting continuous service-rate distribution in Xu et al. (2015))**. *Under the optimal service allocation policy in Xu et al. (2015), the service provider offers a random service rate $\mathcal{M}$, where $\mathcal{M}$ is a random variable following a continuous distribution with bounded support, having probability density*

$$f_{\mathcal{M}}(a) = \frac{2-\rho}{2\rho\mu_0^2}\ a, \qquad a \in \mathbb{S} \equiv \left(\underline{\mu}_\rho, \bar{\mu}_\rho\right) \equiv \left(\frac{2\mu_0(1-\rho)}{2-\rho}, \frac{2\mu_0}{2-\rho}\right). \tag{EC.67}$$

Remark EC.2. First, it is easy to check that the density function given above is indeed well defined, that is, $\int_{a\in\mathbb{S}} f_{\mathcal{M}}(a)da = 1$. Apparently the base service rate μ_0 is in the interior of $\mathbb{S}$, because $\frac{2\mu_0(1-\rho)}{2-\rho} < \mu_0 < \frac{2\mu_0}{2-\rho}$. The spread of the support increases in ρ. Specifically, $\mathbb{S}$ becomes the interval $(0, 2\mu_0)$ as $\rho \to 1$, and $\mathbb{S}$ degenerates to a single point μ_0 as $\rho \to 0$.

Proof: Suppose there are K customer grades. By (24) of Proposition 4 (p.241) in Xu et al. (2015), we know that the optimal service rate assignment satisfies

$$p_k = p_1[(1-\rho)^{\frac{2}{m}}]^{k-1}, \tag{EC.68}$$

$$\mu_k = \mu_1[(1-\rho)^{\frac{1}{m}}]^{k-1}, \quad 1 \le k \le m-1. \tag{EC.69}$$

Normality of $p_1, \ldots, p_K$ implies

$$1 = p_1 + \cdots + p_m = p_1[1 + (1-\rho)^{\frac{2}{m}} + \ldots + (1-\rho)^{\frac{2(k-1)}{m}}] \quad \Rightarrow \quad p_1 = \frac{1-(1-\rho)^{\frac{2}{m}}}{1-(1-\rho)^2}. \tag{EC.70}$$

Similarly, the equal mean condition, along with (EC.68)–(EC.70) imply that

$$\frac{1}{\mu_0} = \frac{p_1}{\mu_1} + \cdots + \frac{p_m}{\mu_m} = \frac{p_1}{\mu_1}\sum_{k=1}^{m}(1-\rho)^{\frac{k-1}{m}} \quad \Rightarrow \quad \mu_1 = \frac{\mu_0\rho}{1-(1-\rho)^2} \cdot \frac{1-(1-\rho)^{\frac{2}{m}}}{1-(1-\rho)^{\frac{1}{m}}}. \tag{EC.71}$$

Let $\mu_1(m)$ be the μ_1 in (EC.71), we have

$$\begin{aligned}
\lim_{m\to\infty}\mu_1(m) &= \frac{\mu_0\rho}{1-(1-\rho)^2}\lim_{m\to\infty}\frac{1-(1-\rho)^{\frac{2}{m}}}{1-(1-\rho)^{\frac{1}{m}}} = \frac{\mu_0\rho}{1-(1-\rho)^2}\lim_{x\to 0}\frac{1-(1-\rho)^{2x}}{1-(1-\rho)^{x}} \\
&= \frac{\mu_0\rho}{1-(1-\rho)^2}\lim_{x\to 0}\frac{-2(1-\rho)^{2x}\log(1-\rho)}{-(1-\rho)^{x}\log(1-\rho)} = \frac{2\mu_0\rho}{1-(1-\rho)^2} = \frac{2\mu_0}{2-\rho} \equiv \mu_1(\infty), \qquad \text{(EC.72)} \\
\lim_{m\to\infty}\mu_m(m) &= \lim_{m\to\infty}\mu_1(m)[(1-\rho)^{\frac{1}{m}}]^{m-1} = \mu_1(\infty)(1-\rho). \qquad \text{(EC.73)}
\end{aligned}$$

Now let the random variable $\mathcal{M}_m$ denote the **random** service rate offered to an arbitrary customer where there are m service grades. According to (EC.72)–(EC.73), we know that as $m \to \infty$, $\mathcal{M}_m$ asymptotically has a bounded domain $\mathbb{S}$ given by (EC.67).

We next show that $\mathcal{M}_m \Rightarrow \mathcal{M}_\infty \equiv \mathcal{M}$ as $m \to \infty$, where the limiting random variable $\mathcal{M}_\infty$ has a continuous support in $\mathbb{S}$. Pick $a \in \mathbb{S}$ and a small $h > 0$, then

$$P(\mathcal{M}_m \in (a, a+h)) = \sum_{k=1}^{m}\mathbf{1}_{\{\mu_k(m)\in(a,a+h)\}} \cdot p_k(m),$$

where $\mu_k(m)$ and $p_k(m)$ are the μ_k and p_k given in (EC.68) and (EC.69). According to (EC.69), we have

$$a < \mu_k = \mu_1(m)(1-\rho)^{\frac{k-1}{m}} < a+h \quad \Leftrightarrow \quad \bar{k}_m \equiv \frac{m\log(\frac{a}{\mu_1(m)})}{\log(1-\rho)} + 1 > k > \frac{m\log(\frac{a+h}{\mu_1(m)})}{\log(1-\rho)} + 1 \equiv \underline{k}_m.$$

Hence, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{M}_m \in (a, a+h)) &= \sum_{k=\underline{k}_m+1}^{\bar{k}_m}\frac{1-(1-\rho)^{\frac{2}{m}}}{1-(1-\rho)^2}[(1-\rho)^{\frac{2}{m}}]^{k-1} \\
&= \frac{1-(1-\rho)^{\frac{2}{m}}}{1-(1-\rho)^2}[(1-\rho)^{\frac{2}{m}}]^{\underline{k}_m} \cdot \frac{1-[(1-\rho)^{\frac{2}{m}}]^{\bar{k}_m-\underline{k}_m-1}}{1-(1-\rho)^{\frac{2}{m}}} \\
&= \frac{(1-\rho)^{\frac{2\log\left(\frac{a+h}{\mu_1(m)}\right)}{\log(1-\rho)}}}{1-(1-\rho)^2}\left(1-(1-\rho)^{\frac{2\left[\log\left(\frac{a}{\mu_1(m)}\right)-\log\left(\frac{a+h}{\mu_1(m)}\right)\right]}{\log(1-\rho)}}\right) \\
&= \frac{\left(\frac{a+h}{\mu_1(m)}\right)^2}{1-(1-\rho)^2}\left(1-\left(\frac{a}{a+h}\right)^2\right).
\end{aligned}$$

Now letting $m \to \infty$ yields

$$\mathbb{P}(\mathcal{M} \in (a, a+h)) = \lim_{m\to\infty} \mathbb{P}(\mathcal{M}_m \in (a, a+h)) = \frac{\left(\frac{a+h}{\mu_1(\infty)}\right)^2}{1-(1-\rho)^2}\left(1-\left(\frac{a}{a+h}\right)^2\right) = \frac{\left(\frac{(a+h)(2-\rho)}{2\mu_0}\right)^2}{1-(1-\rho)^2}\frac{(2a+h)h}{(a+h)^2},$$

where the last equality holds by (EC.72). The probability density function of $\mathcal{M}$ is given by

$$f_{\mathcal{M}}(a) = \lim_{h\downarrow 0} \frac{\mathbb{P}(\mathcal{M} \in (a, a+h))}{h} = \lim_{h\downarrow 0} \frac{\left(\frac{(a+h)(2-\rho)}{2\mu_0}\right)^2}{1-(1-\rho)^2}\frac{(2a+h)}{(a+h)^2} = \frac{2-\rho}{2\rho\mu_0^2}\, a, \qquad a \in \mathbb{S}. \qquad \blacksquare$$

EC.4. Additional Simulations

Nonexponential retrial times. In this paper, we have treated a retrial model with general service times but exponential orbit times. We have showed that the dominance condition (10) is independent with the structure of the service-time distribution beyond its mean. Hence, we conjecture that condition (10) continues to hold for nonexponential orbit times. In the future, we plan to extend to models with nonexponential orbit times. We conduct simulation experiments in Table EC.2 for the $M/H_2/1$ model with *2-phase hyperexponential* (H_2) service times (mixture of two exponential distributions) and H_2 orbit times with SCV $c_s^2 = c_r^2 = 4$, $\theta_0 = \mu_0 = 1$. Table EC.2 shows that SSRD achieves a smaller average delay than homogeneous service when the traffic intensity ρ is close to 1. Specifically, the RRD of SSRD with respect to homogeneous service is 13.3% (13.7%, 4.6%) for $\rho = 0.972$ (0.95, 0.9). But RRD is negative when $\rho \le 0.8$.

Table EC.2 Comparing SSRD and homogeneous service for the $M/H_2/1$ model with H_2 retrial times.

	Homogeneous service			Differentiated service		
ρ	E[No. waiting]	E[No. in service]	E[delay]	E[No. waiting]	E[No. in service]	E[delay]
0.975	100.94±9.67	0.97±4.1E-3	105.03±10.20	94.86±8.67	0.97±4.1E-3	91.03±9.07
rel. diff.	-	-	-	6.02%	0%	**13.33%**
0.95	65.17±7.15	0.95±5.4E-3	68.92±7.54	57.26±5.64	0.95±5.5E-3	59.45±5.65
rel. diff.	-	-	-	12.14%	0%	**13.74%**
0.9	29.20±2.06	0.90±6.4E-3	32.02±2.22	27.92±2.10	0.90±6.6E-3	30.55±2.19
rel. diff.	-	-	-	4.42%	0%	**4.61%**
0.8	11.57±0.60	0.80±6.4E-3	14.53±0.77	11.86±0.66	0.80±6.5E-3	14.72±0.75
rel. diff.	-	-	-	-2.51%	0%	-1.38%
0.7	6.26±0.27	0.70±6.4E-3	8.97±0.37	6.43±0.28	0.70±6.7E-3	9.15±0.38
rel. diff.	-	-	-	-2.64%	0%	-1.97%

Multiple server. Multiserver queueing models have been proven more practical for modeling realistic service systems. Therefore, in the future we plan to extend our service-differentiation policy from single-server framework to multi-server models. We next conduct a simulation example for an $M/M/2$ retrial queueing system. In Table EC.3 we observe that SSRD helps reduce the average delay when the traffic intensity is close to 1. All simulations are conducted with 95% confidence intervals.

Table EC.3 Comparing performance of SSRD and homogeneous service for the 2-server $M/M/2$ model.

	Pure service			Differentiated service		
ρ	E[No. waiting]	E[No. in service]	E[delay]	E[No. waiting]	E[No. in service]	E[delay]
0.975	77.43±3.71	1.95±2.0E-3	39.71±1.89	72.32±3.50	1.95±2.1E-3	36.94±1.77
rel. diff.	-	-	-	6.61%	0%	**6.96%**
0.95	36.16±0.91	1.90±2.1E-3	19.08±0.48	34.71±0.90	1.90±2.1E-3	18.26±0.48
rel. diff.	-	-	-	3.95%	0%	**4.30%**
0.9	15.83±0.23	1.80±2.1E-3	8.82±0.13	15.54±0.23	1.80±2.2E-3	8.64±0.12
rel. diff.	-	-	-	1.83%	0%	**2.11%**
0.8	6.09±5.2E-2	1.60±2.4E-3	3.80±3.2E-2	6.11±5.4E-2	1.60±2.4E-3	3.81±3.5E-2
rel. diff.	-	-	-	-0.36%	0%	-0.42%
0.7	2.94±1.9E-2	1.40±2.0E-3	2.10±1.3E-2	2.99±2.0E-2	1.40±2.3E-3	2.14±1.6E-2
rel. diff.	-	-	-	-1.50%	0%	-1.81%