Appendix A: Analysis of the Fluid Model

A.1. Proof of Theorem 1

We first provide the detailed background information about the theorem which will be used to prove the existence and uniqueness of the solution of the fluid model.

A.1.1. Theorem 3 in §10 of Filippov (1988)

Here, we translate the conditions in part 3 of §10 of Filippov (1988) for applying the theorem to our problem. Suppose that G is an n-dimensional domain, either open or closed, in $\mathbb{R}^n$.

1. Let $s(x)$, $x \in G \subseteq \mathbb{R}^n$, be a continuously differentiable function, and $S = \{x : s(x) = 0\}$ be a smooth surface that separates the domain G into $G^- = \{x : s(x) < 0\}$ and $G^+ = \{x : s(x) > 0\}$. Furthermore, the gradient $\nabla s(x) \neq 0$ on S.
2. Let $u(t,x)$ be a function on $\mathbb{R} \times G$ that is continuous up to the boundary of G^- and G^+ but is discontinuous on S. Let $u^-(t,x)$ and $u^+(t,x)$ be the limiting values of $u(t,x)$, in approaching $x \in S$ from domains G^- and G^+, respectively. Let $U(t,x)$ be an interval with the end points $u^-(t,x)$ and $u^+(t,x)$. Furthermore, $\frac{\partial u(t,x)}{\partial x_i}$, $i = 1, \cdots, n$, are continuous up to the boundary of G^- and G^+.
3. Let $f(t,x,u)$ be a continuous function from $\mathbb{R} \times G \times \mathbb{R}$ to $\mathbb{R}^n$ with continuous $\frac{\partial f}{\partial x_i}$ $(i = 1, \cdots, n)$ and $\frac{\partial f}{\partial u}$. Denote $f^-(t,x)$ and $f^+(t,x)$ to be the limiting values of $f(t,x,u(t,x))$, in approaching $x \in S$ from domains G^- and G^+, respectively. Let f_N, f_N^-, f_N^+ be projections of the vectors f, f^-, f^+ onto the normal to S, e.g., $f_N(t,x,u) = \frac{\nabla s(x) \cdot f(t,x,u)}{|\nabla s(x)|}$.
4. If $x \in S$ and $f_N^-(t,x) \cdot f_N^+(t,x) \leq 0$, then $u(t,x) \in U(t,x)$ and $f_N(t,x,u(t,x)) = 0$.

THEOREM 3 **(Theorem 3 in §10 of Filippov (1988))**. *Suppose that the differential equation*

$$\frac{dx}{dt} = f(t,x,u(t,x)) \tag{21}$$

whose elements are described in 1-4 above satisfies the following conditions

$$S \in C^2; \quad f, \frac{\partial f}{\partial u} \in C^1; \quad u^-(t,x), u^+(t,x) \in C^1;$$
$$\frac{\partial f_N(t,x,u)}{\partial u} \neq 0 \quad \textit{for all } u \in U(t,x).$$

If for each $t \in (a,b)$ at least one of the inequalities $f_N^- > 0$ or $f_N^+ < 0$ (possibly, different inequalities for different t and x) is valid at each point $x \in S$ then for $a < t < b$ in the domain G a solution with the initial data $x(t_0) = x_0 \in G$ exists and right uniqueness holds for (21).

A.1.2. Existence and Uniqueness of the Solution of the Fluid Model

We partition the four-dimensional set defined by Equation (3) into the following three regions.

$$\begin{aligned}\mathbb{S}_I &= \{(z_1, q_1, z_2, q_2) : q_1 > 0\},\\ \mathbb{S}_{II} &= \{(z_1, q_1, z_2, q_2) : z_1 + z_2 = 1,\ q_1 = 0\},\\ \mathbb{S}_{III} &= \{(z_1, q_1, z_2, q_2) : z_1 + z_2 < 1\}.\end{aligned}$$

Note that $\mathbb{S}_{II}$ is the intersection between $\mathbb{S}_I$ and $\mathbb{S}_{III}$. Thus, a solution cannot transit between $\mathbb{S}_I$ and $\mathbb{S}_{III}$ without visiting $\mathbb{S}_{II}$.

LEMMA 1. *The right hand sides of* (4)–(7) *are locally Lipschitz continuous within each region.*

Proof. Since $\beta(t)$ and $\alpha(t)$ are constant in $\mathbb{S}_I$ and $\mathbb{S}_{III}$, they are Lipschitz continuous in $\mathbb{S}_I$ and $\mathbb{S}_{III}$, respectively. Since $z_1(t) + z_2(t) = 1$ in $\mathbb{S}_{II}$, there exists $\delta > 0$ such that the denominator in (8), which can be written as $[\lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + p\mu z_2(t)] + [\lambda_1 + \lambda_2 + \theta q_2(t)](1-p)\mu z_2(t)$, is strictly greater than δ. Thus, $\beta(t)$ and $\alpha(t)$ are locally Lipschitz continuous in $\mathbb{S}_{II}$ since their derivatives or directional derivatives with respect to $x(t)$ are locally bounded, for example,

$$\begin{aligned}\left|\frac{\partial \beta(t)}{\partial q_2(t)}\right| &\le \left|\frac{[\mu_1 z_1(t) + p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t) - \lambda_1]}{\{[\lambda_1 + \lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + \mu z_2(t)] - \lambda_1[\mu_1 z_1(t) + p\mu z_2(t)]\}^2}\right|\\ &\le \left|\frac{[\mu_1 z_1(t) + p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t) - \lambda_1]}{\delta^2}\right| < \infty.\end{aligned}$$

□

Note that Lemma 1 guarantees the existence and uniqueness of the local solution evolving within each of the three regions. Since the right hand sides of (4)–(7) are only piecewise continuous in the whole state space and $\mathbb{S}_{II}$ is a "surface" of discontinuity that separates $\mathbb{S}_I$ from $\mathbb{S}_{III}$, we need to establish that the solution has a unique way transiting between regions. Specifically, we need to examine the behavior of the system when it approaches/crosses/deviates from the surface of discontinuity and rule out the possibilities that a solution starting from a point on the surface can evolve in more than one way. Below, Lemma 2 will first narrow down the possible evolutions by analyzing the values of (4)–(7) in the both-sided neighborhood of $\mathbb{S}_{II}$. Then, Theorem 4 will invoke Theorem 3 which considers the limiting values of the right hand sides of (4)–(7) as a solution enters the surface of discontinuity from both sides to establish the existence and uniqueness of the solution.

LEMMA 2. *Any local solution that starts from a point in* $\mathbb{S}_{II}$ *will either enter* $\mathbb{S}_I$ *or stay in* $\mathbb{S}_{II} \cup \mathbb{S}_{III}$.

Proof. Suppose that $x(\tau) \in \mathbb{S}_{II}$ at some time τ, i.e., $z_1(\tau) + z_2(\tau) = 1$ and $q_1(\tau) = 0$. In this case, $\beta(\tau)$ is given by (8). If we let

$$\zeta(t) = \frac{[\lambda_1 + \lambda_2 + \theta q_2(t) - \mu_1 z_1(t) - p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t)]}{[\lambda_1 + \lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + \mu z_2(t)] - \lambda_1[\mu_1 z_1(t) + p\mu z_2(t)]},$$

$\beta(\tau) = \min\{[\zeta(\tau)]^+, 1\}$. We now discuss the solution in a small time interval after τ for different values of $\zeta(\tau)$.

(i) If $\zeta(\tau) > 1$, i.e., $\mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] < \lambda_1$, then $\beta(\tau) = 1$ and there exists $\tilde{\tau} > \tau$ such that $\mu_1 z_1(t) + \mu[1 - z_1(t)] < \lambda_1$ for all $t \in [\tau, \tilde{\tau})$ by the continuity of $z_1(t)$ in t. Hence, $\beta(t) = 1$ and $q_1'(t) = \lambda_1 - \mu_1 z_1(t) - \mu z_2(t) > 0$ for all $t \in [\tau, \tilde{\tau})$. Thus, $q_1(t) > 0$ for all $t \in [\tau, \tilde{\tau})$ and any local solution, if exists, must enter $\mathbb{S}_I$.

(ii) If $\zeta(\tau) \leq 1$, we show by the following cases (a) and (b) that there exists $\tilde{\tau} > \tau$ such that $\mu_1 z_1(t) + \mu[1 - z_1(t)] > \lambda_1$ for $t \in (\tau, \tilde{\tau})$. Then, by (5), $q_1'(t) \leq 0$ and hence $q_1(t) = 0$ for all $t \in (\tau, \tilde{\tau})$. That is, a local solution, if exists, will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ for $t \in [\tau, \tilde{\tau})$.

(a) If $\zeta(\tau) = 1$, i.e., $\mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] = \lambda_1$, then $\mu < \lambda_1 < \mu_1$, $z_1(\tau) = \frac{\lambda_1 - \mu}{\mu_1 - \mu} \in \left(0, \frac{\lambda_1}{\mu_1}\right)$, $\beta(\tau) = \alpha(\tau) = 1$ and $z_1'(\tau) = \mu z_2(\tau) > \mu(1 - \frac{\lambda_1}{\mu_1}) > 0$. Since the right hand side of (4) is continuous within $\mathbb{S}_{II}$, there exists $\delta > 0$ such that $z_1'(t) > 0$ for any $x(t) \in \mathbb{S}_{II}$ and $||x(t) - x(\tau)|| < \delta$. Furthermore, since $z_1(\tau) \in \left(0, \frac{\lambda_1}{\mu_1}\right)$ and $z_2(\tau) = 1 - z_1(\tau)$ are continuous in t, there exists $\tilde{\tau} > \tau$ such that

$$z_1(t) < \frac{\lambda_1}{\mu_1}, \quad z_2(t) > 1 - \frac{\lambda_1}{\mu_1} > 0, \quad ||x(t) - x(\tau)|| < \delta$$

for all $t \in (\tau, \tilde{\tau})$. Next, we show that $z_1'(t) > 0$ and hence $z_1(t) > z_1(\tau)$ for all $t \in (\tau, \tilde{\tau})$.

- If $x(t) \in \mathbb{S}_I$, then $\beta(t) = \alpha(t) = 1$ and $z_1'(t) = \mu z_2(t) > 0$.
- If $x(t) \in \mathbb{S}_{II}$, then $z_1'(t) > 0$ since $||x(t) - x(\tau)|| < \delta$.
- If $x(t) \in \mathbb{S}_{III}$, then $\beta(t) = \alpha(t) = 0$ and $z_1'(t) = \lambda_1 - \mu_1 z_1(t) > 0$.

Thus, $\mu_1 z_1(t) + \mu[1 - z_1(t)] > \mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] = \lambda_1$ since $z_1(t) > z_1(\tau)$ and $\mu < \mu_1$.

(b) If $\zeta(\tau) < 1$, i.e., $\mu_1 z_1(\tau) + \mu[1 - z_1(\tau)] > \lambda_1$, then $\beta(\tau) < 1$ and there exists $\tilde{\tau} > \tau$ such that $\mu_1 z_1(t) + \mu[1 - z_1(t)] > \lambda_1$ for all $t \in [\tau, \tilde{\tau})$ by the continuity of $z_1(t)$ in t.

□

Note that Lemma 2 has not shown the existence of a local solution starting from a point in $\mathbb{S}_{II}$. It only narrows down the possible evolutions of such a solution to two cases, which simplifies the proof of the existence and uniqueness in the following Theorem 4. Specifically, an explicit solution will be derived for case (i), where the uniqueness is guaranteed by the Lipschitz continuity of the ODEs in $\mathbb{S}_I$. For case (ii), although an explicit solution is almost impossible to obtain due to the non-linearity of the ODEs, we will show the existence and uniqueness simultaneously using Theorem 3.

THEOREM 4. *There exists a unique solution to the differential equations* (4)–(7).

Proof. Since the right hand sides of (4)–(7) are locally Lipschitz continuous within each region by Lemma 1, existence and uniqueness within each region follow directly by the Picard-Lindelöf

theorem (Theorem 2.2 of Teschl 2012). If a solution transits across the regions through some point $x(\tau) \in \mathbb{S}_{II}$ at some time τ, it will either enter $\mathbb{S}_I$ or stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ by Lemma 2. We now show the existence and uniqueness of the solution for $t \in (\tau, \tilde{\tau})$, where $\tilde{\tau}$ is specified in Lemma 2.

(i) If $\zeta(\tau) > 1$, a local solution, if exists, will enter $\mathbb{S}_I$ by Lemma 2. Therefore, we only need to show the existence of a local solution in $\mathbb{S}_I$ as the uniqueness is guaranteed by the Lipschitz continuity of the ODEs in $\mathbb{S}_I$. Solving the linear ODEs (4)–(7) in $\mathbb{S}_I$, we obtain the following local solution in $\mathbb{S}_I$ for $t \in (\tau, \tilde{\tau})$.

$$
\begin{aligned}
z_2(t) &= z_2(\tau)e^{-\mu(t-\tau)}, \\
z_1(t) &= 1 - z_2(t) = 1 - [1 - z_1(\tau)]e^{-\mu(t-\tau)}, \\
q_1(t) &= (\lambda_1 - \mu_1)(t-\tau) + \frac{\mu_1 - \mu}{\mu} z_2(\tau)(1 - e^{-\mu(t-\tau)}), \\
q_2(t) &= q_2(\tau)e^{-\phi\theta(t-\tau)} + (1-\phi)\left[\frac{\lambda_2}{\phi\theta}(1 - e^{-\phi\theta(t-\tau)}) + \frac{(1-p)\mu z_2(\tau)}{\phi\theta - \mu}(e^{-\mu(t-\tau)} - e^{-\phi\theta(t-\tau)})\right].
\end{aligned}
$$

(ii) If $\zeta(\tau) \le 1$, a local solution, if exists, will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ by Lemma 2. Therefore, we only need to prove that there exists a unique local solution $x(t) = \big(z_1(t), 0, z_2(t), q_2(t)\big) \in \mathbb{S}_{II} \cup \mathbb{S}_{III}$. Note that $\mu_1 z_1(t) + \mu[1 - z_1(t)] \ge \lambda_1$ and $\zeta(t) \le 1$ for all $t \in [\tau, \tilde{\tau})$ as shown in case (ii) of Lemma 2.

To apply Theorem 3, we need to relate our setting to the four elements in section A.1.1.

(a) Let $s(x) = z_1 + z_2 - 1$, $G = \{x : q_1 = 0\}$, $S = \mathbb{S}_{II} = \{x : z_1 + z_2 = 1, q_1 = 0\}$, $G^- = \mathbb{S}_{III} = \{x : z_1 + z_2 < 1, q_1 = 0\}$ and $G^+ = \{x : z_1 + z_2 > 1, q_1 = 0\}$. Then, $s(x)$ is continuously differentiable and $\nabla s(x) = (1, 0, 1, 0)^T \neq 0$.

(b) Let $u(t,x) = \beta(t)$, if $x \in S \cup G^-$, and $u(t,x) = c > \frac{\lambda_1}{\lambda_2} + 1$ if $x \in G^+$. Then, $u(t,x) = 0$ in G^-, $u(t,x) = c$ in G^+ and $u(t,x)$ is discontinuous on S. Thus, $u^-(t,x) = 0$, $u^+(t,x) = c$ and $U(t,x) = [0, c]$ for $x \in S$.

(c) Let $f(t, x, u(t,x))$ be the right hand sides of (4)–(7) with $\beta(t)$ replaced by $u(t,x)$ and $\alpha(t)$ replaced by $\frac{\lambda_1 u(t,x)}{\mu_1 z_1 + \mu z_2}$. Then, it is obvious that $\frac{\partial f}{\partial x_i}$ and $\frac{\partial f}{\partial u}$ are continuous, and

$$
\begin{aligned}
f_N\big(t, x, u(t,x)\big) = \frac{1}{\sqrt{2}}\Bigg\{ & [1 - u(t,x)](\lambda_1 + \lambda_2 + \theta q_2) + \frac{\lambda_1 u(t,x)}{\mu_1 z_1 + \mu z_2}(\mu_1 z_1 + \mu z_2) \\
& - \mu_1 z_1 - \left[p + \frac{\lambda_1 u(t,x)}{\mu_1 z_1 + \mu z_2}(1-p)\right]\mu z_2 \Bigg\}, \\
f_N^-\big(t,x\big) = \frac{1}{\sqrt{2}}(\lambda_1 + \lambda_2 + \theta q_2 - \mu_1 z_1 - p\mu z_2), & \\
f_N^+\big(t,x\big) = \frac{1}{\sqrt{2}}\left\{\lambda_1 - (c-1)\lambda_2 - (c-1)\theta q_2 - \mu_1 z_1 - \left[p + \frac{\lambda_1 c}{\mu_1 z_1 + \mu z_2}(1-p)\right]\mu z_2\right\} < 0. &
\end{aligned}
$$

(d) If $x \in S$ and $f_N^-(t,x) \cdot f_N^+(t,x) \leq 0$, then $\lambda_1 + \lambda_2 + \theta q_2 - \mu_1 z_1 - p\mu z_2 \geq 0$ and $\zeta(t) \geq 0$. Thus, $u(t,x) = \beta(t) = \zeta(t) \in [0,1] \subseteq U(t,x)$, $z_1'(t) + z_2'(t) = 0$ and $f_N(t,x,u(t,x)) = 0$.

Given that $\frac{\partial f_N(t,x,u)}{\partial u} = \frac{1}{\sqrt{2}} \left[-(\lambda_2 + \theta q_2) - \frac{\lambda_1(1-p)\mu z_2}{\mu_1 z_1 + \mu z_2} \right] < 0$ and all the conditions in Theorem 3 hold, a unique solution can be found in $G^+ \cup \mathbb{S}_{II} \cup \mathbb{S}_{III}$. Since $f_N^+(t,x) < 0$ for any $x \in \mathbb{S}_{II}$, the solution will only be in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ starting from a point in $\mathbb{S}_{II}$. Therefore, this solution is also the unique solution to (4)–(7) in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$.

Thus, local existence and uniqueness of the solution can be guaranteed. Since the right hand sides of (4)–(7) are bounded, e.g., $|z_1'(t)| = |[1-\beta(t)]\lambda_1 + \alpha(t)[\mu_1 z_1(t) + \mu z_2(t)] - \mu_1 z_1(t)| \leq \lambda_1 + \mu_1 + \mu$, the solution will not go to infinity in a finite amount of time. Therefore, the unique local solution can be extended to the whole space as $t \to \infty$ by Theorem 2.17 of Teschl (2012). □

Based on the above proof, we can summarize the evolution of the solution. At an arbitrary moment τ, the evolution of a solution *within a small amount of time after* τ can be determined as follows.

(i) If $x(\tau) \in \mathbb{S}_I$, the solution will stay in $\mathbb{S}_I$ until it reaches the boundary of $\mathbb{S}_I$, i.e., $\mathbb{S}_{II}$, at some time. A closed form expression of the solution can be obtained by solving (4)–(7) with $\beta(t) = \alpha(t) = 1$.

(ii) If $x(\tau) \in \mathbb{S}_{II}$, its local evolution can be classified into the following cases by the value of $\zeta(\tau)$ defined in Lemma 2.
   - If $\zeta(\tau) > 1$, the solution will enter $\mathbb{S}_I$.
   - If $0 < \zeta(\tau) \leq 1$, the solution will stay in $\mathbb{S}_{II}$.
   - If $\zeta(\tau) = 0$, the solution will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$. Our proof doesn't rule out the possibility that the solution transits infinite times between $\mathbb{S}_{II}$ and $\mathbb{S}_{III}$ within a finite amount of time after τ.
   - If $\zeta(\tau) < 0$, the solution will enter $\mathbb{S}_{III}$.

(iii) If $x(\tau) \in \mathbb{S}_{III}$, the solution will stay in $\mathbb{S}_{III}$ forever or reaches the boundary of $\mathbb{S}_{III}$, i.e., $\mathbb{S}_{II}$, after some time. A closed form expression of the solution can be obtained by solving (4)–(7) with $\beta(t) = \alpha(t) = 0$.

Thus, the local evolution of a solution can be determined at every moment of time according to the above cases. Extending the process in time allows us to obtain the evolution of the global solution. For example, if the initial state $x(0)$ is in $\mathbb{S}_{III}$, the solution will first stay in $\mathbb{S}_{III}$ as in (iii). Then, depending on the system parameters and the initial state, the closed form expression will tell us whether the solution will reach the boundary $\mathbb{S}_{II}$ or not. Suppose that the solution reaches $\mathbb{S}_{II}$ at some time τ. Then, the solution will evolve according to (ii) within a small amount of time after τ, e.g., $(\tau, \tilde{\tau})$, depending on how the solution reaches $\mathbb{S}_{II}$, i.e., the value of $\zeta(\tau)$. For instance, if $\zeta(\tau) > 1$, this solution will enter $\mathbb{S}_I$ immediately after τ. That is, the solution transits

from $\mathbb{S}_{III}$ to $\mathbb{S}_I$ through a point $x(\tau) \in \mathbb{S}_{II}$ without staying in $\mathbb{S}_{II}$. Following the same process, we can determine the evolution of the solution from time $\tilde{\tau}$ until we obtain the global solution as $t \to \infty$.

A.1.3. Convergence to the Steady State

Let $x(\infty)$ denote the limit given in the Theorem. It is easy to verify that $x(\infty)$ is an invariant state, i.e., $x'(t) = 0$ for all $t \geq 0$ if $x(0) = x(\infty)$. We will show that $\lim_{t\to\infty} x(t) = x(\infty)$ for any initial state $x(0)$.

We first show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ after a finite amount of time for any initial state by the following argument. Note that it is impossible for the process $x(\cdot)$ to travel directly between $\mathbb{S}_{III}$ and $\mathbb{S}_I$ without visiting $\mathbb{S}_{II}$.

i. Suppose $x(0) \in \mathbb{S}_I$. We will show that there exists a $\tau < \infty$ such that $q_1(\tau) = 0$. In other words, the solution $x(\tau) \in \mathbb{S}_{II}$ if $x(0) \in \mathbb{S}_I$. Suppose that $x(t) \in \mathbb{S}_I$ in which case $q_1(t) > 0$ and $z_1(t) + z_2(t) = 1$ for all $t \geq 0$. Then, the differential equations (4) and (6) become $z_1'(t) = \mu z_2(t) \geq 0$ and $z_2'(t) = -\mu z_2(t)$, respectively, for all t. Thus, as t increases, $z_2(t)$ decreases while $z_1(t)$ increases at the same rate and $\lim_{t\to\infty} z_1(t) = 1$. In the meantime, the differential equation (5) becomes

$$q_1'(t) = \lambda_1 - [\mu_1 z_1(t) + \mu z_2(t)].$$

Since $\lambda_1 < \mu_1$ by Definition 1, there must exist a finite time τ and $\kappa < 0$ such that $q_1'(t) < \kappa < 0$ for all $t \geq \tau$. This implies that $q_1(\cdot)$ has to hit 0 in a finite amount of time, a contradiction.

So upon returning to $\mathbb{S}_{II}$ at τ, we must have

$$\lambda_1 - [\mu_1 z_1(\tau) + \mu z_2(\tau)] < 0. \tag{22}$$

ii. Suppose $x(0) \in \mathbb{S}_{II}$. For any t such that $x(t) \in \mathbb{S}_{II}$, substituting (8) and (9) into (5), we have

$$q_1'(t) = [\lambda_1 - \mu_1 z_1(t) - \mu z_2(t)]^+. \tag{23}$$

(a) If $\lambda_1 \leq \mu$, then $\lambda_1 - \mu_1 z_1(t) - \mu z_2(t) \leq 0$ because $z_1(t) + z_2(t) = 1$ and $q_1'(t) = 0$ by (23). This implies that the process $x(\cdot)$ will never move from $\mathbb{S}_{II}$ to $\mathbb{S}_I$.

(b) If $\lambda_1 > \mu$, then by (23) $q_1'(t) = 0$ if and only if $z_1(t) \geq z_1^\dagger := \frac{\lambda_1 - \mu}{\mu_1 - \mu}$ and it is possible that the process $x(\cdot)$ will move from $\mathbb{S}_{II}$ to $\mathbb{S}_I$. However, once the process is in $\mathbb{S}_I$, it will move back to $\mathbb{S}_{II}$ in a finite amount of time, say at time τ at which $z_1(\tau) > z_1^\dagger$ by (22). Next we show that $z_1(t) > z_1^\dagger$ for $t > \tau$ so that $x(\cdot)$ will never go back to $\mathbb{S}_I$ again. Suppose there exists a finite $\tau_1 > \tau$ such that $z_1(\tau) \leq z_1^\dagger$. Then, by the mean value theorem, there must exist some $\tau_2 \in (\tau, \tau_1)$ such that $z_1^\dagger < z_1(\tau_2) < \min\{z_1(\tau), \frac{\lambda_1}{\mu_1}\}$ and $z_1'(\tau_2) < 0$. However, for $z_1(t) > z_1^\dagger$ the differential equation (4) becomes

$$z_1'(t) = \lambda_1 - \mu_1 z_1(t), \tag{24}$$

which implies $z_1'(\tau_2) = \lambda_1 - \mu_1 z_1(\tau_2) > 0$, a contradiction. So $z_1(t) > z_1^\dagger$ and $q_1'(t) = 0$ for all $t > \tau$ and $x(\cdot)$ will not go back to $\mathbb{S}_I$ again, i.e., stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$.

In summary, if $x(0) \in \mathbb{S}_{II}$, the process $x(\cdot)$ will either stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ or visit $\mathbb{S}_I$ at most once before coming back to $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ after a finite amount of time.

iii. Suppose $x(0) \in \mathbb{S}_{III}$. If $x(\cdot)$ ever leaves $\mathbb{S}_{III}$, it will first visit $\mathbb{S}_{II}$. As we discussed above, it will stay in $\mathbb{S}_{II} \cup \mathbb{S}_{III}$ after a finite amount of time.

Next, we derive the steady state of the fluid model by assuming that $x(t) \in \mathbb{S}_{II} \cup \mathbb{S}_{III}$ in which $q_1(t) = 0$ and the differential equation (4) becomes (24) or $z_1(t) = \frac{\lambda_1}{\mu_1} - [\frac{\lambda_1}{\mu_1} - z_1(0)]e^{-\mu_1 t}$. So

$$\lim_{t\to\infty} z_1(t) = \frac{\lambda_1}{\mu_1}. \tag{25}$$

To derive the steady state of $z_2(t)$ and $q_2(t)$, we need to consider the following three cases.

1. $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{p\mu} > 1$. We first show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{II}$ and then analyze the steady state of $z_2(t)$ and $q_2(t)$ in $\mathbb{S}_{II}$.

   (a) If $x(0) \in \mathbb{S}_{III}$, then there exists $\tau > 0$ such that $z_1(t) + z_2(t) < 1$ and $x(t) \in \mathbb{S}_{III}$ for all $t \in [0, \tau)$. Then the ODEs (4) and (6) become $z_1'(t) = \lambda_1 - \mu_1 z_1(t)$ and $z_2'(t) = \lambda_2 + \theta q_2(t) - p\mu z_2(t)$, respectively, for $t \in [0, \tau)$. Since $\lambda_1 + \lambda_2 > \lambda_1 + p\mu(1 - \frac{\lambda_1}{\mu_1}) = \lim_{t\to\infty}[\mu_1 z_1(t) + p\mu(1 - z_1(t))]$ by (25), there exist $\tau_1 \geq 0$ and an $\epsilon > 0$ such that

   $$\lambda_1 + \lambda_2 > \mu_1 z_1(t) + p\mu[1 - z_1(t)] + \epsilon \geq \mu_1 z_1(t) + p\mu z_2(t) + \epsilon \tag{26}$$

   for all $t \geq \tau_1$. This implies that $z_1'(t) + z_2'(t) = \lambda_1 + \lambda_2 + \theta q_2(t) - \mu_1 z_1(t) - p\mu z_2(t) > \epsilon$ for all $t \in [\tau_1, \infty)$. Hence $z_1(t) + z_2(t)$ will increase until it reaches 1 or $x(\cdot)$ moves to $\mathbb{S}_{II}$ after a finite amount of time.

   (b) If $x(0) \in \mathbb{S}_{II}$, the process $x(\cdot)$ will go back to $\mathbb{S}_{II}$ even if it moves to $\mathbb{S}_{III}$ as shown above. Thus, there exists a finite $\tau \geq 0$ such that $x(\tau) \in \mathbb{S}_{II}$. We next show that $x(\cdot)$ will then stay in $\mathbb{S}_{II}$ for $t \geq \tau$. By the analysis in i(a) and ii(b),

   $$\lambda_1 < \mu_1 z_1(t) + \mu[1 - z_1(t)] \tag{27}$$

   holds for all $t \geq 0$. Since (26) and (27) hold for $t \geq \tau$, we have

   $$\beta(t) = \frac{[\lambda_1 + \lambda_2 + \theta q_2(t) - \mu_1 z_1(t) - p\mu z_2(t)][\mu_1 z_1(t) + \mu z_2(t)]}{[\lambda_1 + \lambda_2 + \theta q_2(t)][\mu_1 z_1(t) + \mu z_2(t)] - \lambda_1[\mu_1 z_1(t) + p\mu z_2(t)]},$$
   $$\alpha(t) = \frac{\lambda_1 \beta(t)}{\mu_1 z_1(t) + \mu z_2(t)}.$$

   Substituting them into (4) and (6), we obtain $z_2'(t) = -\lambda_1 + \mu_1 z_1(t) = -z_1'(t)$. This implies that $z_1'(t) + z_2'(t) = 0$ for $t \geq \tau$ and $x(\cdot)$ stays in $\mathbb{S}_{II}$.

Substituting (8) and (9) into (7), we have

$$q_2'(t) = \frac{g(x(t))}{[\lambda_1+\lambda_2+\theta q_2(t)][\mu_1 z_1(t)+\mu z_2(t)] - \lambda_1[\mu_1 z_1(t)+p\mu z_2(t)]}, \tag{28}$$

where

$$\begin{aligned} g(x(t)) = {} & (1-\phi)[\lambda_1+\lambda_2+\theta q_2(t)-\mu_1 z_1(t)-p\mu z_2(t)][\mu_1 z_1(t)+\mu z_2(t)][\lambda_2+\theta q_2(t)] \\ & + (1-\phi)[\lambda_1+\lambda_2+\theta q_2(t)-\mu_1 z_1(t)-p\mu z_2(t)]\lambda_1(1-p)\mu z_2(t) \\ & - \theta q_2(t)[\lambda_1+\lambda_2+\theta q_2(t)][\mu_1 z_1(t)+\mu z_2(t)] + \theta q_2(t)\lambda_1[\mu_1 z_1(t)+p\mu z_2(t)]. \end{aligned}$$

For any given $z_1(t)$ and $z_2(t)$, $g(\cdot)$ is a concave quadratic function of $q_2(t)$ and positive at $q_2(t)=0$ by (26) and (27). Furthermore, since the denominator in (28) is positive, there exists a threshold $\hat{q}_2(z_1(t),z_2(t))$ such that $q_2'(t)>0$ if $q_2(t)<\hat{q}_2(z_1(t),z_2(t))$ and $q_2'(t)\le 0$ otherwise. Thus, there exist a C_i such that $\left|\frac{\partial \hat{q}_2(z_1(t),z_2(t))}{\partial z_i(t)}\right| < C_i$ for all $t\ge\tau$ where $i=1,2$.

We are now ready to construct a Lyapunov function to show the convergence. For any $t\ge\tau$, let $V(x(t)) = C_1\left|z_1(t)-\frac{\lambda_1}{\mu_1}\right| + C_2\left|z_2(t)-\left(1-\frac{\lambda_1}{\mu_1}\right)\right| + |q_2(t)-\hat{q}_2(z_1(t),z_2(t))|$, which is zero only at $x(\infty)=\left(\frac{\lambda_1}{\mu_1},0,1-\frac{\lambda_1}{\mu_1},\frac{1-\phi}{\theta\phi}\left[\lambda_2-p\mu\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\right)$ and positive elsewhere. Suppose $q_2(t)>\hat{q}_2(z_1(t),z_2(t))$, then $\frac{dV(x(t))}{dt} = -C_1|z_1'(t)| - C_2|z_2'(t)| + q_2'(t) - \frac{\partial \hat{q}_2(z_1(t),z_2(t))}{\partial z_1(t)}z_1'(t) - \frac{\partial \hat{q}_2(z_1(t),z_2(t))}{\partial z_2(t)}z_2'(t) < 0$. Similarly, we can show $\frac{dV(x(t))}{dt}\le 0$ when $q_2(t)\le\hat{q}_2(z_1(t),z_2(t))$. Thus, $V(x(t))$ is a Lyapunov function and hence $\lim_{t\to\infty} x(t) = x(\infty)$.

Substituting $x(\infty)$ into (8) and (9), we can obtain $\beta(\infty)$ and $\alpha(\infty)$ that satisfy (10)–(12). For example,

$$\beta = \frac{\lambda_2 - p\mu(1-\frac{\lambda_1}{\mu_1})}{\lambda_2 - p\mu(1-\frac{\lambda_1}{\mu_1}) + \phi\mu(1-\frac{\lambda_1}{\mu_1})\frac{\lambda_1+p\mu(1-\frac{\lambda_1}{\mu_1})}{\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})}}. \tag{29}$$

2. $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{p\mu}<1$. We first derive the steady state of $q_2(t)$, $\alpha(t)$ and $\beta(t)$, and then show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{III}$ before deriving the steady state of $z_2(t)$.

When $x(t)\in\mathbb{S}_{II}$, by (25)–(28), for any $\epsilon>0$, there exist $A>0$ and $\tau>0$ such that

$$q_2'(t) \le -Aq_2(t)+\epsilon. \tag{30}$$

When $x(t)\in\mathbb{S}_{III}$, the differential equation (7) is

$$q_2'(t) = -\theta q_2(t). \tag{31}$$

In either case, $\lim_{t\to\infty} q_2(t) = 0$.

Note that $\lambda_1+\lambda_2 < \lambda_1+p\mu(1-\frac{\lambda_1}{\mu_1}) = \lambda_1 + \lim_{t\to\infty}[\mu_1 z_1(t)+p\mu(1-z_1(t))]$ by (25). Thus, after a finite amount of time,

$$\lambda_1+\lambda_2+\theta q_2(t) < \mu_1 z_1(t)+p\mu[1-z_1(t)] \tag{32}$$

and $\beta(t)=\alpha(t)=0$ as long as $x(t)\in\mathbb{S}_{II}\cup\mathbb{S}_{III}$. In this case, $z_1'(t)+z_2'(t)=\lambda_1+\lambda_2+\theta q_2(t)-\mu_1 z_1(t)-p\mu z_2(t)$ after substituting $\beta(t)=\alpha(t)=0$ into (4) and (6).

We now show that $x(\cdot)$ will eventually stay in $\mathbb{S}_{III}$ and derive the steady state of $z_2(t)$. If $x(\tau)\in\mathbb{S}_{II}$, then there exists a small $\delta>0$ such that $x(\tau+t)\in\mathbb{S}_{III}$ for all $t\in(0,\delta]$ (i.e., $x(\cdot)$ will immediately leave $\mathbb{S}_{II}$) as $z_1'(\tau+t)+z_2'(\tau+t)<0$ for $t\in[0,\delta]$ by (32). If $x(\tau)\in\mathbb{S}_{III}$, then $x(\cdot)$ will stay in $\mathbb{S}_{III}$ because $z_1(t)+z_2(t)$ can never increase to 1 by (32). Thus, no matter whether $x(0)$ is in $\mathbb{S}_{II}$ or $\mathbb{S}_{III}$, $x(t)\in\mathbb{S}_{III}$ for t large enough. Let $V(x(t))=\left|z_1(t)-\frac{\lambda_1}{\mu_1}\right|+\left|z_2(t)-\frac{\lambda_2}{p\mu}\right|+q_2(t)$, which is zero only at $x(\infty)=\left(\frac{\lambda_1}{\mu_1},0,\frac{\lambda_2}{p\mu},0\right)$ and positive elsewhere. Then, $\frac{dV(x(t))}{dt}=-|z_1'(t)|-|z_2'(t)|-\theta q_2(t)$ for $z_2(t)\le\frac{\lambda_2}{p\mu}$ and $z_2(t)\ge\frac{\lambda_2+\theta q_2(t)}{p\mu}$, and $\frac{dV(x(t))}{dt}=-|z_1'(t)|+z_2'(t)-\theta q_2(t)=-|z_1'(t)|+\lambda_2-p\mu z_2(t)$ otherwise for $x(t)\in\mathbb{S}_{III}$. $\frac{dV(x(t))}{dt}=0$ only when $x(t)=\left(\frac{\lambda_1}{\mu_1},0,\frac{\lambda_2}{p\mu},0\right)$ and $\frac{dV(x(t))}{dt}<0$ otherwise. Thus, $V(x(t))$ is a Lyapunov function and hence $\lim_{t\to\infty} z_2(t)=\frac{\lambda_2}{p\mu}$.

3. $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{p\mu}=1$. Note that (30) and (31) also hold in this case. Thus, $\lim_{t\to\infty} q_2(t)=0$. For $z_2(t)$, note that

$$\limsup_{t\to\infty} z_2(t)\le 1-\lim_{t\to\infty} z_1(t)=1-\frac{\lambda_1}{\mu_1}=\frac{\lambda_2}{p\mu}$$

since $z_2(t)\le 1-z_1(t)$. On the other hand,

$$\begin{aligned} z_2'(t) &= \begin{cases} \lambda_2+\theta q_2(t)-p\mu z_2(t), & \text{if } z_2(t)<1-z_1(t),\\ \lambda_2+\theta q_2(t)-p\mu z_2(t), & \text{if } \lambda_1+\lambda_2+\theta q_2(t)\le\mu_1 z_1(t)+p\mu z_2(t),\\ \frac{\mu_1}{p\mu}\big(\lambda_2-p\mu z_2(t)\big), & \text{otherwise,}\end{cases}\\ &\ge \min\left\{1,\frac{\mu_1}{p\mu}\right\}\cdot\big(\lambda_2-p\mu z_2(t)\big). \end{aligned}$$

Thus, $\liminf_{t\to\infty} z_2(t)\ge\frac{\lambda_2}{p\mu}$ and $\lim_{t\to\infty} z_2(t)=1-\frac{\lambda_1}{\mu_1}$. It is obvious that $\beta=\alpha=0$.

In all the cases, the limit $TH_2=\lim_{t\to\infty} p\mu z_2(t)=p\mu z_2(\infty)$. □

A.2. Proof of Corollary 1

By Theorem 1, $\beta=0$ when the system is effectively under or critically loaded, and β has a closed-form expression

$$\frac{1}{\beta}=1+\phi\frac{\mu(1-\frac{\lambda_1}{\mu_1})\left[\lambda_1+p\mu(1-\frac{\lambda_1}{\mu_1})\right]}{\left[\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})\right]\left[\lambda_2-p\mu(1-\frac{\lambda_1}{\mu_1})\right]} \tag{33}$$

$$=1-\phi+\frac{\phi}{1-\frac{p\mu}{\lambda_2}(1-\frac{\lambda_1}{\mu_1})}\left[1+\frac{\mu_t}{\mu_2}\cdot\frac{\lambda_1}{\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})}\cdot\frac{p\mu}{\lambda_2}(1-\frac{\lambda_1}{\mu_1})\right] \tag{34}$$

when the system is effectively overloaded. When $\mu_s\le\mu_2$, it can be easily seen from (33) that $\frac{1}{\beta}$ decreases in μ_t since both μ and $p\mu$ decrease in μ_t. When $\mu_s>\mu_2$, substitute $\mu_t=\frac{\mu_s(\mu-\mu_2)}{\mu_s-\mu}$ into (33) and consider $\mu\in[\mu_2,\mu_s]$ where $\mu=\mu_2$ when $\mu_t=0$ and $\mu=\mu_s$ when $\mu_t\to\infty$. Then,

$$\frac{d\{\frac{1}{\beta}\}}{d\mu}=\frac{\phi(1-\frac{\lambda_1}{\mu_1})}{\left[\lambda_1+\mu(1-\frac{\lambda_1}{\mu_1})\right]^2\left[\lambda_2-p\mu(1-\frac{\lambda_1}{\mu_1})\right]^2}\cdot h(\mu).$$

where

$$h(\mu) = -\frac{\mu_2}{\mu_s-\mu_2}\left(\frac{\mu_s}{\mu_s-\mu_2}\lambda_1+\lambda_2\right)\left(1-\frac{\lambda_1}{\mu_1}\right)^2\mu^2 - 2\frac{\mu_2}{\mu_s-\mu_2}\lambda_1\left(1-\frac{\lambda_1}{\mu_1}\right)\left[\lambda_2-\frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\mu + \lambda_1\left[\lambda_1+\frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\left[\lambda_2-\frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right].$$

Note that $h(\mu)$ is a concave quadratic function of μ. Also, $h(\mu)$ is decreasing in μ for $\mu\in[\mu_2,\infty)$ since the symmetric center of the concave quadratic function is below μ_2. So the sign of $h(\cdot)$ on $[\mu_2,\mu_s]$ depends on the value of

$$h(\mu_2) = \frac{\lambda_1\lambda_2}{\mu_s-\mu_2}\left[\lambda_1+\mu_2\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\left\{-\mu_2\left[1+\frac{\mu_2}{\lambda_1}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]+\mu_s\left[1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]\right\},$$
$$h(\mu_s) = \lambda_1^2\lambda_2 - \frac{\mu_s\mu_2}{\mu_s-\mu_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\left[\lambda_1+\mu_s\left(1-\frac{\lambda_1}{\mu_1}\right)\right](\lambda_1+\lambda_2).$$

First, $h(\mu_2)>0$ if and only if

$$-\mu_2\left[1+\frac{\mu_2}{\lambda_1}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]+\mu_s\left[1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)\right]>0,$$

i.e., $1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)>0$ and $\mu_s>\hat{\mu}_s$, where

$$\hat{\mu}_s = \frac{1+\frac{\mu_2}{\lambda_1}\left(1-\frac{\lambda_1}{\mu_1}\right)}{1-\frac{\mu_2}{\lambda_2}\left(1-\frac{\lambda_1}{\mu_1}\right)}\mu_2.$$

Second, $h(\mu_s)\geq 0$ if and only if $-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})^2\mu_2\mu_s^2+\lambda_1\left[\lambda_1\lambda_2-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})\mu_2\right]\mu_s-\lambda_1^2\lambda_2\mu_2\geq 0$. If we treat the left hand size of this inequality as a quadratic function of μ_s, then it holds for some $\mu_s\in[\mu_s^\dagger,\mu_s^\ddagger]$ if and only if its discriminant is non-negative, i.e.,

$$\lambda_1\leq\lambda_2\left[1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}-2\sqrt{1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}}\right],$$

which is equivalent to $\lambda_1>\frac{3\mu_1\mu_2}{\mu_1+3\mu_2}$ and $\lambda_2\geq\hat{\lambda}_2$ where

$$\hat{\lambda}_2 = \frac{\lambda_1}{1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}-2\sqrt{1+\frac{\lambda_1}{\mu_2(1-\frac{\lambda_1}{\mu_1})}}}.$$

Furthermore, the values of $\mu_s^\dagger\leq\mu_s^\ddagger$ can be calculated by the quadratic formula as

$$\frac{\lambda_1\lambda_2-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})\mu_2\pm\sqrt{\left[\lambda_1\lambda_2-(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})\mu_2\right]^2-4(\lambda_1+\lambda_2)\lambda_2(1-\frac{\lambda_1}{\mu_1})^2\mu_2^2}}{2(\lambda_1+\lambda_2)(1-\frac{\lambda_1}{\mu_1})^2\mu_2}\lambda_1.$$

Due to the monotonicity of $h(\cdot)$ on $[\mu_2,\mu_s]$, we know $\hat{\mu}_s<\mu_s^\dagger\leq\mu_s^\ddagger$ when $(\mu_s^\dagger,\mu_s^\ddagger)$ exists. Now we are ready to discuss the sign of $h(\cdot)$.

1. If $\lambda_2 \le \mu_2\big(1-\frac{\lambda_1}{\mu_1}\big)$ or $\frac{1}{\mu_s} \ge \frac{1}{\hat{\mu}_s}$, then $h(\mu_2) \le 0$ and hence $h(\mu) \le 0$ or $\frac{d\{\frac{1}{\beta}\}}{d\mu} \le 0$ for all $\mu \in [\mu_2, \mu_s]$, which implies that β decreases in $\frac{1}{\mu_t}$.
2. If $\lambda_2 > \mu_2\big(1-\frac{\lambda_1}{\mu_1}\big)$ and $\frac{1}{\mu_s} < \frac{1}{\hat{\mu}_s}$, then $h(\mu_2) > 0$. It remains to discuss the sign of $h(\mu_s)$.
   - If $\lambda_1 > \frac{3\mu_1\mu_2}{\mu_1+3\mu_2}$, $\lambda_2 \ge \hat{\lambda}_2$ and $\frac{1}{\mu_s^\ddagger} \le \frac{1}{\mu_s} \le \frac{1}{\mu_s^\dagger}$, then $h(\mu_s) \ge 0$ and $h(\mu) \ge 0$ for all $\mu \in [\mu_2, \mu_s]$, which implies that β always increases in $\frac{1}{\mu_t}$ and hence $\frac{1}{\hat{\mu}_t} = 0$.
   - Otherwise, $h(\mu_s) < 0$ and $h(\mu)$ is first positive and then negative as μ increases from μ_2 to μ_s. This implies that β first decreases and then increases in $\frac{1}{\mu_t}$, and $\frac{1}{\hat{\mu}_t} > 0$.

□

A.3. Proof of Proposition 1

Suppose that the feasible region of Problem (13) is nonempty as the parameters change in all three cases. Since TH_2 is increasing in $\frac{1}{\mu_t}$, the optimization problem reduces to one of finding the largest $\frac{1}{\mu_t}$ that satisfies the delay constraint $\beta \le \eta$. By Corollary 1, β either monotonically decreases in $\frac{1}{\mu_t}$ (as in Figure 2(a)–(b)), in which case $\frac{1}{\mu_t^*} = \infty$, or first decreases and then increases in $\frac{1}{\mu_t}$ (as in Figure 2(c)–(d)). In the latter case, if η is large, $\frac{1}{\mu_t} = \infty$ is feasible and hence optimal. Otherwise, the line $\beta = \eta$ crosses the β curve at most twice or touches its lowest point and $\frac{1}{\mu_t^*}$ is finite and equal to the larger intersection, which lies in the increasing part of the curve.

By (33) and (34), we can easily see that β increases in the cases (2) and (3) in this proposition for a given $1/\mu_t$, i.e., the curves in Figure 2 move upwards as the parameters change in the cases (2) and (3), and hence $\frac{1}{\mu_t^*}$ will either remain as ∞ or decrease as long as the feasible region is still feasible. For the case (1), while keeping $\lambda_1 + \lambda_2 = C$,

$$\frac{d\{\frac{1}{\beta}\}}{d\lambda_1} = -\left[\mu - \frac{\lambda_1(\mu-\mu_2)}{\mu_2}\right] \cdot \frac{\phi(1-\frac{\lambda_1}{\mu_1})\left[\lambda_1 + p\mu(1-\frac{\lambda_1}{\mu_1})\right]}{\left[\lambda_1 + \mu(1-\frac{\lambda_1}{\mu_1})\right]^2\left[\lambda_2 - p\mu(1-\frac{\lambda_1}{\mu_1})\right]} - \frac{d\{\frac{1}{\beta}\}}{d\mu} \le -\frac{d\{\frac{1}{\beta}\}}{d\mu}.$$

This implies that $\frac{d\{\frac{1}{\beta}\}}{d\lambda_1} \le 0$ if $\frac{d\{\frac{1}{\beta}\}}{d\mu} \ge 0$. Thus, the increasing part, where $\frac{d\beta}{d\{\frac{1}{\mu_t}\}} \ge 0$, of the β curve move upwards as λ_1 increases, and hence $\frac{1}{\mu_t^*}$ will either remain as ∞ or decrease as λ_1 increases.

Since $TH_2 = p\mu\big(1-\frac{\lambda_1}{\mu_1}\big) = \frac{\mu_2\mu_s}{\mu_t+\mu_s}\big(1-\frac{\lambda_1}{\mu_1}\big)$, it is easy to see that TH_2^* decreases in cases (1) and (2). In case (3), since both μ_2 and μ_t^* increases, the change of TH_2^* is not known. □

Appendix B: Analysis of the Underlying Stochastic Process

Note that, to obtain the system dynamics, we need to keep track of the status of the unlicensed users in service, i.e., in transmission or sensing, as the actual length of a service session is a phase-type rather than exponential. Although we are able to obtain the system dynamics and the fluid approximation when the length of a service session is a phase-type, in this paper we will only present the system dynamics and all the subsequent analysis as if the length of a service session

were exponential with the same mean for the following reasons. (1) Do not burden the reader with heavy notation and tedious mathematical expressions with only a single phase in each service session. As a result, the dynamics and subsequent analysis are much easier to understand and more intuitive. (2) The fluid approximation with a phase-type (two exponential phases) service session can be obtained following similar arguments. (3) The fluid approximations with an exponential or phase-type service session lead to exactly the same steady-state performance as the rates at which the unlicensed users leave and enter service are the same in both cases.

B.1. System Dynamics

In addition to notation introduced in Section 3, let

$S_i^n(t) =$ total number of type i users who have completed their transmission by t,

$D_2^n(t) =$ total number of service sessions completed by the unlicensed users by t,

$C_2^n(t) =$ total number of times the unlicensed users in the orbit queue have performed sensing by t.

It is easy to see that $S_1^n(t)$, $D_2^n(t)$, and $C_2^n(t)$ are random-time-changed Poisson processes with the rates $\mu_1 Z_1^n(t)$, $\mu Z_2^n(t)$, and $\theta Q_2^n(t)$, respectively. Since an unlicensed user will leave the system at the end of a service session with probability p, $S_2^n(t)$ is a "thinned" Poisson process of $D_2^n(t)$ with a time-varying rate $p\mu Z_2^n(t)$. Next, we derive the dynamics of $Z_i^n(t)$ and $Q_i^n(t)$ for $i = 1, 2$.

Note that the number of licensed users in service increases whenever an arriving licensed user sees an idle channel or a waiting licensed user sees a licensed user completing service or an unlicensed user finishing a session, and decreases whenever a licensed user completes his service. Likewise, the queue length of the licensed users increases whenever an arriving licensed user sees a busy system and decreases whenever a waiting licensed user sees a service or session completion. Thus, we have

$$\begin{aligned} Z_1^n(t) =& Z_1^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)>0\}} d\Lambda_1^n(s) + \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} d[S_1^n(s) + D_2^n(s)] - S_1^n(t), \\ Q_1^n(t) =& Q_1^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)=0\}} d\Lambda_1^n(s) - \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} d[S_1^n(s) + D_2^n(s)]. \end{aligned}$$

The dynamics of the unlicensed users is more complex as they may go back and forth between in service and waiting. The number of unlicensed users in service increases whenever a new arrival or waiting unlicensed user sees an idle channel and decreases whenever an unlicensed user finishes his transmission or is interrupted. The number of unlicensed users in the orbit queue increases whenever an arriving unlicensed user sees a busy system or an unlicensed user is interrupted but is willing to wait and decreases whenever a waiting unlicensed user enters service or abandons the system. Then,

$$Z_2^n(t) = Z_2^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)>0\}} d[\Lambda_2^n(s) + C_2^n(s)] - \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}} d[D_2^n(s) - S_2^n(s)] - S_2^n(t),$$

$$Q_2^n(t) = Q_2^n(0) + \int_0^t \mathbf{1}_{\{I^n(s)=0\}}[1 - B_\Lambda^n(s)]d\Lambda_2^n(s) + \int_0^t \mathbf{1}_{\{Q_1^n(s)>0\}}[1 - B_D^n(s)]d[D_2^n(s) - S_2^n(s)]$$
$$- \int_0^t \left[\mathbf{1}_{\{I^n(s)>0\}} + \mathbf{1}_{\{I^n(s)=0\}} B_C^n(s)\right] dC_2^n(s),$$

where $B_\Lambda^n(s)$, $B_C^n(s)$ and $B_D^n(s)$ are Bernoulli random variables with parameter ϕ at any s.

B.2. Proof of Theorem 2

Step 1: Martingale Representation. Let

$$\bar{M}_{\Lambda,i} = \bar{\Lambda}_i^n(t) - \bar{\lambda}_i^n t, \qquad \bar{M}_{S,1}^n = \bar{S}_1^n(t) - \int_0^t \mu_1 \bar{Z}_1^n(s)ds, \quad \bar{M}_{S,2}^n = \bar{S}_2^n(t) - \int_0^t p\mu \bar{Z}_2^n(s)ds,$$
$$\bar{M}_{C,2}^n = \bar{C}_2^n(t) - \int_0^t \theta \bar{Q}_2^n(s)ds, \quad \bar{M}_{D,2}^n = \bar{D}_2^n(t) - \int_0^t \mu \bar{Z}_2^n(s)ds$$

be the martingales corresponding to the processes. Recall $m^n(t)$ defined in (15). Then, we can rewrite the system dynamics as

$$\bar{Z}_1^n(t) = \bar{Z}_1^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)<0\}} d\bar{M}_{\Lambda,1}^n(s) + \int_0^t \mathbf{1}_{\{m^n(s)>0\}} d[\bar{M}_{S,1}^n(s) + \bar{M}_{D,2}^n(s)] - \bar{M}_{S,1}^n(t)$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)<0\}} \bar{\lambda}_1^n ds + \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [\mu_1 \bar{Z}_1^n(s) + \mu \bar{Z}_2^n(s)]ds - \int_0^t \mu_1 \bar{Z}_1^n(s)ds,$$
$$\bar{Q}_1^n(t) = \bar{Q}_1^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} d\bar{M}_{\Lambda,1}^n(s) - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} d[\bar{M}_{S,1}^n(s) + \bar{M}_{D,2}^n(s)]$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} \bar{\lambda}_1^n ds - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [\mu_1 \bar{Z}_1^n(s) + \mu \bar{Z}_2^n(s)]ds,$$
$$\bar{Z}_2^n(t) = \bar{Z}_2^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)<0\}} d[\bar{M}_{\Lambda,2}^n(s) + \bar{M}_{C,2}^n(s)] - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} d[\bar{M}_{D,2}^n(s) - \bar{M}_{S,2}^n(s)] - \bar{M}_{S,2}^n(t)$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)<0\}} \left[\bar{\lambda}_2^n + \theta \bar{Q}_2^n(s)\right] ds - \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [\mu \bar{Z}_2^n(s) - p\mu \bar{Z}_2^n(s)]ds - \int_0^t p\mu \bar{Z}_2^n(s)ds,$$
$$\bar{Q}_2^n(t) = \bar{Q}_2^n(0) + \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} [1 - B_\Lambda^n(s)] d\bar{M}_{\Lambda,2}^n(s) + \int_0^t \mathbf{1}_{\{m^n(s)>0\}} [1 - B_D^n(s)] d[\bar{M}_{D,2}^n(s) - \bar{M}_{S,2}^n(s)]$$
$$- \int_0^t \left[\mathbf{1}_{\{m^n(s)<0\}} + \mathbf{1}_{\{m^n(s)\geq 0\}} B_C^n(s)\right] d\bar{M}_{C,2}^n(s) + \int_0^t \mathbf{1}_{\{m^n(s)\geq 0\}} (1-\phi)\bar{\lambda}_2^n ds$$
$$+ \int_0^t \mathbf{1}_{\{m^n(s)>0\}} (1-\phi)[\mu \bar{Z}_2^n(s) - p\mu \bar{Z}_2^n(s)]ds - \int_0^t \left[\mathbf{1}_{\{m^n(s)<0\}} + \mathbf{1}_{\{m^n(s)\geq 0\}} \phi\right] \theta \bar{Q}_2^n(s)ds.$$

The dynamics of the process depends on the state of $m^n(t) \in \mathbb{Z}$. We compactify $\mathbb{Z}$ by letting $\bar{\mathbb{Z}} = \mathbb{Z} \cup \{\pm\infty\}$ (e.g., Perry and Whitt 2013) and denote by $\mathbb{M}$ the space of all measures ν on $[0,\infty) \times \bar{\mathbb{Z}}$ satisfying $\nu([0,t] \times \bar{\mathbb{Z}}) = t$. Consider the random measure $\nu^n(\cdot) \in \mathbb{M}$ defined by

$$\nu^n((0,t) \times \Gamma) = \int_0^t \mathbf{1}_{\{m^n(u)\in\Gamma\}} du \tag{35}$$

for all $t \in (0,\infty)$ and measurable $\Gamma \subset \bar{\mathbb{Z}}$.

Step 2: Tightness. Let $\mathbb{D}_{\mathbb{R}^4}[0,\infty)$ be the space of all right-continuous $\mathbb{R}^4$-valued functions with left limits defined on the real line. We show that the sequence $\{\bar{X}^n(\cdot),\nu^n\}$ is relatively compact in $\mathbb{D}_{\mathbb{R}^4}[0,\infty)\times\mathbb{M}$ by showing that both $\{\bar{X}^n(\cdot)\}$ and $\{\nu^n\}$ are relatively compact.

$\{\nu^n\}$ is relatively compact due to the compactness of $\mathbb{M}$, which follows from the compactness of $\bar{\mathbb{Z}}$ by Prohorov's theorem (cf. Theorem 11.6.1 in Whitt 2002). $\{\bar{X}^n(\cdot)\}$ is relatively compact in $\mathbb{D}_{\mathbb{R}^4}[0,\infty)$ if it satisfies the conditions (6.3) and (6.4) of Theorem 11.6.3 in Whitt (2002). For any $\epsilon>0$, there exists a $c>0$ such that

$$\mathbb{P}(|\bar{X}^n(0)|>c)<\epsilon,\ \text{for all } n\geq 1,$$

since $\bar{X}^n(0)\Rightarrow x(0)$. Thus, the initial states are stochastically bounded and hence condition (6.3) is satisfied.

To show that condition (6.4) is satisfied, for any $\delta>0$, we define the modulus of continuity for a function $y(\cdot)$ as

$$w(y(\cdot),\delta,T)=\sup_{|t-s|\leq\delta,\ s,t\in[0,T]}|y(t)-y(s)|,$$

and show that, for any $\epsilon,\eta,T>0$, there exists a δ such that

$$\mathbb{P}\left(w(\bar{X}^n(\cdot),\delta,T)>\epsilon\right)<\eta, \tag{36}$$

for all n large enough. To do so, we decompose the oscillations of the process $X^n(t)$. Take the component $Q_2^n(t)$ for example,

$$\begin{aligned}|\bar{Q}_2^n(t)-\bar{Q}_2^n(s)|\leq&|\bar{M}_{\Lambda,2}^n(t)-\bar{M}_{\Lambda,2}^n(s)|+|\bar{M}_{D,2}^n(t)-\bar{M}_{D,2}^n(s)|+|\bar{M}_{C,2}^n(t)-\bar{M}_{C,2}^n(s)|\\&+\int_s^t(1-\phi)\left[\bar{\lambda}_2^n+(1-p)\mu\right]du+\int_s^t\theta\bar{Q}_2^n(u)du.\end{aligned}$$

Since the fourth term on the right hand side is deterministic and uniformly continuous, there exists a $\delta'>0$ such that it is less than $\frac{\epsilon}{5}$, i.e., $\mathbb{P}\left(w\left(\int_0^t(1-\phi)\left[\bar{\lambda}_2^n+(1-p)\mu\right]du,\delta',T\right)>\frac{\epsilon}{5}\right)=0$. Furthermore, since $\bar{M}_{\Lambda,2}^n$, $\bar{M}_{D,2}^n$ and $\bar{M}_{C,2}^n$ are square-integrable martingales, they weakly converge to 0 as $n\to\infty$ by Doob's inequality and hence their oscillations can also be controlled, e.g., $\mathbb{P}\left(w\left(\bar{M}_{\Lambda,2}^n(\cdot),\delta',T\right)>\frac{\epsilon}{5}\right)<\frac{\eta}{5}$ for large enough n. For the last term, we can bound the process $\bar{Q}_2^n(t)$ by a stable and bounded auxiliary one with simple dynamics. Thus, there exists a constant c such that $\mathbb{P}\left(\sup_{t\in[0,T]}\{\bar{Q}_2^n(t)\}>c\right)\leq\frac{\eta}{5}$ for all large n. Let $\delta=\min\left\{\frac{\epsilon}{5\theta c},\delta'\right\}$. Then,

$$\mathbb{P}\left(\sup_{|t-s|\leq\delta,\ s,t\in[0,T]}\left\{\int_s^t\theta\bar{Q}_2^n(u)du\right\}>\frac{\epsilon}{5}\right)\leq\frac{\eta}{5}$$

and

$$\mathbb{P}\left(w\left(\bar{Q}_2^n(t),\delta,T\right)>\epsilon\right)\leq\frac{\eta}{5}+\frac{\eta}{5}+\frac{\eta}{5}+0+\frac{\eta}{5}<\eta$$

for large enough n. Following a similar procedure, we can control the oscillations of $\bar{Z}_1^n(t)$ and $\bar{Z}_2^n(t)$, which implies (36) and condition (6.4) are satisfied. By Theorem 11.6.3 of Whitt (2002), $\{\bar{X}^n(\cdot)\}$ is relatively compact.

Step 3: The Limiting Process. Since $\{\bar{X}^n(\cdot), \nu^n\}$ is relatively compact, there exists a convergent subsequence whose limit is denoted by $\{x(\cdot), \nu\}$. Then, by the continuous mapping theorem, the subsequence satisfies

$$z_1(t) = z_1(0) + \lambda_1 \nu([0,t] \times \bar{\mathbb{Z}}^-) + \int_{[0,t]\times\bar{\mathbb{Z}}^+} [\mu_1 z_1(s) + \mu z_2(s)] \nu(ds \times dy) - \int_0^t \mu_1 z_1(s) ds, \tag{37}$$

$$q_1(t) = q_1(0) + \lambda_1 \nu([0,t] \times \bar{\mathbb{N}}) - \int_{[0,t]\times\bar{\mathbb{Z}}^+} [\mu_1 z_1(s) + \mu z_2(s)] \nu(ds \times dy) - \int_0^t \mu_1 z_1(s) ds, \tag{38}$$

$$\begin{aligned} z_2(t) = z_2(0) &+ \int_{[0,t]\times\bar{\mathbb{Z}}^-} [\lambda_2 + \theta q_2(s)] \nu(ds \times dy) - \int_{[0,t]\times\bar{\mathbb{Z}}^+} (1-p)\mu z_2(s) \nu(ds \times dy) \\ &- \int_0^t p\mu z_2(s) ds, \end{aligned} \tag{39}$$

$$\begin{aligned} q_2(t) = q_2(0) &+ \int_{[0,t]\times\bar{\mathbb{N}}} (1-\phi)[\lambda_2 + \theta q_2(s)] \nu(ds \times dy) + \int_{[0,t]\times\bar{\mathbb{Z}}^+} (1-\phi)(1-p)\mu z_2(s) \nu(ds \times dy) \\ &- \int_0^t \theta q_2(s) \nu(ds \times dy), \end{aligned} \tag{40}$$

where $\bar{\mathbb{N}} = \{0, 1, 2, \cdots, +\infty\}$, $\bar{\mathbb{Z}}^+ = \{1, 2, \cdots, +\infty\}$ and $\bar{\mathbb{Z}}^- = \{-1, -2, \cdots, -\infty\}$.

Kurtz (1992) shows in Lemma 1.4 that the limit measure $\nu(\cdot)$ can be separated into a product form. That is, for any Borel set $\Gamma_1 \subset [0, \infty)$ and $\Gamma_2 \subset \bar{\mathbb{Z}}$,

$$\nu(\Gamma_1 \times \Gamma_2) = \int_{\Gamma_1} \pi_s(\Gamma_2) ds, \tag{41}$$

where π_s is a probability measure on $\bar{\mathbb{Z}}$. Next, we complete the proof of Theorem 2 by deriving the expression of $\pi_s(\cdot)$. Let $\{m(\cdot|x) : x = (z_1, z_2, q_2) \in \mathbb{R}_+^3\}$ be a family of continuous-time Markov chains with transition rates dependent on x as follows:

$$m(\cdot|x) \to \begin{cases} m(\cdot|x) + 1, \text{ at the rate } \mathbf{1}_{\{m(\cdot|x)<0\}}(\lambda_1 + \lambda_2 + \theta q_2) + \mathbf{1}_{\{m(\cdot|x)\geq 0\}}\lambda_1, \\ m(\cdot|x) - 1, \text{ at the rate } \mathbf{1}_{\{m(\cdot|x)\leq 0\}}(\mu_1 z_1 + p\mu z_2) + \mathbf{1}_{\{m(\cdot|x)>0\}}(\mu_1 z_1 + \mu z_2). \end{cases}$$

We now show that π_s is the stationary distribution of $m(\cdot|x(s))$ for $s \in (0, \infty)$.

For any bounded continuous function f on $\bar{\mathbb{Z}}$,

$$\begin{aligned} \frac{f(m^n(t))}{n} =& \frac{f(m^n(0))}{n} + \int_0^t [f(m^n(s)+1) - f(m^n(s))] \left\{ d\bar{M}_{\Lambda,1}^n(s) + \mathbf{1}_{\{m^n(s)<0\}} d[\bar{M}_{\Lambda,2}^n(s) + \bar{M}_{C,2}^n(s)] \right\} \\ &+ \int_0^t [f(m^n(s)-1) - f(m^n(s))] \left\{ \bar{M}_{S,1}^n(s) + \mathbf{1}_{\{m^n(s)\leq 0\}} d\bar{M}_{S,2}^n(s) + \mathbf{1}_{\{m^n(s)>0\}} d\bar{M}_{D,2}^n \right\} \\ &+ \int_0^t [f(m^n(s)+1) - f(m^n(s))] \left\{ \bar{\lambda}_1^n + \mathbf{1}_{\{m^n(s)<0\}} [\bar{\lambda}_2^n + \theta \bar{Q}_2^n(s)] \right\} ds \\ &+ \int_0^t [f(m^n(s)-1) - f(m^n(s))] \left\{ \mu_1 \bar{Z}_1^n(s) + \mathbf{1}_{\{m^n(s)\leq 0\}} p\mu \bar{Z}_2^n(s) + \mathbf{1}_{\{m^n(s)>0\}} \mu \bar{Z}_2^n(s) \right\} ds. \end{aligned}$$

As $n \to \infty$, the martingale parts (the second and third terms) converge to zero by Doob's inequality and $\frac{f(m^n(t))-f(m^n(0))}{n} \to 0$ since f is bounded. Therefore, the sum of the last two terms should also converge to zero, which, by the continuous mapping theorem, leads to

$$\int_{[0,t]\times\bar{\mathbb{Z}}} [f(y+1)-f(y)]\left\{\mathbf{1}_{\{y<0\}}[\lambda_1+\lambda_2+\theta q_2(s)]+\mathbf{1}_{\{y\geq 0\}}\lambda_1\right\}\nu(ds\times dy)$$
$$+\int_{[0,t]\times\bar{\mathbb{Z}}} [f(y-1)-f(y)]\left\{\mathbf{1}_{\{y\leq 0\}}[\mu_1 z_1(s)+p\mu z_2(s)]+\mathbf{1}_{\{y>0\}}[\mu_1 z_1(s)+\mu z_2(s)]\right\}\nu(ds\times dy)=0,$$

for any t. Hence, by (41),

$$\int_{\bar{\mathbb{Z}}} [f(y+1)-f(y)]\left\{\mathbf{1}_{\{y<0\}}[\lambda_1+\lambda_2+\theta q_2(s)]+\mathbf{1}_{\{y\geq 0\}}\lambda_1\right\}$$
$$+[f(y-1)-f(y)]\left\{\mathbf{1}_{\{y\leq 0\}}[\mu_1 z_1(s)+p\mu z_2(s)]+\mathbf{1}_{\{y>0\}}[\mu_1 z_1(s)+\mu z_2(s)]\right\}\pi_s(dy)=0$$

for almost all s and it follows from Proposition 4.9.2 of Ethier and Kurtz (1986) that π_s is the stationary (invariant) measure for $m(\cdot|x(s))$. Thus, the steady-state probability can be obtained as follows:

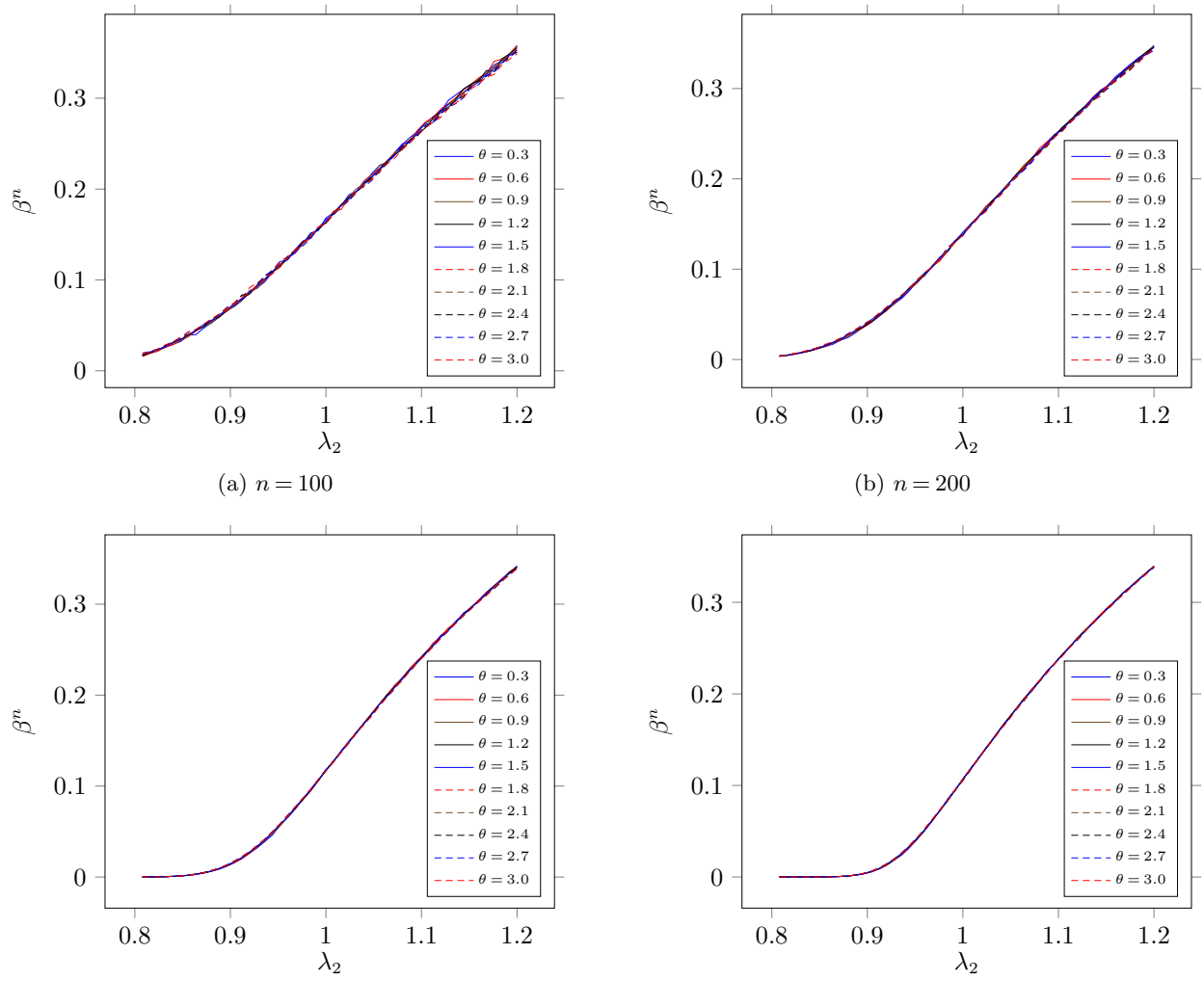
- For $q_1(s)>0$, $m(\cdot|x(s))=\infty$, $\pi_s(\bar{\mathbb{N}})=\pi_s(\bar{\mathbb{Z}}^+)=1$ and $\pi_s(\bar{\mathbb{Z}}^-)=0$.
- For $z_1(s)+z_2(s)<1$, $m(\cdot|x(s))=-\infty$, $\pi_s(\bar{\mathbb{N}})=\pi_s(\bar{\mathbb{Z}}^+)=0$ and $\pi_s(\bar{\mathbb{Z}}^-)=1$.
- For $q_1(s)=0$ and $z_1(s)+z_2(s)=1$,
  $\pi_s(\bar{\mathbb{N}}) = \min\left\{\left(\frac{[\lambda_1+\lambda_2+\theta q_2(s)-\mu_1 z_1(s)-p\mu z_2(s)][\mu_1 z_1(s)+\mu z_2(s)]}{[\lambda_1+\lambda_2+\theta q_2(s)][\mu_1 z_1(s)+\mu z_2(s)]-\lambda_1[\mu_1 z_1(s)+p\mu z_2(s)]}\right)^+, 1\right\}$, $\pi_s(\bar{\mathbb{Z}}^-) = 1-\pi_s(\bar{\mathbb{N}})$ and $\pi_s(\bar{\mathbb{Z}}^+) = \min\left\{\frac{\lambda_1}{\mu_1 z_1(s)+\mu z_2(s)}\pi_s(\bar{\mathbb{N}}), 1\right\}$.

By (35) and (41), $\pi_s(\bar{\mathbb{N}}) = \lim_{n\to\infty}\mathbb{P}(I^n(s)=0)$ and $\pi_s(\bar{\mathbb{Z}}^+) = \lim_{n\to\infty}\mathbb{P}(Q_1^n(s)>0)$. If we let $\beta(s):=\pi_s(\bar{\mathbb{N}})$ and $\alpha(s):=\pi_s(\bar{\mathbb{Z}}^+)$, then $\beta(s)$ and $\alpha(s)$ represent the instantaneous probability that an arriving licensed user is delayed and the probability that an unlicensed user has to release the channel after a service session, respectively. By substituting them into (37)–(40) and taking the derivative with respect to t, we can easily show that the limit $x(t)$ satisfies the differential equations (4)–(7).

Appendix C: A Numerical Study on the Impact of the Sensing Frequency

We simulate the delay probability and throughput rate for $n \in \{100, 200, 500, 1000\}$, $\lambda_1 \in \{0.02, 0.05\}, \lambda_2 \in [0.8, 1.2]$, $\frac{1}{\mu_t} \in \{\infty, 0.5, 0.25, 0.125\}$. For each combination, the system performance is almost identical when we vary $\theta \in \{0.3, 0.6, \cdots, 3.0\}$. This shows that the performance of unscaled systems is indeed insensitive to the sensing frequency and our fluid limits represent the actual systems accurately.

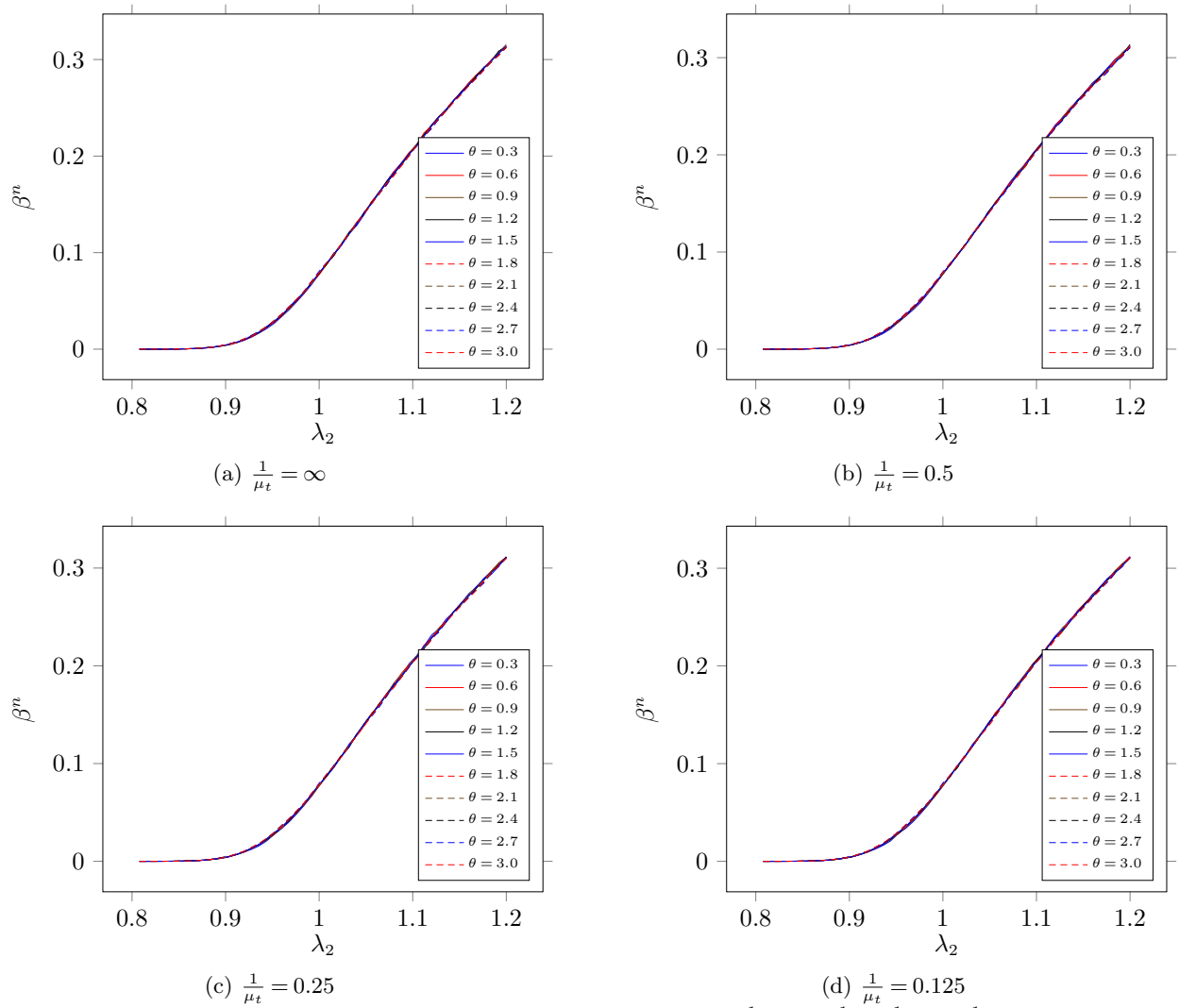
For illustration purposes, we consider systems with $\frac{1}{\mu_1}=\frac{1}{\mu_2}=1$, $\frac{1}{\mu_s}=0.0001$, $\phi=0.5$ and $\lambda_i^n = n\lambda_i$. We plot the delay probability as a function of λ_2 for (1) different n when $\lambda_1=0.05$ and $\frac{1}{\mu_t}=0.5$ in Figure 9, and (2) different $\frac{1}{\mu_t}$ when $n=500$ and $\lambda_1=0.02$ in Figure 10. As one can see, the delay probability curves are almost identical for $\theta \in \{0.3, 0.6, \cdots, 3.0\}$.



(a) $n = 100$ (b) $n = 200$

(c) $n = 500$ (d) $n = 1000$

Figure 9 The delay probability as a function of λ_2 for different θ and n when $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 1$, $\frac{1}{\mu_s} = 0.0001$, $\phi = 0.5$, $\lambda_i^n = n\lambda_i$, $\lambda_1 = 0.05$ and $\frac{1}{\mu_t} = 0.5$



(a) $\frac{1}{\mu_t} = \infty$ (b) $\frac{1}{\mu_t} = 0.5$

(c) $\frac{1}{\mu_t} = 0.25$ (d) $\frac{1}{\mu_t} = 0.125$

Figure 10 The delay probability as a function of λ_2 for different θ and $\frac{1}{\mu_t}$ when $\frac{1}{\mu_1} = \frac{1}{\mu_2} = 1$, $\frac{1}{\mu_s} = 0.0001$, $\phi = 0.5$, $\lambda_i^n = n\lambda_i$, $n = 500$ and $\lambda_1 = 0.02$