Appendix A Proof of Propositions

A.1 Proof of Proposition 1

The expected wait time under the dedicated delivery system is

$$\begin{aligned} W_{\mathrm{S}}^{\mathrm{D}} &= \frac{\lambda_1}{\lambda_1+\lambda_2} W_{\mathrm{S},1}^{\mathrm{D}} + \frac{\lambda_2}{\lambda_1+\lambda_2} W_{\mathrm{S},2}^{\mathrm{D}} \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2} \frac{r}{k_1 - 2r\lambda_1} \left(\frac{2r\lambda_1}{k_1}\right)^{\sqrt{2(k_1+1)}-1} + \frac{\lambda_2}{\lambda_1+\lambda_2} \frac{r}{k_2 - 2r\lambda_2} \left(\frac{2r\lambda_2}{k_2}\right)^{\sqrt{2(k_2+1)}-1}, \end{aligned}$$

and the expected wait time under the fully flexible delivery system is

$$W_{\mathrm{S}}^{\mathrm{F}} = \frac{r}{k_1+k_2-2r(\lambda_1+\lambda_2)} \left(\frac{2r(\lambda_1+\lambda_2)}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}-1}.$$

We define $\rho_{\mathrm{S},1}^{\mathrm{D}} = \frac{\lambda_1}{k_1\mu_{\mathrm{S}}} = \frac{2r\lambda_1}{k_1}$ and $\rho_{\mathrm{S},2}^{\mathrm{D}} = \frac{\lambda_2}{k_2\mu_{\mathrm{S}}} = \frac{2r\lambda_2}{k_2}$. For notational simplicity, $\rho_{\mathrm{S},1}^{\mathrm{D}}$ and $\rho_{\mathrm{S},2}^{\mathrm{D}}$ are simplified to ρ_1 and ρ_2, respectively, throughout the proof process in this section. We have

$$W_{\mathrm{S}}^{\mathrm{D}} - W_{\mathrm{S}}^{\mathrm{F}} = \frac{\lambda_1}{\lambda_1+\lambda_2} \frac{r\rho_1^{\sqrt{2(k_1+1)}-1}}{k_1(1-\rho_1)} + \frac{\lambda_2}{\lambda_1+\lambda_2} \frac{r\rho_2^{\sqrt{2(k_2+1)}-1}}{k_2(1-\rho_2)} - \frac{r\left(\frac{\rho_1 k_1+\rho_2 k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}-1}}{k_1(1-\rho_1)+k_2(1-\rho_2)},$$

which must be larger than or equal to 0 if the following inequality holds:

$$\lambda_1\rho_1^{\sqrt{2(k_1+1)}-1} + \lambda_2\rho_2^{\sqrt{2(k_2+1)}-1} - (\lambda_1+\lambda_2)\left(\frac{\rho_1 k_1+\rho_2 k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}-1} \geq 0. \tag{A1}$$

Since $\lambda_1 = \frac{k_1\rho_1}{2r}$ and $\lambda_2 = \frac{k_2\rho_2}{2r}$, the above inequality in (A1) can be re-written as

$$\frac{k_1\rho_1}{2r}\rho_1^{\sqrt{2(k_1+1)}-1} + \frac{k_2\rho_2}{2r}\rho_2^{\sqrt{2(k_2+1)}-1} - \frac{k_1\rho_1+k_2\rho_2}{2r}\left(\frac{\rho_1 k_1+\rho_2 k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}-1} \geq 0,$$

which is equivalent to

$$\frac{k_1}{k_1+k_2}\rho_1^{\sqrt{2(k_1+1)}} + \frac{k_2}{k_1+k_2}\rho_2^{\sqrt{2(k_2+1)}} - \left(\frac{\rho_1 k_1+\rho_2 k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}} \geq 0. \tag{A2}$$

Since $f(\rho) = \rho^{\sqrt{2(k+1)}}$ is a convex function, by Jensen's inequality, we can obtain

$$\left(\frac{\rho_1 k_1+\rho_2 k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}} \leq \frac{k_1}{k_1+k_2}\rho_1^{\sqrt{2(k_1+k_2+1)}} + \frac{k_2}{k_1+k_2}\rho_2^{\sqrt{2(k_1+k_2+1)}}. \tag{A3}$$

Therefore, we can verify the inequality in (A1) indeed holds since

$$\begin{aligned} &\frac{k_1}{k_1+k_2}\rho_1^{\sqrt{2(k_1+1)}} + \frac{k_2}{k_1+k_2}\rho_2^{\sqrt{2(k_2+1)}} - \left(\frac{\rho_1 k_1+\rho_2 k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}} \\ \geq\ &\frac{k_1}{k_1+k_2}\left(\rho_1^{\sqrt{2(k_1+1)}-1} - \rho_1^{\sqrt{2(k_1+k_2+1)}-1}\right) + \frac{k_2}{k_1+k_2}\left(\rho_2^{\sqrt{2(k_2+1)}-1} - \rho_2^{\sqrt{2(k_1+k_2+1)}-1}\right) \geq 0 \end{aligned}$$

where the first inequality follows (A3), and the second inequality follows $\rho_1^{\sqrt{2(k_1+k_2+1)}} < \rho_1^{\sqrt{2(k_1+1)}}$ and $\rho_2^{\sqrt{2(k_1+k_2+1)}} < \rho_1^{\sqrt{2(k_2+1)}}$. As a result, we have $W_{\mathrm{S}}^{\mathrm{D}} - W_{\mathrm{S}}^{\mathrm{F}} \geq 0$, and this completes the proof.

$\square$

When the two zones are asymmetric in terms of traveling distance, i.e., $r_1 \neq r_2$, Proposition 1 still holds. A similar line of reasoning proceeds as follows.

The service rates for the two zones are $1/(2r_1)$ and $1/(2r_2)$, respectively. For the fully flexible system, the service rate $\mu_{\mathrm{S}}^{\mathrm{F}}$ is $\frac{1}{\frac{\lambda_1}{\lambda_1+\lambda_2}2r_1 + \frac{\lambda_2}{\lambda_1+\lambda_2}2r_2}$, which is $\frac{\lambda_1+\lambda_2}{2(\lambda_1 r_1+\lambda_2 r_2)}$. The summation of coefficients of variation remains 1. The expected wait time under the dedicated delivery system is

$$W_{\mathrm{S}}^{\mathrm{D}} = \frac{\lambda_1}{\lambda_1+\lambda_2} \frac{r_1}{k_1 - 2r_1\lambda_1} \left(\frac{2r_1\lambda_1}{k_1}\right)^{\sqrt{2(k_1+1)}-1} + \frac{\lambda_2}{\lambda_1+\lambda_2} \frac{r_2}{k_2 - 2r_2\lambda_2} \left(\frac{2r_2\lambda_2}{k_2}\right)^{\sqrt{2(k_2+1)}-1},$$

and the expected wait time under the fully flexible delivery system is

$$W_{\rm S}^{\rm F} = \frac{1}{2[(k_1+k_2)\mu_{\rm S}^{\rm F} - (\lambda_1+\lambda_2)]}\left(\frac{\lambda_1+\lambda_2}{(k_1+k_2)\mu_{\rm S}^{\rm F}}\right)^{\sqrt{2(k_1+k_2+1)}-1}.$$

We can now define $\rho_1 = \frac{\lambda_1}{k_1\mu_1} = \frac{2r_1\lambda_1}{k_1}$ and $\rho_2 = \frac{2r_2\lambda_2}{k_2}$. Then the load factor of the fully flexible system is $\frac{\lambda_1+\lambda_2}{(k_1+k_2)\mu_{\rm S}^{\rm F}}$, which is $\frac{k_1\rho_1+k_2\rho_2}{k_1+k_2}$. We have

$$\begin{aligned}
W_{\rm S}^{\rm D} - W_{\rm S}^{\rm F} &= \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{r_1\rho_1^{\sqrt{2(k_1+1)}-1}}{k_1(1-\rho_1)} + \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{r_2\rho_2^{\sqrt{2(k_2+1)}-1}}{k_2(1-\rho_2)} - \frac{\lambda_1 r_1+\lambda_2 r_2}{\lambda_1+\lambda_2}\frac{\left(\frac{\rho_1k_1+\rho_2k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}-1}}{k_1(1-\rho_1)+k_2(1-\rho_2)}\\
&= \frac{1}{2(\lambda_1+\lambda_2)}\frac{\rho_1^{\sqrt{2(k_1+1)}}}{(1-\rho_1)} + \frac{1}{2(\lambda_1+\lambda_2)}\frac{\rho_2^{\sqrt{2(k_2+1)}}}{(1-\rho_2)} - \frac{k_1+k_2}{2(\lambda_1+\lambda_2)}\frac{\left(\frac{\rho_1k_1+\rho_2k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}}}{k_1(1-\rho_1)+k_2(1-\rho_2)},
\end{aligned}$$

which must be larger than or equal to 0 if the following inequality holds:

$$\frac{\rho_1^{\sqrt{2(k_1+1)}}}{(1-\rho_1)} + \frac{\rho_2^{\sqrt{2(k_2+1)}}}{(1-\rho_2)} - \frac{(k_1+k_2)\left(\frac{\rho_1k_1+\rho_2k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}}}{k_1(1-\rho_1)+k_2(1-\rho_2)} \ge 0.$$

Since $f(\rho) = \frac{\rho^\beta}{1-\rho}$ is a convex function when $\rho\in(0,1)$ and $\beta\ge 2$, by applying Jensen's inequality, we have

$$\begin{aligned}
&\frac{\rho_1^{\sqrt{2(k_1+1)}}}{(1-\rho_1)} + \frac{\rho_2^{\sqrt{2(k_2+1)}}}{(1-\rho_2)} - \frac{(k_1+k_2)\left(\frac{\rho_1k_1+\rho_2k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}}}{k_1(1-\rho_1)+k_2(1-\rho_2)}\\
\ge&\frac{\rho_1^{\sqrt{2(k_1+k_2+1)}}}{(1-\rho_1)} + \frac{\rho_2^{\sqrt{2(k_1+k_2+1)}}}{(1-\rho_2)} - \frac{(k_1+k_2)\left(\frac{\rho_1k_1+\rho_2k_2}{k_1+k_2}\right)^{\sqrt{2(k_1+k_2+1)}}}{k_1(1-\rho_1)+k_2(1-\rho_2)} \ge 0.
\end{aligned}$$

Therefore, we can conclude that if the two zones are asymmetric in terms of traveling distance, the fully flexible delivery system still outperforms the dedicated delivery system.

A.2 Proof of Proposition 2

With slight abuse of notations, we denote $k := k_1 = k_2$, $\theta := \theta_1 = \theta_2$, and $\rho_{\rm B}^{\rm D} := \rho_{\rm B,1}^{\rm D} = \rho_{\rm B,2}^{\rm D}$. Since the two zones are symmetric, we use λ to represent the order arrival rate for simplicity in this proof. We denote the load factor of the fully flexible delivery system by

$$\rho_{\rm B}^{\rm F} := \frac{\mu_{\rm B}(\theta,r)}{\mu_{\rm B}(2\theta,r)}\rho_{\rm B}^{\rm D} = \frac{2\theta+6}{\theta+6}\rho_{\rm B}^{\rm D},$$

where the last equality follows Equation (10).

Based on Equation (7), (10), and (11), we have

$$W_{\rm B}^{\rm D} - W_{\rm B}^{\rm F} = \frac{1}{4\lambda} + \frac{r}{12k}\left\{\frac{\theta^2+6\theta+18}{\theta+6}\frac{(\rho_{\rm B}^{\rm D})^{\sqrt{2(k+1)}-1}}{1-\rho_{\rm B}^{\rm D}} - \frac{2\theta^2+6\theta+9}{2(\theta+3)}\frac{(\rho_{\rm B}^{\rm F})^{\sqrt{2(2k+1)}-1}}{1-\rho_{\rm B}^{\rm F}}\right\}. \tag{A4}$$

(a) To assist the analysis on Equation (A4), we define

$$\begin{aligned}
a(\theta,\rho_{\rm B}^{\rm D},k) &= \frac{\frac{\theta^2+6\theta+18}{\theta+6}(\rho_{\rm B}^{\rm D})^{\sqrt{2(k+1)}-1}\left(1-\frac{2\theta+6}{\theta+6}\rho_{\rm B}^{\rm D}\right)}{\frac{2\theta^2+6\theta+9}{2(\theta+3)}(1-\rho_{\rm B}^{\rm D})\left(\frac{2\theta+6}{\theta+6}\rho_{\rm B}^{\rm D}\right)^{\sqrt{2(2k+1)}-1}}\\
&= \left(1+\frac{27\theta+54}{2\theta^3+18\theta^2+45\theta+54}\right)\cdot\frac{(\rho_{\rm B}^{\rm D})^{\sqrt{2(k+1)}-\sqrt{2(2k+1)}}}{\left(1+\frac{\theta}{\theta+6}\right)^{\sqrt{2(2k+1)}-1}}\cdot\frac{1-\left(1+\frac{\theta}{\theta+6}\right)\rho_{\rm B}^{\rm D}}{1-\rho_{\rm B}^{\rm D}}.
\end{aligned} \tag{A5}$$

We highlight the following two properties of Equation (A5). First, Equation (A5) decreases as θ increases because $1+\frac{27\theta+54}{2\theta^3+18\theta^2+45\theta+54}$ decreases as θ increases, $\left(1+\frac{\theta}{\theta+6}\right)^{\sqrt{2(2k+1)}-1}$ increases as θ increases, $1-\left(1+\frac{\theta}{\theta+6}\right)\rho_{\mathrm{B}}^{\mathrm{D}}$ decreases as θ increases, and all the terms are positive. Second, Equation (A5) decreases as $\rho_{\mathrm{B}}^{\mathrm{D}}$ increases because $(\rho_{\mathrm{B}}^{\mathrm{D}})^{\sqrt{2(k+1)}-\sqrt{2(2k+1)}}$ decreases as $\rho_{\mathrm{B}}^{\mathrm{D}}$ increases (it can be easily shown by taking log operation) and $\frac{1-\left(1+\frac{\theta}{\theta+6}\right)\rho_{\mathrm{B}}^{\mathrm{D}}}{1-\rho_{\mathrm{B}}^{\mathrm{D}}}$ decreases as $\rho_{\mathrm{B}}^{\mathrm{D}}$ increases, and all the terms are positive (note that $0<\left(1+\frac{\theta}{\theta+6}\right)<1$).

By substituting the value of $\rho_{\mathrm{B}}^{\mathrm{D}}=(\theta+6)/(4\theta+12)$ into $a(\theta,\rho_{\mathrm{B}}^{\mathrm{D}},k)$, we obtain

$$\begin{aligned} a\left(\theta,\frac{\theta+6}{4\theta+12},k\right) &= \frac{\frac{\theta^2+6\theta+18}{\theta+6}\left(\frac{\theta+6}{4\theta+12}\right)^{\sqrt{2(k+1)}-1}\frac{1}{2}}{\frac{2\theta^2+6\theta+9}{2(\theta+3)}\left(1-\frac{\theta+6}{4\theta+12}\right)\left(\frac{1}{2}\right)^{\sqrt{2(2k+1)}-1}} \\ &= \left(1+\frac{27\theta+54}{2\theta^3+18\theta^2+45\theta+54}\right)\cdot\left(\frac{\theta+6}{4\theta+12}\right)^{\sqrt{2(k+1)}-1}\cdot\left(\frac{1}{2}\right)^{2-\sqrt{2(2k+1)}}\cdot\frac{4\theta+12}{3\theta+6}. \end{aligned} \tag{A6}$$

If $\left(\frac{\theta+6}{4\theta+12}\right)^{\sqrt{2(k+1)}-1}\cdot\left(\frac{1}{2}\right)^{2-\sqrt{2(2k+1)}}>1$, Equation (A6) will be larger than 1. Therefore, we can conclude that Equation (A5) is larger than 1 when $\rho_{\mathrm{B}}^{\mathrm{D}}\le(\theta+6)/(4\theta+12)$. Thus, Equation (A4) is larger than 0 when $\rho_{\mathrm{B}}^{\mathrm{D}}\le(\theta+6)/(4\theta+12)$. Define function $f(\theta,k)=\left(\frac{\theta+6}{4\theta+12}\right)^{\sqrt{2(k+1)}-1}\cdot\left(\frac{1}{2}\right)^{2-\sqrt{2(2k+1)}}$. One can easily verify that we must have $0.41<\frac{\theta+6}{4\theta+12}<0.50$ since $\theta\in\left(0,\frac{\pi}{2}\right]$. Note that $f(\theta,k)$ decreases as θ increases and increases as k increases. If $\theta=0$, we have $f(0,k)>1$, for all $k\ge3$. Therefore, there must exist $\theta^{\#}$, for all $0<\theta<\theta^{\#}$ and for all $k\ge3$, $f(\theta,k)>1$. Equation (A4) is larger than 0, meaning that the fully flexible delivery system leads to a shorter expected wait time compared to the dedicated delivery system when $\rho_{\mathrm{B}}^{\mathrm{D}}\le(\theta+6)/(4\theta+12)$.

(b) We have $0<\rho_{\mathrm{B}}^{\mathrm{D}}<1$ and $0<\rho_{\mathrm{B}}^{\mathrm{D}}<\rho_{\mathrm{B}}^{\mathrm{F}}=\frac{2\theta+6}{\theta+6}\rho_{\mathrm{B}}^{\mathrm{D}}<1$. When $\frac{2\theta+6}{\theta+6}\rho_{\mathrm{B}}^{\mathrm{D}}\to1$, i.e., $\rho_{\mathrm{B}}^{\mathrm{D}}\to\frac{\theta+6}{2\theta+6}$, the second term of Equation (A4) tends to infinity. This indicates that there exists a $\bar{\rho}$, such that the fully flexible delivery system leads to a longer expected customer wait time compared to the dedicated delivery system (i.e., $W_{\mathrm{B}}^{\mathrm{D}}-W_{\mathrm{B}}^{\mathrm{F}}<0$) when $\rho_{\mathrm{B}}^{\mathrm{D}}\to\bar{\rho}$. □

A.3 Proof of Proposition 3

Following Theorem 1 and Equation (17)–(19), we have

$$W_{\mathrm{S},1}^{\mathrm{P}}=\frac{r}{k_1-q_1-2r(1-p_1)\lambda_1}\left(\frac{2r(1-p_1)\lambda_1}{k_1-q_1}\right)^{\sqrt{2(k_1-q_1+1)}-1}$$

$$W_{\mathrm{S},2}^{\mathrm{P}}=\frac{r}{k_2-q_2-2r(1-p_2)\lambda_2}\left(\frac{2r(1-p_2)\lambda_2}{k_2-q_2}\right)^{\sqrt{2(k_2-q_2+1)}-1}$$

$$W_{\mathrm{S},3}^{\mathrm{P}}=\frac{r}{q_1+q_2-2r(p_1\lambda_1+p_2\lambda_2)}\left(\frac{2r(p_1\lambda_1+p_2\lambda_2)}{q_1+q_2}\right)^{\sqrt{2(q_1+q_2+1)}-1}$$

$$W_{\mathrm{S}}^{\mathrm{P}}=\frac{(1-p_1)\lambda_1}{\lambda_1+\lambda_2}W_{\mathrm{S},1}^{\mathrm{P}}+\frac{(1-p_2)\lambda_2}{\lambda_1+\lambda_2}W_{\mathrm{S},2}^{\mathrm{P}}+\frac{p_1\lambda_1+p_2\lambda_2}{\lambda_1+\lambda_2}W_{\mathrm{S},3}^{\mathrm{P}},$$

where we use $W_{\mathrm{S},3}^{\mathrm{P}}$ to denote the expected wait time of the customers served by these flexible couriers. Actually, the partially flexible delivery system comprises three queues: two dedicated to serving their respective zones and one formed by the flexible couriers.

We define $\rho_{\mathrm{S},1}^{\mathrm{P}}=\frac{2r(1-p_1)\lambda_1}{k_1-q_1}$, $\rho_{\mathrm{S},2}^{\mathrm{P}}=\frac{2r(1-p_2)\lambda_2}{k_2-q_2}$, and $\rho_{\mathrm{S},3}^{\mathrm{P}}=\frac{2r(p_1\lambda_1+p_2\lambda_2)}{q_1+q_2}$. Then, we have

$$W_{\mathrm{S}}^{\mathrm{P}}=\frac{1}{2(\lambda_1+\lambda_2)}\left\{\frac{(\rho_{\mathrm{S},1}^{\mathrm{P}})^{\sqrt{2(k_1-q_1+1)}}}{1-\rho_{\mathrm{S},1}^{\mathrm{P}}}+\frac{(\rho_{\mathrm{S},2}^{\mathrm{P}})^{\sqrt{2(k_2-q_2+1)}}}{1-\rho_{\mathrm{S},2}^{\mathrm{P}}}+\frac{(\rho_{\mathrm{S},3}^{\mathrm{P}})^{\sqrt{2(q_1+q_2+1)}}}{1-\rho_{\mathrm{S},3}^{\mathrm{P}}}\right\}.$$

We further define $\rho_{\mathrm{S}}^{\mathrm{F}} = \frac{2r(\lambda_1+\lambda_2)}{k_1+k_2}$, which equals $\frac{k_1-q_1}{k_1+k_2}\rho_{\mathrm{S},1}^{\mathrm{P}} + \frac{k_2-q_2}{k_1+k_2}\rho_{\mathrm{S},2}^{\mathrm{P}} + \frac{q_1+q_2}{k_1+k_2}\rho_{\mathrm{S},3}^{\mathrm{P}}$. The difference in the expected wait time between the partially flexible delivery system and the fully flexible delivery system can be reformulated as

$$W_{\mathrm{S}}^{\mathrm{P}} - W_{\mathrm{S}}^{\mathrm{F}} = \frac{1}{2(\lambda_1+\lambda_2)}\left\{\frac{(\rho_{\mathrm{S},1}^{\mathrm{P}})^{\sqrt{2(k_1-q_1+1)}}}{1-\rho_{\mathrm{S},1}^{\mathrm{P}}} + \frac{(\rho_{\mathrm{S},2}^{\mathrm{P}})^{\sqrt{2(k_2-q_2+1)}}}{1-\rho_{\mathrm{S},2}^{\mathrm{P}}} + \frac{(\rho_{\mathrm{S},3}^{\mathrm{P}})^{\sqrt{2(q_1+q_2+1)}}}{1-\rho_{\mathrm{S},3}^{\mathrm{P}}} - \frac{(\rho_{\mathrm{S}}^{\mathrm{F}})^{\sqrt{2(k_1+k_2+1)}}}{1-\rho_{\mathrm{S}}^{\mathrm{F}}}\right\}. \tag{A8}$$

It is easy to prove that $f(x) = \frac{x^{\sqrt{2a}}}{1-x}$ is a convex function when $0 < x < 1$ and $a > 0$. According to Jensen's inequality, we have

$$f(\rho_{\mathrm{S}}^{\mathrm{F}}) \leq \frac{k_1-q_1}{k_1+k_2} f(\rho_{\mathrm{S},1}^{\mathrm{P}}) + \frac{k_2-q_2}{k_1+k_2} f(\rho_{\mathrm{S},2}^{\mathrm{P}}) + \frac{q_1+q_2}{k_1+k_2} f(\rho_{\mathrm{S},3}^{\mathrm{P}}),$$

which is

$$\frac{(\rho_{\mathrm{S}}^{\mathrm{F}})^{\sqrt{2(k_1+k_2+1)}}}{1-\rho_{\mathrm{S}}^{\mathrm{F}}} \leq \frac{k_1-q_1}{k_1+k_2}\frac{(\rho_{\mathrm{S},1}^{\mathrm{P}})^{\sqrt{2(k_1+k_2+1)}}}{1-\rho_{\mathrm{S},1}^{\mathrm{P}}} + \frac{k_2-q_2}{k_1+k_2}\frac{(\rho_{\mathrm{S},2}^{\mathrm{P}})^{\sqrt{2(k_1+k_2+1)}}}{1-\rho_{\mathrm{S},2}^{\mathrm{P}}} + \frac{q_1+q_2}{k_1+k_2}\frac{(\rho_{\mathrm{S},3}^{\mathrm{P}})^{\sqrt{2(k_1+k_2+1)}}}{1-\rho_{\mathrm{S},3}^{\mathrm{P}}}.$$

Because $0 < \rho_{\mathrm{S},1}^{\mathrm{P}}, \rho_{\mathrm{S},2}^{\mathrm{P}}, \rho_{\mathrm{S},3}^{\mathrm{P}} < 1$, $q_1 < k_1$, and $q_2 < k_2$, we have $(\rho_{\mathrm{S},1}^{\mathrm{P}})^{\sqrt{2(k_1+k_2+1)}} \leq (\rho_{\mathrm{S},1}^{\mathrm{P}})^{\sqrt{2(k_1-q_1+1)}}$, $(\rho_{\mathrm{S},2}^{\mathrm{P}})^{\sqrt{2(k_1+k_2+1)}} \leq (\rho_{\mathrm{S},2}^{\mathrm{P}})^{\sqrt{2(k_2-q_2+1)}}$, and $(\rho_{\mathrm{S},3}^{\mathrm{P}})^{\sqrt{2(q_1+q_2+1)}} \leq (\rho_{\mathrm{S},3}^{\mathrm{P}})^{\sqrt{2(k_1+k_2+1)}}$, which indicates

$$\frac{(\rho_{\mathrm{S}}^{\mathrm{F}})^{\sqrt{2(k_1+k_2+1)}}}{1-\rho_{\mathrm{S}}^{\mathrm{F}}} \leq \frac{k_1-q_1}{k_1+k_2}\frac{(\rho_{\mathrm{S},1}^{\mathrm{P}})^{\sqrt{2(k_1-q_1+1)}}}{1-\rho_{\mathrm{S},1}^{\mathrm{P}}} + \frac{k_2-q_2}{k_1+k_2}\frac{(\rho_{\mathrm{S},2}^{\mathrm{P}})^{\sqrt{2(k_2-q_2+1)}}}{1-\rho_{\mathrm{S},2}^{\mathrm{P}}} + \frac{q_1+q_2}{k_1+k_2}\frac{(\rho_{\mathrm{S},3}^{\mathrm{P}})^{\sqrt{2(q_1+q_2+1)}}}{1-\rho_{\mathrm{S},3}^{\mathrm{P}}}.$$

Therefore, we must have Equation (A8) ≥ 0, which means $W_{\mathrm{S}}^{\mathrm{P}} - W_{\mathrm{S}}^{\mathrm{F}} \geq 0$ always holds. □

A.4 Proof of Proposition 4

(a) Suppose that one courier delivers two orders in a batch, $k_1 = k_2 = k$, and $\theta_1 \neq \theta_2$. Under the half-flexible delivery system, we have $q_1 = q_2 = k/2$. Plugging the values of $\mu_{\mathrm{B}}(\theta_1, r)$, $\mu_{\mathrm{B}}(\theta_2, r)$, and $\mu_{\mathrm{B}}(\theta_1+\theta_2, r)$, we define the load factors of three types of delivery systems as follows:

$$\rho_{\mathrm{B},1}^{\mathrm{P}} = \frac{(1-p_1)\lambda r\theta_1(\theta_1+6)}{3k},\ \rho_{\mathrm{B},2}^{\mathrm{P}} = \frac{(1-p_2)\lambda r\theta_2(\theta_2+6)}{3k},\ \rho_{\mathrm{B},3}^{\mathrm{P}} = \frac{\lambda r(p_1\theta_1+p_2\theta_2)(\theta_1+\theta_2+6)}{6k},$$

$$\rho_{\mathrm{B},1}^{\mathrm{D}} = \frac{\lambda r\theta_1(\theta_1+6)}{6k},\ \rho_{\mathrm{B},2}^{\mathrm{D}} = \frac{\lambda r\theta_2(\theta_2+6)}{6k},$$

$$\rho_{\mathrm{B}}^{\mathrm{F}} = \frac{\lambda r(\theta_1+\theta_2)(\theta_1+\theta_2+6)}{12k}.$$

If $\rho_{\mathrm{B},i}^{\mathrm{P}} < \rho_{\mathrm{B},1}^{\mathrm{D}}$ and $\rho_{\mathrm{B},i}^{\mathrm{P}} < \rho_{\mathrm{B}}^{\mathrm{F}}$, for all $i \in \{1,2,3\}$, the half-flexible delivery system has an advantage among the three types of delivery systems under heavy traffic. To achieve $\rho_{\mathrm{B},i}^{\mathrm{P}} < \rho_{\mathrm{B},1}^{\mathrm{D}}$ and $\rho_{\mathrm{B},i}^{\mathrm{P}} < \rho_{\mathrm{B}}^{\mathrm{F}}$, we must satisfy the following conditions:

$$p_1 > \frac{1}{2} \tag{A9a}$$

$$p_2 > 1 - \frac{\theta_1(\theta_1+6)}{2\theta_2(\theta_2+6)} \tag{A9b}$$

$$p_1\theta_1 + p_2\theta_2 < \frac{\theta_1(\theta_1+6)}{\theta_1+\theta_2+6} \tag{A9c}$$

$$p_1 > 1 - \frac{(\theta_1+\theta_2)(\theta_1+\theta_2+6)}{4\theta_1(\theta_1+6)} \tag{A9d}$$

$$p_2 > 1 - \frac{(\theta_1+\theta_2)(\theta_1+\theta_2+6)}{4\theta_2(\theta_2+6)} \tag{A9e}$$

$$p_1\theta_1 + p_2\theta_2 < \frac{\theta_1+\theta_2}{2}. \tag{A9f}$$

We show that there exists an allocation rule (i.e., a configuration of p_1 and p_2) such that inequalities (A9a)–(A9f) are satisfied, with θ_1 and θ_2 constrained to $(0,\pi/2]$. Indeed, one can verify that the configuration of $p_1 = 0.6$, $p_2 = 0.2$, $\theta_1 = 1.4$, and $\theta_2 = 0.8$ satisfies inequalities (A9a)–(A9f). In fact, if we restrict $p_1 = 0.6$ and $p_2 = 0.2$, the values of θ_1 and θ_2 satisfying inequalities (A9a)–(A9f) are shown in the blue region of Figure A1. Thus, we reach the conclusion that there exists a configuration of p_1, p_2, θ_1, and θ_2, such that $\rho^{\mathrm{P}}_{\mathrm{B},i} < \rho^{\mathrm{D}}_{\mathrm{B},1}$ and $\rho^{\mathrm{P}}_{\mathrm{B},i} < \rho^{\mathrm{F}}_{\mathrm{B}}$ for all $i \in \{1,2,3\}$.

(b) The load factors satisfy

$$\rho^{\mathrm{P}}_{\mathrm{B},1} = 2(1-p_1)\,\rho^{\mathrm{D}}_{\mathrm{B},1},\;\; \rho^{\mathrm{P}}_{\mathrm{B},2} = 2(1-p_2)\,\rho^{\mathrm{D}}_{\mathrm{B},2},\;\; \rho^{\mathrm{P}}_{\mathrm{B},3} = 2\frac{p_1\theta_1 + p_2\theta_2}{\theta_1+\theta_2}\,\rho^{\mathrm{F}}_{\mathrm{B}}.$$

Conditions (A9a)–(A9f) yield

$$2(1-p_1) < 1, \qquad 2\frac{(1-p_2)\theta_2(\theta_2+6)}{\theta_1(\theta_1+6)} < 1, \qquad 2\frac{p_1\theta_1 + p_2\theta_2}{\theta_1+\theta_2} < 1.$$

Therefore, there exists a constant $c \in (0,1)$ depending only on $(p_1, p_2, \theta_1, \theta_2)$ such that

$$\rho^{\mathrm{P}}_{\mathrm{B},i} \le c < 1, \qquad \forall i \in \{1,2,3\},$$

along any sequence of systems for which $\rho^{\mathrm{D}}_{\mathrm{B},1} \to 1$ and $\rho^{\mathrm{F}}_{\mathrm{B}} \to 1$. For example, we can set

$$c := \max\left\{2(1-p_1),\, 2\frac{(1-p_2)\theta_2(\theta_2+6)}{\theta_1(\theta_1+6)},\, 2\frac{p_1\theta_1 + p_2\theta_2}{\theta_1+\theta_2}\right\} < 1.$$

Then, we have

$$\rho^{\mathrm{P}}_{\mathrm{B},i} \le c \cdot \max\{\rho^{\mathrm{D}}_{\mathrm{B},1}, \rho^{\mathrm{F}}_{\mathrm{B}}\} < 1, \qquad i \in \{1,2,3\}.$$

Thus, all three queues in the partially flexible system remain strictly stable with a uniform margin $1-c>0$, while the dedicated and fully flexible systems reach their stability boundaries. □

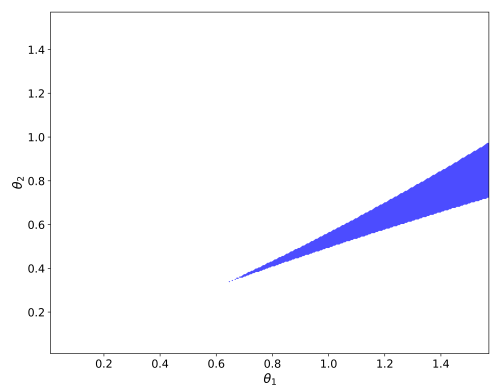


Figure A1 The values of θ_1 and θ_2 satisfying inequalities (A9a)–(A9f) given $p_1 = 0.6$ and $p_2 = 0.2$

A.5 Proof of Proposition 5

Denote by a random variable X_{T}, where the subscript abbreviates "three", the travel distance to deliver a batch of three orders. We have

$$\begin{aligned}\mathbb{E}[X_{\mathrm{T}}(\theta,r)] &= \frac{1}{\theta}\frac{1}{\theta}\frac{1}{\theta}\int_0^\theta\int_0^\theta\int_0^\theta (\max(|\theta_1-\theta_2|,|\theta_1-\theta_3|,|\theta_2-\theta_3|)r + 2r)\mathrm{d}\theta_3\mathrm{d}\theta_2\mathrm{d}\theta_1 \\ &= \frac{(\theta+4)r}{2}, \text{ and}\end{aligned} \tag{A10}$$

$$\begin{aligned}\mathbb{E}[X^2_{\mathrm{T}}(\theta,r)] &= \frac{1}{\theta}\frac{1}{\theta}\frac{1}{\theta}\int_0^\theta\int_0^\theta\int_0^\theta (\max(|\theta_1-\theta_2|,|\theta_1-\theta_3|,|\theta_2-\theta_3|)r + 2r)^2\mathrm{d}\theta_3\mathrm{d}\theta_2\mathrm{d}\theta_1 \\ &= \left(\frac{3\theta^2}{10} + 2\theta + 4\right) r^2.\end{aligned} \tag{A11}$$

The service rate and the summation of coefficients of variations can be calculated as follows:

$$\mu_{\mathrm{T}}(\theta,r)=\frac{1}{\mathbb{E}[X_{\mathrm{T}}(\theta,r)]}=\frac{2}{(\theta+4)r},\ \theta\in\left(0,\frac{\pi}{2}\right], \tag{A12}$$

$$C_{\mathrm{T}}(\theta)=\frac{1}{3}+\frac{\mathbb{E}[X_{\mathrm{T}}^2(\theta,r)]]-(\mathbb{E}[X_{\mathrm{T}}(\theta,r)])^2}{(\mathbb{E}[X_{\mathrm{T}}(\theta,r)])^2}=\frac{28\theta^2+200\theta+400}{15\theta^2+120\theta+240},\ \theta\in\left(0,\frac{\pi}{2}\right]. \tag{A13}$$

According to Theorem 2, the expected customer wait time when a courier delivers a batch of three consecutive orders is

$$w_{\mathrm{T}}(\lambda,\mu,C,k)=\frac{1}{\lambda}+\frac{C}{2(3k\mu-\lambda)}\left(\frac{\lambda}{3k\mu}\right)^{\sqrt{2(k+1)}-1}+\frac{\theta r}{9}+\frac{2r}{3}. \tag{A14}$$

Recall that λ is the per unit order arrival rate in the analysis with order batching. For the dedicated delivery system, the expected customer wait time of each zone is

$$W_{\mathrm{T},i}^{\mathrm{D}}=w_{\mathrm{T}}(\lambda\theta_i,\mu_{\mathrm{T}}(\theta_i,r),C_{\mathrm{T}}(\theta_i),k_i),\ i\in\{1,2\}, \tag{A15}$$

and the overall expected customer wait time of the dedicated delivery system is

$$W_{\mathrm{T}}^{\mathrm{D}}=\sum_{i\in\{1,2\}}\frac{\theta_i}{\theta_1+\theta_2}w_{\mathrm{T}}(\lambda\theta_i,\mu_{\mathrm{T}}(\theta_i,r),C_{\mathrm{T}}(\theta_i),k_i). \tag{A16}$$

We define the load factor of the dedicated delivery system under batch delivery as

$$\rho_{\mathrm{T},i}^{\mathrm{D}}=\frac{\lambda\theta_i}{3k_i\mu_{\mathrm{T}}(\theta_i,r)},\ i\in\{1,2\}. \tag{A17}$$

For the fully flexible delivery system, the expected customer wait time is calculated by

$$W_{\mathrm{T}}^{\mathrm{F}}=w_{\mathrm{T}}(\lambda\theta_1+\lambda\theta_2,\mu_{\mathrm{T}}(\theta_1+\theta_2,r),C_{\mathrm{T}}(\theta_1+\theta_2),k_1+k_2), \tag{A18}$$

and the load factor of the fully flexible delivery system under batch delivery is

$$\rho_{\mathrm{T}}^{\mathrm{F}}=\frac{\lambda\theta_1+\lambda\theta_2}{3(k_1+k_2)\mu_{\mathrm{T}}(\theta_1+\theta_2,r)}. \tag{A19}$$

For notational simplicity, we set $k_1=k_2=k$, $\theta_1=\theta_2=\theta$, and $\rho_{\mathrm{T},1}^{\mathrm{D}}=\rho_{\mathrm{T},2}^{\mathrm{D}}=\rho_{\mathrm{T}}^{\mathrm{D}}$. Since the two zones are symmetric, we denote λ as the order arrival rate for simplicity, with a slight abuse of notation. The load factor of the fully flexible delivery system $\rho_{\mathrm{T}}^{\mathrm{F}}$ equals $\frac{\mu_{\mathrm{T}}(\theta,r)}{\mu_{\mathrm{T}}(2\theta,r)}\rho_{\mathrm{T}}^{\mathrm{D}}$ (i.e., $\frac{2\theta+4}{\theta+4}\rho_{\mathrm{T}}^{\mathrm{D}}$ according to Equation (A12)).

Based on Equation (A15), (A16), and (A18), we have

$$\begin{aligned}W_{\mathrm{T}}^{\mathrm{D}}-W_{\mathrm{T}}^{\mathrm{F}}&=\frac{1}{2\lambda}+\frac{r}{45k}\left\{\frac{7\theta^2+50\theta+100}{\theta+4}\frac{(\rho_{\mathrm{T}}^{\mathrm{D}})^{\sqrt{2(k+1)}-1}}{1-\rho_{\mathrm{T}}^{\mathrm{D}}}-\frac{7\theta^2+25\theta+25}{\theta+2}\frac{(\rho_{\mathrm{T}}^{\mathrm{F}})^{\sqrt{2(2k+1)}-1}}{1-\rho_{\mathrm{T}}^{\mathrm{F}}}\right\}-\frac{\theta r}{9}\\&=\frac{(\theta+4)r}{12k\rho_{\mathrm{T}}^{\mathrm{D}}}+\frac{r}{45k}\left\{\frac{7\theta^2+50\theta+100}{\theta+4}\frac{(\rho_{\mathrm{T}}^{\mathrm{D}})^{\sqrt{2(k+1)}-1}}{1-\rho_{\mathrm{T}}^{\mathrm{D}}}-\frac{7\theta^2+25\theta+25}{\theta+2}\frac{(\rho_{\mathrm{T}}^{\mathrm{F}})^{\sqrt{2(2k+1)}-1}}{1-\rho_{\mathrm{T}}^{\mathrm{F}}}\right\}-\frac{\theta r}{9}.\end{aligned} \tag{A20}$$

(a) To assist the analysis on Equation (A20), we define

$$\begin{aligned}b(\theta,\rho_{\mathrm{T}}^{\mathrm{D}},k)&=\frac{\frac{7\theta^2+50\theta+100}{\theta+4}(\rho_{\mathrm{T}}^{\mathrm{D}})^{\sqrt{2(k+1)}-1}(1-\frac{2\theta+4}{\theta+4}\rho_{\mathrm{T}}^{\mathrm{D}})}{\frac{7\theta^2+25\theta+25}{\theta+2}(1-\rho_{\mathrm{T}}^{\mathrm{D}})(\frac{2\theta+4}{\theta+4}\rho_{\mathrm{T}}^{\mathrm{D}})^{\sqrt{2(2k+1)}-1}}\\&=\left(1+\frac{11\theta^2+75\theta+100}{7\theta^3+53\theta^2+125\theta+100}\right)\cdot\frac{(\rho_{\mathrm{T}}^{\mathrm{D}})^{\sqrt{2(k+1)}-\sqrt{2(2k+1)}}}{\left(1+\frac{\theta}{\theta+4}\right)^{\sqrt{2(2k+1)}-1}}\cdot\frac{1-\left(1+\frac{\theta}{\theta+4}\right)\rho_{\mathrm{T}}^{\mathrm{D}}}{1-\rho_{\mathrm{T}}^{\mathrm{D}}}.\end{aligned} \tag{A21}$$

Equation (A21) decreases as θ increases and decreases as $\rho_{\mathrm{T}}^{\mathrm{D}}$ increases.

Substituting $\rho_{\mathrm{T}}^{\mathrm{D}}=(\theta+4)/(4\theta+8)$, we have

$$b\left(\theta, \frac{\theta+4}{4\theta+8}, k\right) = \left(1 + \frac{11\theta^2+75\theta+100}{7\theta^3+53\theta^2+125\theta+100}\right) \cdot \left(\frac{\theta+4}{4\theta+8}\right)^{\sqrt{2(k+1)}-1} \cdot \left(\frac{1}{2}\right)^{2-\sqrt{2(2k+1)}} \cdot \frac{4\theta+8}{3\theta+4}. \tag{A22}$$

If $\left(\frac{\theta+4}{4\theta+8}\right)^{\sqrt{2(k+1)}-1} \cdot \left(\frac{1}{2}\right)^{2-\sqrt{2(2k+1)}} > 1$, Equation (A22) will be larger than 1. Therefore, Equation (A21) is larger than 1 when $\rho_{\mathrm{T}}^{\mathrm{D}} \le (\theta+4)/(4\theta+8)$, which indicates Equation (A20) is larger than 0 when $\rho_{\mathrm{T}}^{\mathrm{D}} \le (\theta+4)/(4\theta+8)$. For the sake of brevity, we set $f(\theta,k) = \left(\frac{\theta+4}{4\theta+8}\right)^{\sqrt{2(k+1)}-1} \cdot \left(\frac{1}{2}\right)^{2-\sqrt{2(2k+1)}}$. One can easily verify that we have $0.32 < \frac{\theta+4}{4\theta+8} < 0.50$ since $\theta \in \left(0, \frac{\pi}{2}\right]$. Note that $f(\theta,k)$ decreases as θ increases and increases as k increases. If $\theta = 0$, we have $f(0,k) > 1$, for all $k \ge 3$. Therefore, there must exist $\theta^{\#}$, for all $0 < \theta < \theta^{\#}$ and for all $k \ge 3$, $f(\theta,k) > 1$.

Then we pay attention to the remaining two terms in Equation (A20). If the remaining two terms in Equation (A20) are larger than 0, the following conditions should be satisfied:

$$9\theta + 36 - 12k\rho_{\mathrm{T}}^{\mathrm{D}}\theta > 0,$$

which means

$$k < \frac{9\theta+36}{12\rho_{\mathrm{T}}^{\mathrm{D}}\theta} < \frac{9\theta+36}{12\frac{\theta+4}{4\theta+8}\theta} = \frac{(9\theta+36)(\theta+2)}{3\theta(\theta+4)} \in [7, 72].$$

Therefore, there must exist $\bar{k}$ and $\theta^{\#}$, such that $W_{\mathrm{T}}^{\mathrm{D}} - W_{\mathrm{T}}^{\mathrm{F}} > 0$, for all $0 < \theta < \theta^{\#}$ and for all $3 \le k \le \bar{k}$.

(b) We have $0 < \rho_{\mathrm{T}}^{\mathrm{D}} < 1$ and $0 < \rho_{\mathrm{T}}^{\mathrm{D}} < \rho_{\mathrm{T}}^{\mathrm{F}} = \frac{2\theta+4}{\theta+4}\rho^{\mathrm{D}} < 1$. When $\frac{2\theta+4}{\theta+4}\rho_{\mathrm{T}}^{\mathrm{D}} \to 1$, i.e., $\rho_{\mathrm{T}}^{\mathrm{D}} \to \frac{\theta+4}{2\theta+4}$, the second term of Equation (A20) tends to infinity. This indicates that there exists a $\bar{\rho}_{\mathrm{T}}$, such that the fully flexible delivery system leads to a longer expected customer wait time compared to the dedicated delivery system (i.e., $W_{\mathrm{T}}^{\mathrm{D}} - W_{\mathrm{T}}^{\mathrm{F}} < 0$) when $\rho_{\mathrm{T}}^{\mathrm{D}} \to \bar{\rho}_{\mathrm{T}}$. □

A.6 Proof of Proposition 6

When a batch consists of three orders, we denote the load factors of the two dedicated delivery systems in the partially flexible delivery system as $\rho_{\mathrm{T},1}^{\mathrm{P}}$ and $\rho_{\mathrm{T},2}^{\mathrm{P}}$, the load factor of the flexible delivery system in the partially flexible delivery system as $\rho_{\mathrm{T},3}^{\mathrm{P}}$. We have

$$\rho_{\mathrm{T},i}^{\mathrm{P}} = \frac{(1-p_i)\lambda\theta_i}{3(k_i - q_i)\mu_{\mathrm{T}}(\theta_i)}, \; i \in \{1,2\}, \tag{A23}$$

$$\rho_{\mathrm{T},3}^{\mathrm{P}} = \frac{p_1\lambda\theta_1 + p_2\lambda\theta_2}{3(k_1+k_2)\mu_{\mathrm{T}}(\theta_1+\theta_2)}. \tag{A24}$$

(a) Suppose that one courier delivers three orders in a batch, $k_1 = k_2 = k$, and $\theta_1 \ne \theta_2$. Under the half-flexible delivery system, we have $q_1 = q_2 = k/2$. Plugging the values of $\mu_{\mathrm{T}}(\theta_1, r)$, $\mu_{\mathrm{T}}(\theta_2, r)$, and $\mu_{\mathrm{T}}(\theta_1+\theta_2, r)$, we define the load factors of three types of delivery systems as follows:

$$\rho_{\mathrm{T},1}^{\mathrm{P}} = \frac{(1-p_1)\lambda r\theta_1(\theta_1+4)}{3k}, \; \rho_{\mathrm{T},2}^{\mathrm{P}} = \frac{(1-p_2)\lambda r\theta_2(\theta_2+4)}{3k}, \; \rho_{\mathrm{T},3}^{\mathrm{P}} = \frac{\lambda r(p_1\theta_1+p_2\theta_2)(\theta_1+\theta_2+4)}{6k},$$

$$\rho_{\mathrm{T},1}^{\mathrm{D}} = \frac{\lambda r\theta_1(\theta_1+4)}{6k}, \; \rho_{\mathrm{T},2}^{\mathrm{D}} = \frac{\lambda r\theta_2(\theta_2+4)}{6k},$$

$$\rho_{\mathrm{T}}^{\mathrm{F}} = \frac{\lambda r(\theta_1+\theta_2)(\theta_1+\theta_2+4)}{12k}. \tag{A25}$$

If $\rho_{\mathrm{T},i}^{\mathrm{P}} < \rho_{\mathrm{T},1}^{\mathrm{D}}$ and $\rho_{\mathrm{T},i}^{\mathrm{P}} < \rho_{\mathrm{T}}^{\mathrm{F}}$, for all $i \in \{1,2,3\}$, the half-flexible delivery system has an advantage among the three types of delivery systems under heavy traffic. To achieve $\rho_{\mathrm{T},i}^{\mathrm{P}} < \rho_{\mathrm{T},1}^{\mathrm{D}}$ and $\rho_{\mathrm{T},i}^{\mathrm{P}} < \rho_{\mathrm{T}}^{\mathrm{F}}$, we must satisfy the following conditions:

$$p_1 > \frac{1}{2} \tag{A26a}$$

$$p_2 > 1 - \frac{\theta_1(\theta_1+4)}{2\theta_2(\theta_2+4)} \tag{A26b}$$

$$p_1\theta_1 + p_2\theta_2 < \frac{\theta_1(\theta_1+4)}{\theta_1+\theta_2+4} \tag{A26c}$$

$$p_1 > 1 - \frac{(\theta_1+\theta_2)(\theta_1+\theta_2+4)}{4\theta_1(\theta_1+4)} \tag{A26d}$$

$$p_2 > 1 - \frac{(\theta_1+\theta_2)(\theta_1+\theta_2+4)}{4\theta_2(\theta_2+4)} \tag{A26e}$$

$$p_1\theta_1 + p_2\theta_2 < \frac{\theta_1+\theta_2}{2}. \tag{A26f}$$

Similar to the proof process in online Appendix A.4, we can easily verify that the configuration of $p_1 = 0.6$, $p_2 = 0.2$, $\theta_1 = 1.4$, and $\theta_2 = 0.8$ satisfies inequalities (A26a)–(A26f) simultaneously. If we restrict $p_1 = 0.6$ and $p_2 = 0.2$, the values of θ_1 and θ_2 satisfying inequalities (A26a)–(A26f) are shown in the blue region of Figure A2. Thus, we reach the conclusion that there exists a configuration of p_1, p_2, θ_1, and θ_2, such that $\rho^{\mathrm{P}}_{\mathrm{T},i} < \rho^{\mathrm{D}}_{\mathrm{T},1}$ and $\rho^{\mathrm{P}}_{\mathrm{T},i} < \rho^{\mathrm{F}}_{\mathrm{T}}$ for all $i \in \{1,2,3\}$.

(b) The proof follows the same argument as in Proposition 4(b) in Appendix A.4 and is therefore omitted for brevity. □

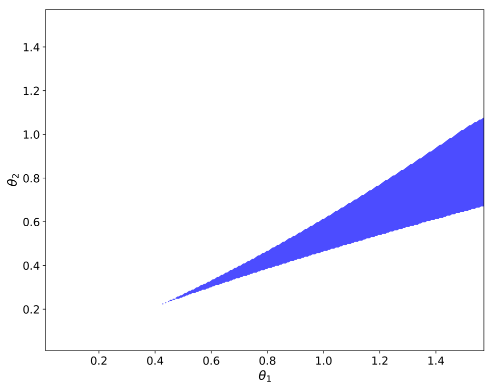


Figure A2 **The values of θ_1 and θ_2 satisfying inequalities (A26a)–(A26f) given $p_1 = 0.6$ and $p_2 = 0.2$**

A.7 Proof of Proposition 7

(a) Fix any pair $j \in [n]$, which consists of zones $2j-1$ and $2j$. Within this pair, the dedicated system yields the weighted average wait time

$$\frac{\lambda_{2j-1}}{\Lambda_j} w_{\mathrm{S}}(\lambda_{2j-1}, \mu_{\mathrm{S}}, C_{\mathrm{S}}, k_{2j-1}) + \frac{\lambda_{2j}}{\Lambda_j} w_{\mathrm{S}}(\lambda_{2j}, \mu_{\mathrm{S}}, C_{\mathrm{S}}, k_{2j}),$$

whereas the pair-pooled system yields

$$W^{\mathrm{Pair}}_{\mathrm{S},j} = w_{\mathrm{S}}(\Lambda_j, \mu_{\mathrm{S}}, C_{\mathrm{S}}, K_j).$$

By Proposition 1, pooling two zones cannot increase the expected customer wait time. Therefore,

$$\frac{\lambda_{2j-1}}{\Lambda_j} w_{\mathrm{S}}(\lambda_{2j-1}, \mu_{\mathrm{S}}, C_{\mathrm{S}}, k_{2j-1}) + \frac{\lambda_{2j}}{\Lambda_j} w_{\mathrm{S}}(\lambda_{2j}, \mu_{\mathrm{S}}, C_{\mathrm{S}}, k_{2j}) \ge W^{\mathrm{Pair}}_{\mathrm{S},j}, \qquad \forall j \in [n]. \tag{A27}$$

Multiplying both sides of (A27) by

$$\frac{\Lambda_j}{\sum_{m=1}^n \Lambda_m},$$

and summing over all $j \in [n]$, we obtain

$$\sum_{j=1}^n \frac{\Lambda_j}{\sum_{m=1}^n \Lambda_m} \left(\frac{\lambda_{2j-1}}{\Lambda_j} w_{\mathrm{S}}(\lambda_{2j-1}, \mu_{\mathrm{S}}, C_{\mathrm{S}}, k_{2j-1}) + \frac{\lambda_{2j}}{\Lambda_j} w_{\mathrm{S}}(\lambda_{2j}, \mu_{\mathrm{S}}, C_{\mathrm{S}}, k_{2j}) \right) \ge \sum_{j=1}^n \frac{\Lambda_j}{\sum_{m=1}^n \Lambda_m} W^{\mathrm{Pair}}_{\mathrm{S},j}.$$

The left-hand side is exactly $W_{\mathrm{S}}^{\mathrm{D}}$, and the right-hand side is exactly $W_{\mathrm{S}}^{\mathrm{Pair}}$. Hence,

$$W_{\mathrm{S}}^{\mathrm{D}} \geq W_{\mathrm{S}}^{\mathrm{Pair}},$$

which proves the result.

(b) Treat each pair as a super-zone. Then the pair-pooled system can be viewed as a dedicated system with n super-zones, where super-zone j has arrival rate Λ_j and K_j couriers. Correspondingly, the fully flexible system pools all n super-zones together. Applying the same argument in Proposition 1 to the n super-zones yields $W_{\mathrm{S}}^{\mathrm{F}} \leq W_{\mathrm{S}}^{\mathrm{Pair}}$. □

A.8 Proof of Proposition 8

(a) Under the symmetric multi-zone system, the pair-pooled system can be viewed as n identical adjacent pairs, each with arrival rate 2λ, $2k$ couriers, and service-region angle 2θ. Hence,

$$W_{\mathrm{B}}^{\mathrm{Pair}} = w_{\mathrm{B}}(2\lambda, \mu_{\mathrm{B}}(2\theta, r), C_{\mathrm{B}}(2\theta), 2k).$$

The fully flexible system pools all $2n$ zones together, and thus

$$W_{\mathrm{B}}^{\mathrm{F}} = w_{\mathrm{B}}(2n\lambda, \mu_{\mathrm{B}}(2n\theta, r), C_{\mathrm{B}}(2n\theta), 2nk).$$

Therefore, we have

$$\begin{aligned} W_{\mathrm{B}}^{\mathrm{Pair}} - W_{\mathrm{B}}^{\mathrm{F}} = \left(\frac{1}{4\lambda} - \frac{1}{4n\lambda}\right) &+ \frac{C_{\mathrm{B}}(2\theta)}{2\big(4k\mu_{\mathrm{B}}(2\theta, r) - 2\lambda\big)} \left(\frac{2\lambda}{4k\mu_{\mathrm{B}}(2\theta, r)}\right)^{\sqrt{2(2k+1)}-1} \\ &- \frac{C_{\mathrm{B}}(2n\theta)}{2\big(4nk\mu_{\mathrm{B}}(2n\theta, r) - 2n\lambda\big)} \left(\frac{2n\lambda}{4nk\mu_{\mathrm{B}}(2n\theta, r)}\right)^{\sqrt{2(2nk+1)}-1}. \end{aligned} \tag{A28}$$

A sufficient condition for $W_{\mathrm{B}}^{\mathrm{Pair}} - W_{\mathrm{B}}^{\mathrm{F}} \geq 0$ is

$$\frac{1}{4\lambda} - \frac{1}{4n\lambda} \geq \frac{C_{\mathrm{B}}(2n\theta)}{2\big(4nk\mu_{\mathrm{B}}(2n\theta, r) - 2n\lambda\big)} \left(\frac{2n\lambda}{4nk\mu_{\mathrm{B}}(2n\theta, r)}\right)^{\sqrt{2(2nk+1)}-1}.$$

By definition,

$$\rho_{\mathrm{B}}^{\mathrm{F}} := \frac{2n\lambda}{4nk\mu_{\mathrm{B}}(2n\theta, r)},$$

which implies

$$\lambda = \frac{2nk\mu_{\mathrm{B}}(2n\theta, r)}{n}\rho_{\mathrm{B}}^{\mathrm{F}} = \frac{6k}{r(2n\theta+6)}\rho_{\mathrm{B}}^{\mathrm{F}}.$$

Therefore,

$$\frac{1}{4\lambda} - \frac{1}{4n\lambda} = \frac{n-1}{4n\lambda} = \frac{(n-1)r(2n\theta+6)}{24nk\,\rho_{\mathrm{B}}^{\mathrm{F}}}.$$

Substituting $C_{\mathrm{B}}(2n\theta)$ and $\mu_{\mathrm{B}}(2n\theta, r)$ into the last term of Equation (A28) yields

$$\frac{r(4n^2\theta^2 + 12n\theta + 18)}{24nk(2n\theta+6)} \cdot \frac{(\rho_{\mathrm{B}}^{\mathrm{F}})^{\sqrt{2(2nk+1)}-1}}{1-\rho_{\mathrm{B}}^{\mathrm{F}}}.$$

Then, the sufficient condition for $W_{\mathrm{B}}^{\mathrm{Pair}} - W_{\mathrm{B}}^{\mathrm{F}} \geq 0$ becomes

$$\frac{(n-1)r(2n\theta+6)}{24nk\,\rho_{\mathrm{B}}^{\mathrm{F}}} \geq \frac{r(4n^2\theta^2 + 12n\theta + 18)}{24nk(2n\theta+6)} \cdot \frac{(\rho_{\mathrm{B}}^{\mathrm{F}})^{\sqrt{2(2nk+1)}-1}}{1-\rho_{\mathrm{B}}^{\mathrm{F}}}.$$

By rearranging, we obtain

$$\frac{(n-1)(2n\theta+6)^2}{4n^2\theta^2 + 12n\theta + 18} \geq \frac{(\rho_{\mathrm{B}}^{\mathrm{F}})^{\sqrt{2(2nk+1)}}}{1-\rho_{\mathrm{B}}^{\mathrm{F}}}.$$

Therefore, if $\rho_{\mathrm{B}}^{\mathrm{F}} \leq \frac{1}{2}$, then

$$\frac{1}{1-\rho_{\mathrm{B}}^{\mathrm{F}}} \leq 2,$$

and it suffices to require

$$2(\rho_{\mathrm{B}}^{\mathrm{F}})^{\sqrt{2(2nk+1)}} \le \frac{(n-1)(2n\theta+6)^2}{4n^2\theta^2+12n\theta+18},$$

that is,

$$\rho_{\mathrm{B}}^{\mathrm{F}} \le \left(\frac{(n-1)(2n\theta+6)^2}{2(4n^2\theta^2+12n\theta+18)}\right)^{\frac{1}{\sqrt{2(2nk+1)}}}.$$

Combining this with the requirement $\rho_{\mathrm{B}}^{\mathrm{F}} \le \frac{1}{2}$, we conclude that if

$$\rho_{\mathrm{B}}^{\mathrm{F}} \le \min\left\{\frac{1}{2}, \left(\frac{(n-1)(2n\theta+6)^2}{2(4n^2\theta^2+12n\theta+18)}\right)^{\frac{1}{\sqrt{2(2nk+1)}}}\right\},$$

then

$$W_{\mathrm{B}}^{\mathrm{Pair}} - W_{\mathrm{B}}^{\mathrm{F}} \ge 0.$$

This completes the proof.

(b) We have $0 < \rho_{\mathrm{B}}^{\mathrm{Pair}} < 1$ and $0 < \rho_{\mathrm{B}}^{\mathrm{Pair}} < \rho_{\mathrm{B}}^{\mathrm{F}} = \frac{n\theta+3}{\theta+3}\rho_{\mathrm{B}}^{\mathrm{Pair}} < 1$. When $\frac{n\theta+3}{\theta+3}\rho_{\mathrm{B}}^{\mathrm{Pair}} \to 1$, i.e., $\rho_{\mathrm{B}}^{\mathrm{Pair}} \to \frac{\theta+3}{n\theta+3}$, the last term of Equation (A28) tends to infinity. This indicates that there exists a $\bar{\rho}$, such that the fully flexible delivery system leads to a longer expected customer wait time compared to the dedicated delivery system (i.e., $W_{\mathrm{B}}^{\mathrm{Pair}} - W_{\mathrm{B}}^{\mathrm{F}} < 0$) when $\rho_{\mathrm{B}}^{\mathrm{Pair}} \to \bar{\rho}$. □

Appendix B Supplementary Figures and Tables

In Table B1, the dependent variable was the average customer wait time per hour, while the independent variables included order count, courier count, flexibility percentage, and a peak hour indicator. Order count and peak hour are positively associated with wait time, while courier count is negatively associated. Notably, a higher proportion of flexible couriers is also associated with longer wait times.

Table B1 Regression results for average customer wait time

Variable	Coef.	Std. Err.	t	P$>|t|$
const	8.489	2.935	2.892	0.004
order_count	0.001	0.000	2.104	0.037
courier_count	−0.005	0.002	−2.765	0.006
flexibility_percentage	9.945	3.522	2.824	0.005
peak_hour	5.559	0.902	6.161	0.000

Table B2 shows the simulation results with travel time in batch delivery. We observe that when $\rho_{\mathrm{B},1}^{\mathrm{D}}$ and $\rho_{\mathrm{B},2}^{\mathrm{D}}$ are relatively small, $W_{\mathrm{B}}^{\mathrm{F}} < W_{\mathrm{B}}^{\mathrm{D}}$. Otherwise, we have $W_{\mathrm{B}}^{\mathrm{F}} > W_{\mathrm{B}}^{\mathrm{D}}$.

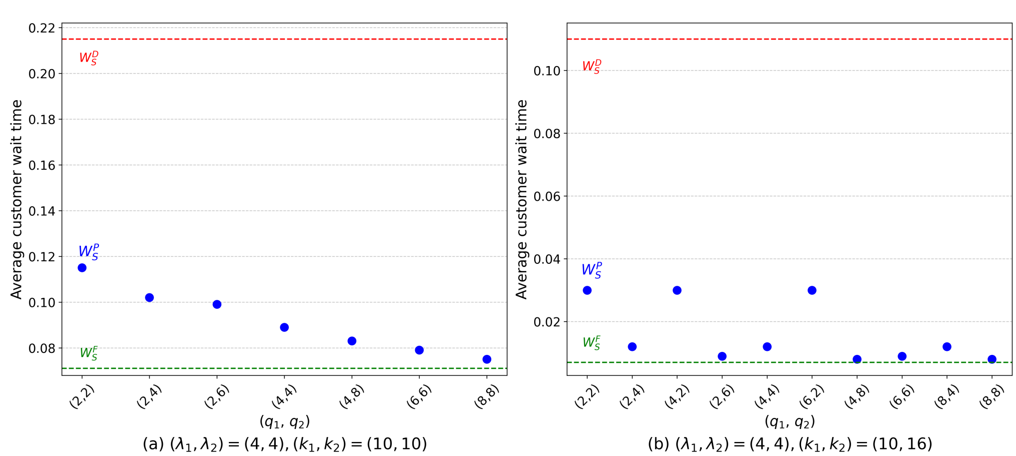


Figure B1 Simulation results of the partially flexible delivery system in single-order delivery

Table B2 **Simulation results with travel time in batch delivery**

$(\lambda_1^{\mathrm{D}}, \lambda_2^{\mathrm{D}})$	(k_1, k_2)	(θ_1, θ_2)	$(\rho_{\mathrm{B},1}^{\mathrm{D}}, \rho_{\mathrm{B},2}^{\mathrm{D}})$	With travel time		Without travel time	
				$W_{\mathrm{B}}^{\mathrm{D}}$	$W_{\mathrm{B}}^{\mathrm{F}}$	$W_{\mathrm{B}}^{\mathrm{D}}$	$W_{\mathrm{B}}^{\mathrm{F}}$
(2,2)	(10,10)	$(\frac{\pi}{4},\frac{\pi}{4})$	(0.178,0.178)	1.407	1.385	0.803	0.647
(4,4)	(10,10)	$(\frac{\pi}{4},\frac{\pi}{4})$	(0.355,0.355)	1.278	1.328	0.654	0.575
(6,6)	(10,10)	$(\frac{\pi}{4},\frac{\pi}{4})$	(0.533,0.533)	1.231	1.299	0.604	0.552
(2,2)	(10,10)	$(\frac{\pi}{2},\frac{\pi}{2})$	(0.396,0.396)	1.401	1.541	0.654	0.575
(3,3)	(10,10)	$(\frac{\pi}{2},\frac{\pi}{2})$	(0.595,0.595)	1.359	1.509	0.608	0.554
(4,4)	(10,10)	$(\frac{\pi}{2},\frac{\pi}{2})$	(0.793,0.793)	1.406	1.685	0.665	0.828
(2,6)	(10,15)	$(\frac{2\pi}{5},\frac{\pi}{2})$	(0.304,0.793)	1.299	1.362	0.612	0.541
(5,4)	(10,10)	$(\frac{2\pi}{5},\frac{\pi}{2})$	(0.760,0.793)	1.330	1.536	0.650	0.712
(8,7)	(16,15)	$(\frac{2\pi}{5},\frac{\pi}{2})$	(0.760,0.925)	1.391	1.768	0.718	1.069

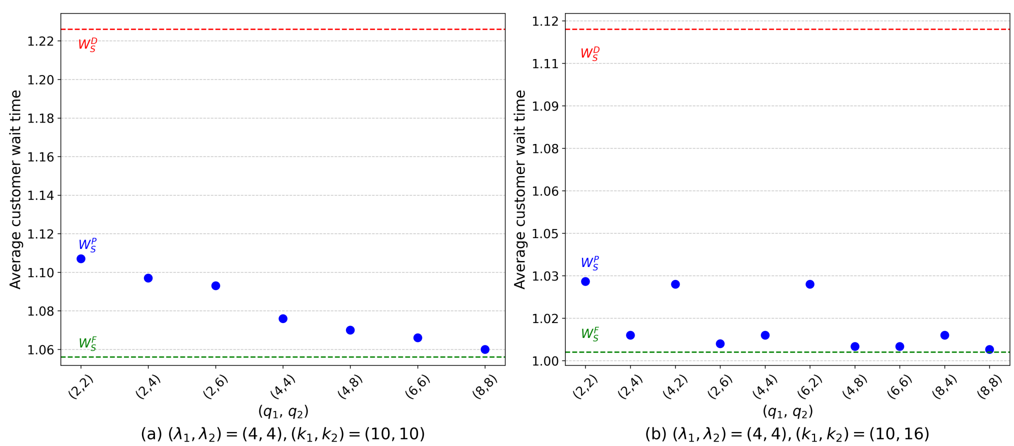


Figure B2 **Simulation results of the partially flexible delivery system in single-order delivery with travel time**

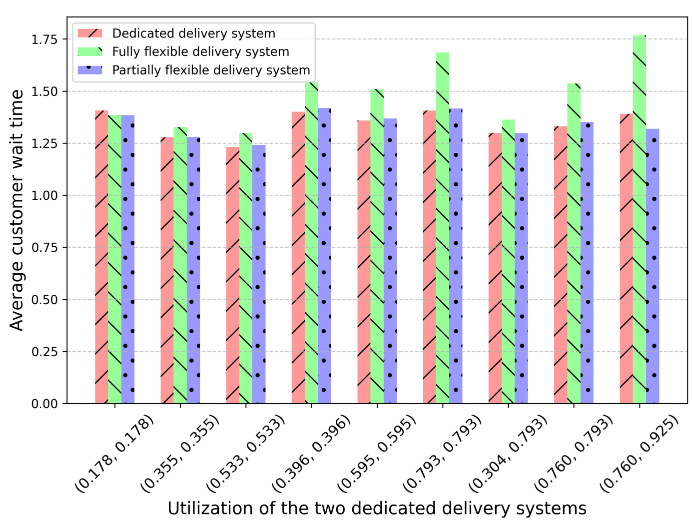


Figure B3 **Simulation results of the partially flexible delivery system in batch delivery with travel time**

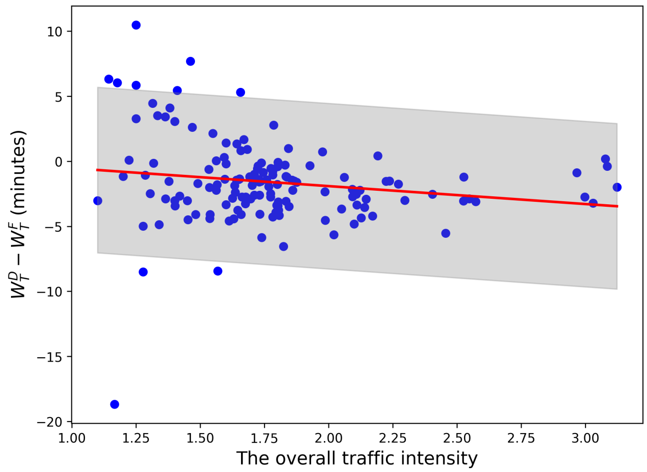


Figure B4 Wait time difference between dedicated and fully flexible delivery systems in a batch of three orders

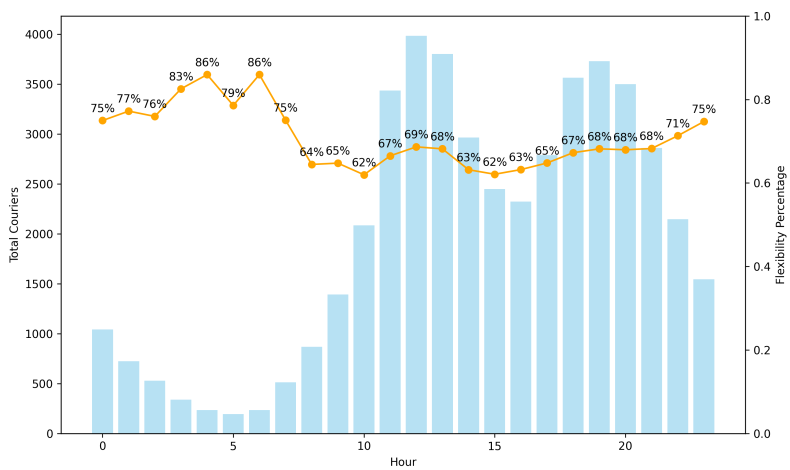


Figure B5 Flexibility distribution between hours

Appendix C Derivations of the Disk-Shaped Region

With a slight abuse of notation, we use the same symbols as in the main text, but in this section, all parameters and variables are defined within a disk-shaped region. Following the proof logic of Proposition 2, we can set $\theta_1 = \theta_2 = \pi/4$ and $\theta_1 = \theta_2 = \pi/2$, respectively, to examine the difference between $W_{\mathrm{B}}^{\mathrm{D}}$ and $W_{\mathrm{B}}^{\mathrm{F}}$ when each courier delivers two consecutive orders under the disk-shaped region with varying sizes. We still assume that $k_1 = k_2 = k$.

When $\theta_1 = \theta_2 = \pi/4$, we have $\mathbb{E}[Y_{\mathrm{B}}(\frac{\pi}{4}, r)] = 1.742r$, $\mathbb{E}[Y_{\mathrm{B}}^2(\frac{\pi}{4}, r)] = 3.180r^2$, $\mathbb{E}[Y_{\mathrm{B}}(\frac{\pi}{2}, r)] = 1.933r$, and $\mathbb{E}[Y_{\mathrm{B}}^2(\frac{\pi}{2}, r)] = 3.976r^2$. Thus, $C_{\mathrm{B}}\left(\frac{\pi}{4}\right) = 0.548$ and $C_{\mathrm{B}}\left(\frac{\pi}{2}\right) = 0.564$. Plugging the values into Equation (13) and Equation (15), we obtain

$$W_{\mathrm{B}}^{\mathrm{D}} - W_{\mathrm{B}}^{\mathrm{F}} = \frac{1}{4\lambda} + \frac{0.477r}{2k - 1.742\lambda r}\left(\frac{1.742\lambda r}{2k}\right)^{\sqrt{2(k+1)}-1} - \frac{0.273r}{2k - 1.933\lambda r}\left(\frac{1.933\lambda r}{2k}\right)^{\sqrt{2(2k+1)}-1}.$$

We define $\rho_{\mathrm{B}}^{\mathrm{D}} = \frac{1.742\lambda r}{2k}$ and assume $r = 1$, the above formula can be further simplified as

$$W_{\mathrm{B}}^{\mathrm{D}} - W_{\mathrm{B}}^{\mathrm{F}} = \frac{1}{k}\left\{\frac{1.742}{8\rho_{\mathrm{B}}^{\mathrm{D}}} + \frac{0.477}{2 - 2\rho_{\mathrm{B}}^{\mathrm{D}}}(\rho_{\mathrm{B}}^{\mathrm{D}})^{\sqrt{2(k+1)}-1} - \frac{0.273}{2 - 2.219\rho_{\mathrm{B}}^{\mathrm{D}}}(1.110\rho_{\mathrm{B}}^{\mathrm{D}})^{\sqrt{2(2k+1)}-1}\right\}.$$

Similarly, when $\theta_1 = \theta_2 = \pi/2$, we have $\mathbb{E}[Y_{\mathrm{B}}(\pi, r)] = 2.238r$ and $\mathbb{E}[Y_{\mathrm{B}}^2(\pi, r)] = 5.428r^2$. Therefore, $C_{\mathrm{B}}(\pi) = 0.583$. We finally have

$$W_{\mathrm{B}}^{\mathrm{D}} - W_{\mathrm{B}}^{\mathrm{F}} = \frac{1}{k}\left\{\frac{1.933}{8\rho_{\mathrm{B}}^{\mathrm{D}}} + \frac{0.545}{2-2\rho_{\mathrm{B}}^{\mathrm{D}}}(\rho_{\mathrm{B}}^{\mathrm{D}})^{\sqrt{2(k+1)}-1} - \frac{0.326}{2-2.316\rho_{\mathrm{B}}^{\mathrm{D}}}(1.158\rho_{\mathrm{B}}^{\mathrm{D}})^{\sqrt{2(2k+1)}-1}\right\}.$$

Figure C1 shows the results of $W_{\mathrm{B}}^{\mathrm{D}} - W_{\mathrm{B}}^{\mathrm{F}}$ when $\theta = \pi/4$ and $\theta = \pi/2$. The results remain consistent with Proposition 2: when the delivery systems are busy, the dedicated delivery system achieves shorter expected customer wait times; whereas, the fully flexible delivery system demonstrates greater advantages. The threshold that determines the sign of $W_{\mathrm{B}}^{\mathrm{D}} - W_{\mathrm{B}}^{\mathrm{F}}$ is affected by θ and is insensitive to k.

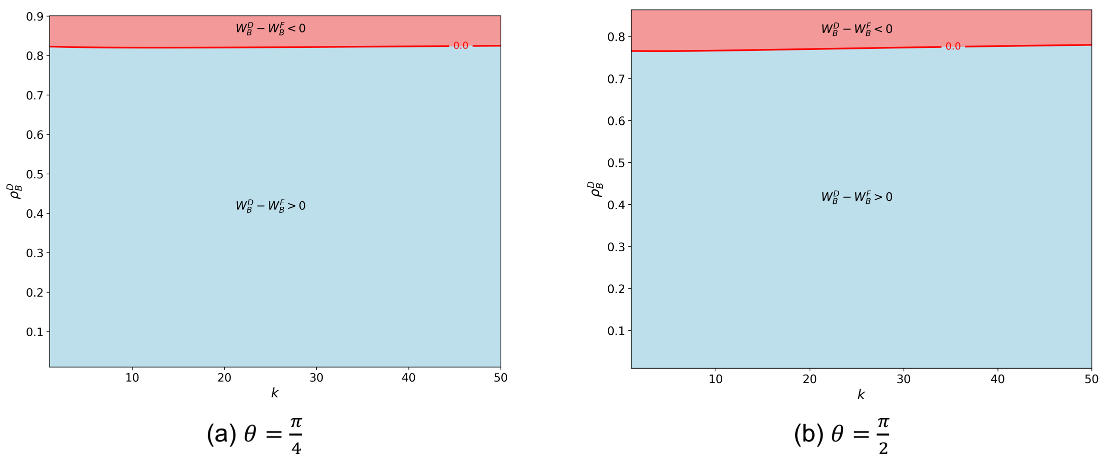


Figure C1 **Results of $W_{\mathrm{B}}^{\mathrm{D}} - W_{\mathrm{B}}^{\mathrm{F}}$ in the disk-shaped region**

For Proposition 4, we can also find configurations of θ_1 and θ_2 such that $\rho_{\mathrm{B},i}^{\mathrm{P}} < \rho_{\mathrm{B},1}^{\mathrm{D}}$ and $\rho_{\mathrm{B},i}^{\mathrm{P}} < \rho_{\mathrm{B}}^{\mathrm{F}}$, $\forall i \in \{1,2,3\}$ under the half-flexible delivery system in the disk-shaped region. For example, if $\theta_1 = \theta_2 = \pi/4$, the feasible conditions for these inequalities to hold are $p_1 > 0.445$, $p_2 > 0.445$, and $p_1 + p_2 < 1$, which are feasible. Therefore, when $\rho_{\mathrm{B},1}^{\mathrm{D}} \to 1$ and $\rho_{\mathrm{B},1}^{\mathrm{F}} \to 1$, only the half-flexible delivery system leads to a finite expected wait time.

Appendix D Analysis and Case Study of Within-Zone Batching

Suppose that the dedicated and flexible delivery systems can optimize their respective batching region sizes, denoted by ϕ^{D} and ϕ^{F}, in order to minimize customer wait time. That is, we have

$$\phi^{\mathrm{D}*} = \underset{\phi^{\mathrm{D}} \in (0,\frac{\pi}{2}]}{\arg\min}\; w_{\mathrm{B}}\left(\lambda\phi^{\mathrm{D}}, \mu_{\mathrm{B}}(\phi^{\mathrm{D}}, r), C_{\mathrm{B}}(\phi^{\mathrm{D}}), k\right), \text{ and } \phi^{\mathrm{F}*} = \underset{\phi^{\mathrm{F}} \in (0,\frac{\pi}{2}]}{\arg\min}\; w_{\mathrm{B}}\left(\lambda\phi^{\mathrm{F}}, \mu_{\mathrm{B}}(\phi^{\mathrm{F}}, r), C_{\mathrm{B}}(\phi^{\mathrm{F}}), 2k\right).$$

The optimal values of $\phi^{\mathrm{D}*}$ and $\phi^{\mathrm{F}*}$ are determined by grid search. The corresponding load factors are denoted by ρ_{D}^* and ρ_{F}^*, respectively.

We report $\frac{\phi^{\mathrm{F}*}}{\phi^{\mathrm{D}*}}$, $\frac{w_{\mathrm{B}}\left(\lambda\phi^{\mathrm{F}*}, \mu_{\mathrm{B}}(\phi^{\mathrm{F}*}, r), C_{\mathrm{B}}(\phi^{\mathrm{F}*}), 2k\right)}{w_{\mathrm{B}}\left(\lambda\phi^{\mathrm{D}*}, \mu_{\mathrm{B}}(\phi^{\mathrm{D}*}, r), C_{\mathrm{B}}(\phi^{\mathrm{D}*}), k\right)}$, ρ_{F}^*, and ρ_{D}^* in Figure D1. We find that $\phi^{\mathrm{F}*}$ is approximately twice as large as $\phi^{\mathrm{D}*}$. This finding implies that batching decisions informed by zone size are reasonable; nevertheless, restricting batching to the same zone does not yield the optimal outcome. Consistent with our theoretical expectations, under the optimal $\phi^{\mathrm{D}*}$ and $\phi^{\mathrm{F}*}$ settings, full flexibility consistently yields shorter wait times. That is, we always have $\frac{w_{\mathrm{B}}\left(\lambda\phi^{\mathrm{F}*}, \mu_{\mathrm{B}}(\phi^{\mathrm{F}*}, r), C_{\mathrm{B}}(\phi^{\mathrm{F}*}), 2k\right)}{w_{\mathrm{B}}\left(\lambda\phi^{\mathrm{D}*}, \mu_{\mathrm{B}}(\phi^{\mathrm{D}*}, r), C_{\mathrm{B}}(\phi^{\mathrm{D}*}), k\right)} < 1$. We observe that as k increases, the ratio of wait times between the full flexibility and dedicated strategies approaches one. As k increases, the system's service capacity improves, reducing the likelihood of queue buildup under both the full flexibility and dedicated strategies. In such cases, wait times are primarily determined by the service rate rather than the flexibility policy. Consequently, the performance difference between the two strategies diminishes, and their wait time ratio would converge to one.

However, in practice, the order arrival rate λ is inherently variable due to predictable and unpredictable uncertainties. Next, we examine how variations in the arrival rate impact the system when operating at a

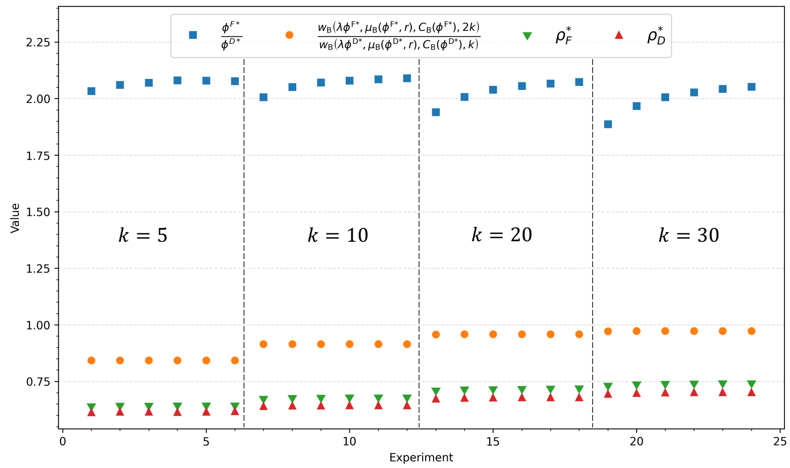


Figure D1 **Numerical results for location-based batching**

predetermined optimal batch size. Results are presented in Figure D2, which reveals that the dedicated strategy can outperform the full flexibility strategy when $\tilde{\lambda}$ deviates from its nominal value under a predetermined batch zone size. This observation highlights that variability in the arrival rate could reverse the relationship between the fully flexible and dedicated designs. In particular, when λ is large, a dedicated system leads to short wait times compared to the fully flexible system, which is consistent with our previous discussion. Therefore, while spatial batching is effective in mitigating the prolonged travel distance under full flexibility, the optimal batch zone size is susceptible to variable arrival rate, which could undermine the efficiency of the fully flexible policy.

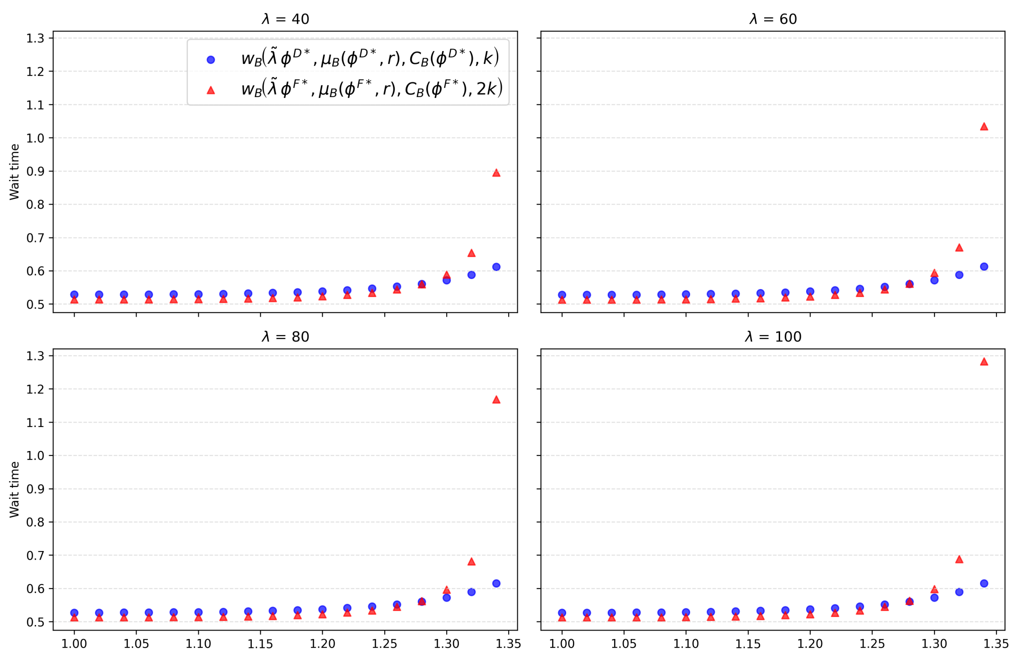
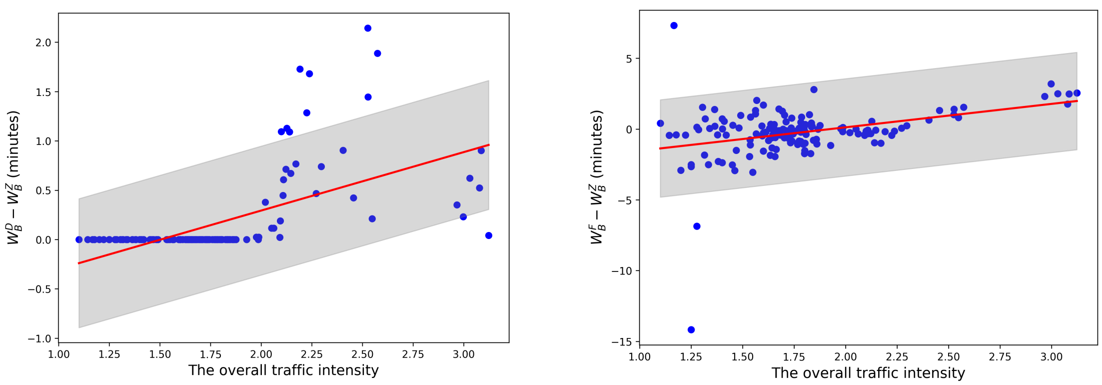


Figure D2 **Wait time dynamics with respect to variable λ**

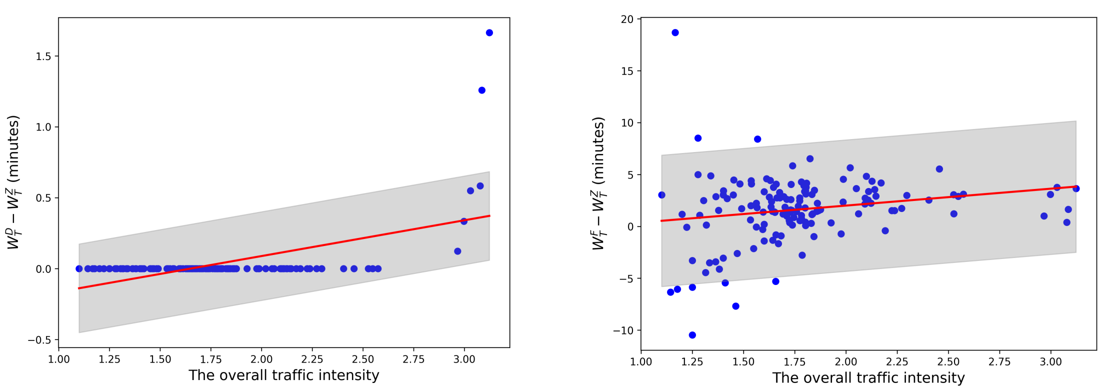
To facilitate the presentation, we use $W_{\mathrm{B}}^{\mathrm{Z}}$ to represent the average customer wait time under the batch-within-zone flexibility setting. Figure D3 reports the comparative performance of different flexibility designs using Meituan's dataset. Figure D3(a) shows that within-zone batch flexibility consistently outperforms the dedicated system, and the advantage becomes more pronounced as the overall traffic intensity increases. Figure D3(b) compares within-zone batch flexibility with full flexibility. The results reveal that there exists a positive relationship between the overall traffic intensity and $W_{\mathrm{B}}^{\mathrm{F}} - W_{\mathrm{B}}^{\mathrm{Z}}$, which indicates that full flexibility is superior under low traffic intensity, where order accumulation time is the dominant bottleneck. However, under high traffic intensity, within-zone batch flexibility achieves shorter wait times, as improving the effective service rate becomes the key driver of performance. We also extend the case study to the setting with a batch size of three orders, and the findings remain unchanged (see Figure D4).

Next, we briefly discuss the case in which couriers can strategically decide whether to wait to form a same-zone batch. Suppose a courier currently holds one order. The courier chooses between (i) departing immediately with the current order, or (ii) waiting for the next order from the same zone and then serving the two orders as a batch. Orders arrive according to a Poisson process with rate λ. Assume the courier earns revenue R for serving one order, and incurs a time cost c per unit time. If the courier departs immediately, the expected payoff is $R - c \times 2r$, which we assume to be non-negative. If the courier waits for the next same-zone order, the expected payoff is $2R - c\left(\frac{1}{\lambda} + 2r + \frac{\theta r}{3}\right)$. Then, the optimal decision is to wait to form a batch if $\lambda > \frac{3c}{3R - \theta c r}$. This implies that, from the courier's perspective, batching naturally arises when the arrival rate is high, which also aligns with the setting for the result in Proposition 2(b). Therefore, we expect the key insight to hold in the presence of strategic behaviors.



(a) $W_B^D - W_B^Z$ (b) $W_B^F - W_B^Z$

Figure D3 Customer wait time differences under different flexibility designs



(a) $W_T^D - W_T^Z$ (b) $W_T^F - W_T^Z$

Figure D4 Customer wait time differences under different flexibility designs with a batch size of three