To Pool or Not to Pool: Queueing Design for Large-Scale Service Systems (E-Companion)

We provide the proofs of all theorems and propositions in this e-companion, along with additional numerical examples. In particular, the proof of Theorem 5 is given in Section EC.5, where performance formulas for all $\rho > 1$ under the two queue structures are provided.

Before presenting the proofs, let us introduce the following notions to be used in the analysis of the fluid model: For a function $f : [0,\infty) \to \mathbb{R}$, we say $t \ge 0$ is *regular* if f is differentiable at t. In the proofs below, we implicitly assume t to be a regular point of f when we write $f'(t)$. We say f converges to $a \in \mathbb{R}$ at rate $\theta > 0$, if there exists some $c > 0$ such that $|f(t) - a| \le c \cdot \mathrm{e}^{-\theta t}$ for all $t \ge 0$.

EC.1. Proof of Theorem 1

Existence. We prove the existence of a solution by construction. Let us consider the following dynamical system:

$$\begin{cases} \bar{P}_i(t) = \bar{P}_i(0) - \bar{V}_{i-1}(t) + \bar{V}_i(t) - (\mu + \theta(i-1)) \displaystyle\int_0^t (\bar{P}_i(s) - \bar{P}_{i+1}(s))\,\mathrm{d}s, & \text{(EC.1)} \\ \bar{P}_i(t) \ge 0, & \text{(EC.2)} \\ \bar{P}_N(t) = 1, & \text{(EC.3)} \\ \bar{V}_0(t) = \rho\mu t, & \text{(EC.4)} \\ \bar{V}_i \text{ is non-decreasing with } \bar{V}_i(0) = 0, & \text{(EC.5)} \\ \displaystyle\int_0^\infty \mathbb{1}_{\{\bar{P}_i(t-)>0\}}\,\mathrm{d}\bar{V}_i(t) = 0, & \text{(EC.6)} \end{cases}$$

for $i = 1, \ldots, N-1$. Write $\bar{\mathbf{P}}^N(t) := (\bar{P}_i(t) : i = 1, \ldots, N-1)$ and $\bar{\mathbf{V}}^N(t) = (\bar{V}_i(t) : i = 1, \ldots, N-1)$. Equation (EC.1) can be written into a vector form:

$$\bar{\mathbf{P}}^N(t) = \bar{\mathbf{P}}^N(0) + \bar{\mathbf{E}}^N(t) - \int_0^t \mathbf{H}^N \bar{\mathbf{P}}^N(s)\,\mathrm{d}s + \mathbf{R}^N \bar{\mathbf{V}}^N(t),$$

where $\bar{\mathbf{E}}^N(t) := (\bar{E}_i^N(t) : i = 1, \ldots, N-1)$ is given by

$$\bar{E}_i^N(t) := \begin{cases} -\rho\mu t, & i = 1, \\ 0, & i = 2, \ldots, N-2, \\ (\mu + \theta(N-2))t, & i = N-1, \end{cases}$$

$\mathbf{H}^N$ is an $(N-1) \times (N-1)$ matrix with the (i,j)th entry given by

$$H_{ij} := \begin{cases} \mu + \theta(i-1), & i = j, \\ -(\mu + \theta(i-1)), & i = 1, \ldots, N-2,\ j = i+1, \\ 0, & \text{otherwise}, \end{cases}$$

and $\mathbf{R}^N$ is an $(N-1)\times(N-1)$ matrix with the (i,j)th entry given by

$$R_{ij} := \begin{cases} 1, & i=j, \\ -1, & i=2,\dots,N-1,\ j=i-1, \\ 0, & \text{otherwise.} \end{cases}$$

Since the inverse of $\mathbf{R}^N$ is a lower triangular matrix with all nonzero entries being one, it follows from Proposition 2 in Reed and Ward (2004) that the above dynamical system has a unique solution, with both $\bar{\mathbf{P}}^N(t)$ and $\bar{\mathbf{V}}^N(t)$ being continuous in t. We may also write (EC.1) as

$$\bar{P}_i(t) = \bar{P}_i(0) + \bar{V}_i(t) + (\mu+\theta(i-1))\int_0^t \bar{P}_{i+1}(s)\,\mathrm{d}s - \bar{V}_{i-1}(t) - (\mu+\theta(i-1))\int_0^t \bar{P}_i(s)\,\mathrm{d}s,$$

where $\bar{P}_i$ is the difference of two nondecreasing continuous functions. Then, $\bar{P}_i$ is of bounded variation, thus differentiable almost everywhere. Hence, $\bar{V}_i$ is differentiable almost everywhere too.

Let us prove $\bar{V}'_{i+1}(t) \le \bar{V}'_i(t)$ for $i=0,\dots,N-2$. If $\bar{P}_{i+1}(t)>0$, it is true because $\bar{V}'_{i+1}(t)=0$. If $\bar{P}_{i+1}(t)=0$, then $\bar{P}'_{i+1}(t)=0$, so that $\bar{V}'_i(t)-\bar{V}'_{i+1}(t)=(\mu+i\theta)\bar{P}_{i+2}(t)\ge 0$.

Next, we prove that $\bar{P}_{N-1}(t)\ne 0$ almost everywhere. If t is a regular point of $\bar{P}_{N-1}$ such that $\bar{P}_{N-1}(t)=0$, we must have $\bar{P}'_{N-1}(t)=0$. However, by (EC.1) and the fact that $\bar{V}'_{N-2}(t)\le \bar{V}'_0(t)=\rho\mu$,

$$\bar{P}'_{N-1}(t) = -\bar{V}'_{N-2}(t)+\bar{V}'_{N-1}(t)+(\mu+\theta(N-2)) \ge (\mu+\theta(N-2))-\rho\mu \ge (\mu+\theta\bar{q})-\rho\mu > 0.$$

This contradiction implies that $\bar{P}_{N-1}(t)\ne 0$ if t is regular. Therefore, $\bar{V}_{N-1}(t)=0$ for $t\ge 0$.

We construct a solution to (11)–(16) as follows:

1. For $i=1,\dots,N-1$, let $\bar{Q}_i(t):=1-\bar{P}_i(t)$ and $\bar{U}_i(t):=\bar{V}_i(t)$ for $t\ge 0$.
2. For $i\ge N$, let $\bar{Q}_i(t):=0$ and $\bar{U}_i(t):=0$ for $t\ge 0$.

Clearly, $(\bar{\mathbf{Q}},\bar{\mathbf{U}})$ satisfies (11)–(14) and (16), and $\bar{Q}_i(t)\le 1$ for all i.

Let us prove $\bar{Q}_{i+1}(t)\le \bar{Q}_i(t)$ for $t\ge 0$, by which we can deduce that $\bar{Q}_i(t)\ge 0$ for all i. We use backward induction, assuming that $\bar{Q}_{j+1}(t)\le \bar{Q}_j(t)$ for some $j\in\mathbb{N}$ and all $t\ge 0$. Clearly, the assumption holds for $j\ge N$. Suppose that there exists some $t_0\ge 0$ such that $\bar{Q}_j(t_0)>\bar{Q}_{j-1}(t_0)$. Because $\bar{Q}_{j-1}(0)-\bar{Q}_j(0)\ge 0$ and $\bar{Q}_{j-1}(t)-\bar{Q}_j(t)$ is continuous in t, we may find a regular point $t_1\in(0,t_0]$ such that $\bar{Q}'_{j-1}(t_1)-\bar{Q}'_j(t_1)<0$ and $\bar{Q}_{j-1}(t_1)-\bar{Q}_j(t_1)<0$. Then by (11) and (16),

$$\bar{Q}'_{j-1}(t_1) = \bar{U}'_{j-2}(t_1)-\bar{U}'_{j-1}(t_1)-(\mu+\theta(j-2))(\bar{Q}_{j-1}(t_1)-\bar{Q}_j(t_1))>0.$$

Similarly,

$$\bar{Q}'_j(t_1) = \bar{U}'_{j-1}(t_1)-\bar{U}'_j(t_1)-(\mu+\theta(j-1))(\bar{Q}_j(t_1)-\bar{Q}_{j+1}(t_1)).$$

Since $\bar{Q}_{j-1}(t_1)<1$, we have $\bar{U}'_{j-1}(t_1)=0$ by (14), which implies that $\bar{Q}'_j(t_1)\le 0$. Then, we deduce that $\bar{Q}'_{j-1}(t_1)-\bar{Q}'_j(t_1)>0$, a contradiction.

The Lipschitz continuity of the solution follows from (11)–(12) and (15)–(16).

Uniqueness. Suppose that there is another solution $(\check{\mathbf{Q}}, \check{\mathbf{U}})$ different from $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$, where $\check{\mathbf{Q}}(t) := (\check{Q}_i(t) : i \in \mathbb{N})$ and $\check{\mathbf{U}}(t) := (\check{U}_i(t) : i \in \mathbb{N}_0)$. Then, we may find some $\tau_0 > 0$ such that $(\check{\mathbf{Q}}(\tau_0), \check{\mathbf{U}}(\tau_0)) \neq (\bar{\mathbf{Q}}(\tau_0), \bar{\mathbf{U}}(\tau_0))$. Let $\tau_1 := \inf\{t \geq 0 : \check{Q}_N(t) \geq 1/2\}$ and $\tau_2 := \sup\{t \in [0, \tau_0 \wedge \tau_1] : (\check{\mathbf{Q}}(t), \check{\mathbf{U}}(t)) = (\bar{\mathbf{Q}}(t), \bar{\mathbf{U}}(t))\}$. Since $(\check{\mathbf{Q}}(0), \check{\mathbf{U}}(0)) = (\bar{\mathbf{Q}}(0), \bar{\mathbf{U}}(0))$, we have $\tau_2 < \infty$.

Using the fact that $\check{\mathbf{Q}}(\tau_2) = \bar{\mathbf{Q}}(\tau_2)$, we obtain $\check{Q}_i(\tau_2) = 0$ for $i \geq N$. Because $\check{Q}_N$ is right continuous, we may find some $\varepsilon_0 > 0$ such that $\check{Q}_N(t) < 1$ for $0 \leq t \leq \tau_2 + \varepsilon_0$. Then, $\check{U}_i(t) = 0$ for $i \geq N$ and $0 \leq t \leq \tau_2 + \varepsilon_0$. It follows from (11) and (15) that $\check{Q}_i(t) = 0$ for $i \geq N+1$ and $0 \leq t \leq \tau_2 + \varepsilon_0$.

Put

$$\check{\mathbf{P}}^{N+1}(t) := (1 - \check{Q}_i(t) : i = 1, \ldots, N), \quad \check{\mathbf{V}}^{N+1}(t) := (\check{U}_i(t) : i = 1, \ldots, N),$$
$$\bar{\mathbf{P}}^{N+1}(t) := (1 - \bar{Q}_i(t) : i = 1, \ldots, N), \quad \bar{\mathbf{V}}^{N+1}(t) := (\bar{U}_i(t) : i = 1, \ldots, N).$$

Both $(\check{\mathbf{P}}^{N+1}, \check{\mathbf{V}}^{N+1})$ and $(\bar{\mathbf{P}}^{N+1}, \bar{\mathbf{V}}^{N+1})$ satisfy (EC.1)–(EC.6) for $i = 1, \ldots, N$ and $0 \leq t \leq \tau_2 + \varepsilon_0$. Then, they must be identical because this dynamical system has a unique solution. This implies that $(\check{\mathbf{Q}}(t), \check{\mathbf{U}}(t)) = (\bar{\mathbf{Q}}(t), \bar{\mathbf{U}}(t))$ for $0 \leq t \leq \tau_2 + \varepsilon_0$, which contradicts the definition of τ_2.

EC.2. Proof of Theorem 2

We first prove (20). For $k \leq \bar{q} - 1$, let us write $\bar{Y}_k(t) := \sum_{i=1}^{k} \bar{Q}_i(t)$. Then,

$$\bar{Y}_k(t) = \bar{Y}_k(0) + \rho\mu t - \bar{U}_k(t) - \int_0^t \big(\mu\bar{Q}_1(s) + \theta(\bar{Y}_k(s) - \bar{Q}_1(s))\big)\,\mathrm{d}s + (\mu + \theta(k-1)) \int_0^t \bar{Q}_{k+1}(s)\,\mathrm{d}s.$$

If $\bar{Y}_k(t) < k$ for some $t \geq 0$, we have $\bar{Q}_k(t) < 1$, and thus $\bar{U}_k'(t) = 0$ by (14). Because $\mu\bar{Q}_1(t) + \theta(\bar{Y}_k(t) - \bar{Q}_1(t)) \leq \mu + \theta(k-1)$, then $\bar{Y}_k'(t) \geq (\rho - 1)\mu - \theta(k-1) = \theta(q + 1 - k) > 0$. We must have $\bar{Y}_k(t) = k$, and thus $\bar{Q}_k(t) = 1$, for $t \geq (k - \bar{Y}_k(0))/(\theta(q+1-k))$. This assertion also holds for $k = \bar{q}$ if q is not an integer.

Put $\bar{Z}(t) := \sum_{i=\bar{q}+2}^{N-1} \bar{Q}_i(t)$. If t is a regular point of $\bar{Z}$ such that $\bar{Q}_{\bar{q}+1}(t) < 1$, then by (11) and (14)–(15),

$$\bar{Z}'(t) = -\sum_{i=\bar{q}+2}^{N-1} (\mu + (i-1)\theta)(\bar{Q}_i(t) - \bar{Q}_{i+1}(t)) \leq -\theta\bar{Z}(t).$$

If $\bar{Q}_{\bar{q}+1}(t) = 1$, then $\bar{Q}_i(t) = 1$ for $i = 1, \ldots, \bar{q}$, and thus $\bar{Q}_i'(t) = 0$ for $i = 1, \ldots, \bar{q} + 1$. By (17), $\bar{X}'(t) = \rho\mu - \mu\bar{Q}_1(t) - \theta(\bar{X}(t) - \bar{Q}_1(t)) = (\rho - 1)\mu - \bar{q}\theta - \theta\bar{Z}(t) \leq -\theta(1 - r) - \theta\bar{Z}(t) \leq -\theta\bar{Z}(t)$. Therefore, $\bar{Z}'(t) = \bar{X}'(t) - \sum_{i=1}^{\bar{q}+1} \bar{Q}_i'(t) \leq -\theta\bar{Z}(t)$. Because $\bar{Z}'(t) \leq -\theta\bar{Z}(t)$ always holds, $\bar{Z}$ must converge to zero at rate θ. Hence, each $\bar{Q}_i$ will also converge to zero at the same rate for $i = \bar{q} + 2, \ldots, N - 1$. When q is an integer, the above argument is also valid if we take $\bar{Z}(t) := \sum_{i=\bar{q}+1}^{N-1} \bar{Q}_i(t)$. In this case, $\bar{Q}_{\bar{q}+1}$ will converge to zero at rate θ.

Since $\bar{Q}_1(t) = 1$ for $t \geq (1 - \bar{Q}_1(0))/(\theta q)$, it follows from (17) that $\bar{X}'(t) = (\rho - 1)\mu - \theta(\bar{X}(t) - 1)$, so that $\bar{X}(t)$ will converge to $q + 1$ at rate θ. By the previous results, we deduce that $\bar{Q}_{\bar{q}+1}(t)$ will converge to r at rate θ when q is not an integer, and that $\bar{Q}_{\bar{q}}(t)$ will converge to one at rate θ when q is an integer. Now, we obtain the convergence rate specified by (20) for $i = 1, \ldots, N-1$.

Clearly, $\mathbf{q}^*$ is an invariant state in $\mathbb{S}_N$. Let $\mathbf{q}$ be an arbitrary invariant state in $\mathbb{S}_N$. Then, $\bar{\mathbf{Q}}(t) = \mathbf{q}$ if we take $\bar{\mathbf{Q}}(0) = \mathbf{q}$. By (20), we must have $\mathbf{q} = \mathbf{q}^*$, so that $\mathbf{q}^*$ is the unique invariant state in $\mathbb{S}_N$.

EC.3. Proof of Theorem 3

We prove part (i) using the following tightness result, the proof of which is given later in this section.

LEMMA EC.1. *Under the conditions of part (i) of Theorem 3,* $\{(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) : n \in \mathbb{N}\}$ *is tight and the limit of any weakly convergent subsequence is a fluid solution, i.e., a solution to* (11)–(16) *for* $t \geq 0$ *almost everywhere.*

By Theorem 1, the dynamical system (11)–(16) has a unique solution $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$, which implies that all weakly convergent subsequences of $\{(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) : n \in \mathbb{N}\}$ must have the same limit. Therefore, $(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) \Rightarrow (\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ as $n \to \infty$.

We prove part (ii) in three steps. First, we show that $\mathbf{Q}^n$ has a unique steady-state distribution for each $n \in \mathbb{N}$. Second, we prove that $\bar{Q}_i^n(\infty) \Rightarrow 0$ as $n \to \infty$ when i is sufficiently large. Third, we prove that $\{\bar{\mathbf{Q}}^n(\infty) : n \in \mathbb{N}\}$ is tight and that the limit of any weakly convergent subsequence must be $\mathbf{q}^*$. We would thus obtain $\bar{\mathbf{Q}}^n(\infty) \Rightarrow \mathbf{q}^*$ as $n \to \infty$.

Step 1. As an irreducible continuous-time Markov chain, $\mathbf{Q}^n$ has a unique steady-state distribution if the empty state $\mathbf{0} := (0 : i \in \mathbb{N})$ is positive recurrent. This steady-state distribution will also be the limiting distribution. With $\mathbf{Q}^n(0) = \mathbf{0}$, let $\tau^n(\mathbf{0})$ be the first hitting time of state $\mathbf{0}$ by $\mathbf{Q}^n$ from other states. Since $\mathbf{Q}^n(t) = \mathbf{0}$ if and only if $X^n(t) = 0$, $\tau^n(\mathbf{0})$ is also the first hitting time of 0 by X^n from other states. We need to prove $\mathbb{E}[\tau^n(\mathbf{0})] < \infty$.

At time t, the instantaneous rate of customers leaving the system (either by service completion or by abandonment) satisfies $\mu Q_1^n(t) + \theta \sum_{i=2}^{\infty} (i-1)(Q_i^n(t) - Q_{i+1}^n(t)) \geq (\mu \wedge \theta) X^n(t)$. Consider an M/M/$\infty$ system that has arrival rate λ^n, mean service time $1/(\mu \wedge \theta)$, and initial condition $X_\infty^n(0) = 0$, where $X_\infty^n(t)$ is the number of customers at time t. Using the coupling method in the proof of Lemma 3 in Dong et al. (2015), we establish that

$$\{X^n(t) : t \geq 0\} \leq_{st} \{X_\infty^n(t) : t \geq 0\}, \tag{EC.7}$$

where $\leq_{st}$ denotes the standard stochastic order. (Please refer to Lemma EC.2 below for a more general stochastic order result, where $X_{\infty,1}^n + X_{\infty,2}^n$ is equal in distribution to X_∞^n. The details of the coupling method are given in the proof of Lemma EC.2.) Clearly, X_∞^n is positive recurrent. Let $\tau_\infty^n(0)$ be the first hitting time of zero by X_∞^n from other states. The above stochastic order implies that $\mathbb{E}[\tau^n(\mathbf{0})] \leq \mathbb{E}[\tau_\infty^n(0)] < \infty$.

Step 2. Let M be a positive integer such that $M > \max\{\lambda^n/(n(\mu\wedge\theta)) : n \in \mathbb{N}\}$. We will prove that $\bar{Q}_i^n(\infty) \Rightarrow 0$ as $n \to \infty$ for $i > M$. As a result, if $\{\bar{\mathbf{Q}}^n(\infty) : n \in \mathbb{N}\}$ has a weak limit, it must belong to $\mathbb{S}_{M+1}$.

We introduce a sequence of auxiliary systems each having two server pools. In the nth auxiliary system, there are nM servers at the first pool and infinitely many servers at the second pool. All servers are identical. The arrival process of the nth auxiliary system is identical to that of the nth DQ–JSQ system. Upon arrival, each customer will join the first pool if there are idle servers; otherwise, the customer will join the second pool. Service times are exponentially distributed with mean $1/(\mu\wedge\theta)$ at both pools. In other words, the nth auxiliary system is an M/M/∞ system as described in Step 1, with nM servers having priority to take incoming customers.

Let $X_{\infty,1}^n(t)$ and $X_{\infty,2}^n(t)$ be the respective numbers of customers at the two server pools at time t. The next lemma establishes a stochastic order between the nth DQ–JSQ system and the nth auxiliary system. The proof is also given later in this section.

Lemma EC.2. *Assume that $\sum_{i=1}^{M} Q_i^n(0) \le_{st} X_{\infty,1}^n(0)$ and $\sum_{i=M+1}^{\infty} Q_i^n(0) \le_{st} X_{\infty,2}^n(0)$. Then under the conditions of part (ii) of Theorem 3,*

$$\left\{\Big(\sum_{i=1}^{\infty} Q_i^n(t), \sum_{i=M+1}^{\infty} Q_i^n(t)\Big) : t \ge 0\right\} \le_{st} \left\{\big(X_{\infty,1}^n(t) + X_{\infty,2}^n(t), X_{\infty,2}^n(t)\big) : t \ge 0\right\}.$$

Note that $X_{\infty,1}^n(t)$ corresponds to the number of customers in an M/M/nM/nM loss system at time t and $X_{\infty,1}^n(t) + X_{\infty,2}^n(t)$ corresponds to the number of customers in the M/M/∞ system. Both processes are positive recurrent continuous-time Markov chains. Therefore, there exists a random vector $(X_{\infty,1}^n(\infty), X_{\infty,2}^n(\infty))$ such that $(X_{\infty,1}^n(t), X_{\infty,2}^n(t)) \Rightarrow (X_{\infty,1}^n(\infty), X_{\infty,2}^n(\infty))$ as $t \to \infty$, where $(X_{\infty,1}^n(\infty), X_{\infty,2}^n(\infty))$ follows the unique steady-state distribution of $(X_{\infty,1}^n, X_{\infty,2}^n)$. By Lemma EC.2,

$$\sum_{i=M+1}^{\infty} Q_i^n(\infty) \le_{st} X_{\infty,2}^n(\infty).$$

Put $\bar{X}_{\infty,k}^n(\infty) \coloneqq X_{\infty,k}^n(\infty)/n$ for $k = 1, 2$. We next show that $\bar{X}_{\infty,2}^n(\infty) \Rightarrow 0$ as $n \to \infty$. To this end, we consider an M/M/k/ℓ + M system with both mean service time and mean patience time being $1/(\mu\wedge\theta)$. With $\ell = \infty$, this model is identical to the aforementioned M/M/∞ system. By Theorem 2.3 in Whitt (2004), $\bar{X}_{\infty,1}^n(\infty) + \bar{X}_{\infty,2}^n(\infty) \Rightarrow \lambda/(\mu\wedge\theta)$ as $n \to \infty$. With $k = \ell = nM$, this model is identical to the M/M/nM/nM loss system. Using Theorem 2.3 in Whitt (2004) again, $\bar{X}_{\infty,1}^n(\infty) \Rightarrow \lambda/(\mu\wedge\theta)$ as $n \to \infty$. These results imply that $\bar{X}_{\infty,2}^n(\infty) \Rightarrow 0$ as $n \to \infty$. Then by the above stochastic order, we obtain

$$\sum_{i=M+1}^{\infty} \bar{Q}_i^n(\infty) \Rightarrow 0 \quad \text{as } n \to \infty, \tag{EC.8}$$

so that $\bar{Q}_i^n(\infty) \Rightarrow 0$ as $n \to \infty$ for $i > M$.

Step 3. The tightness of $\{\bar{\mathbf{Q}}^n(\infty) : n \in \mathbb{N}\}$ follows from the fact that $0 \le \bar{Q}_i^n(\infty) \le 1$ for all $i \in \mathbb{N}$. With slight abuse of notation, we also use $\{\bar{\mathbf{Q}}^n(\infty) : n \in \mathbb{N}\}$ to denote a weakly convergent subsequence, i.e., $\bar{\mathbf{Q}}^n(\infty) \Rightarrow \bar{\mathbf{Q}}(\infty)$ as $n \to \infty$ for some $\mathbb{R}^\infty$-valued random vector $\bar{\mathbf{Q}}(\infty)$. It remains to prove $\bar{\mathbf{Q}}(\infty) = \mathbf{q}^*$.

Assume that all DQ–JSQ systems start with their steady states—that is, $\bar{\mathbf{Q}}^n(0)$ has the same distribution as $\bar{\mathbf{Q}}^n(\infty)$ for all $n \in \mathbb{N}$. Since $\bar{\mathbf{Q}}^n(0) \Rightarrow \bar{\mathbf{Q}}(\infty)$, we have $\bar{\mathbf{Q}}^n(t) \Rightarrow \bar{\mathbf{Q}}(\infty)$ as $n \to \infty$ for all $t \ge 0$. By (EC.8), $\bar{\mathbf{Q}}(\infty) \in \mathbb{S}_{M+1}$. It follows from part (i) of Theorem 3 that $(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) \Rightarrow (\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ as $n \to \infty$, where $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ is a fluid solution. Comparing these convergence results, we deduce that $\bar{\mathbf{Q}}(t) = \bar{\mathbf{Q}}(\infty)$ for all $t \ge 0$. Since $\bar{\mathbf{Q}}(\infty)$ is an invariant state, we must have $\bar{\mathbf{Q}}(\infty) = \mathbf{q}^*$ by Theorem 2.

Let us present the proofs of Lemmas EC.1 and EC.2 below to complete the proof of Theorem 3.

Proof of Lemma EC.1. To obtain the tightness result, it suffices to prove the tightness of $\{\bar{Q}_i^n : n \in \mathbb{N}\}$ and $\{\bar{U}_i^n : n \in \mathbb{N}\}$ for $i \in \mathbb{N}$ (see Proposition 3.2.4 in Ethier and Kurtz 1986). To this end, we define the fluid-scaled versions of some processes by

$$\bar{A}^n(t) := \bar{U}_0^n(t) = \frac{1}{n}A^n(t), \quad \bar{D}_i^n(t) := \frac{1}{n}D_i^n(t), \quad \bar{G}_i^n(t) := \frac{1}{n}G_i^n(t).$$

In addition, we write

$$\bar{S}_i^n(t) := \frac{1}{n}S_i(nt) \quad \text{and} \quad \bar{F}_i^n(t) := \frac{1}{n}F_i(nt),$$

where $\{S_i, F_i : i \in \mathbb{N}\}$ is a set of independent Poisson processes with rate one. Clearly, $\{\bar{A}^n : n \in \mathbb{N}\}$ is tight and $\{(\bar{S}_i^n, \bar{F}_i^n) : n \in \mathbb{N}\}$ is tight for each $i \in \mathbb{N}$. Since $\bar{U}_i^n(t) \le \bar{A}^n(t)$ and $0 \le \bar{U}_i^n(t) - \bar{U}_i^n(s) \le \bar{A}^n(t) - \bar{A}^n(s)$ for $0 \le s \le t$, $\{\bar{U}_i^n : n \in \mathbb{N}\}$ is tight for $i \in \mathbb{N}$. Similarly, the tightness of $\{(\bar{D}_i^n, \bar{G}_i^n) : n \in \mathbb{N}\}$ follows from (7)–(8) and the fact that $0 \le \bar{Q}_i^n(t) \le 1$ for $i \in \mathbb{N}$ and $t \ge 0$. Then, we obtain the tightness of $\{\bar{Q}_i^n : n \ge 1\}$ using these tightness results, along with the dynamical equation (9).

Now let us prove that the limit of a weakly convergent subsequence of $\{(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) : n \in \mathbb{N}\}$ is a fluid solution. With slight abuse of notation, we also use $\{(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) : n \in \mathbb{N}\}$ to denote such a subsequence, with $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ being the limit. By Skorohod's representation theorem (see, e.g., Theorem 6.7 in Billingsley 1999), we may further assume that $\{(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) : n \in \mathbb{N}\}$ and $(\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ are defined on a common probability space, with $(\bar{\mathbf{Q}}^n, \bar{\mathbf{U}}^n) \to (\bar{\mathbf{Q}}, \bar{\mathbf{U}})$ as $n \to \infty$ on every sample path. Then,

$$\left\{ \int_0^t (\bar{Q}_i^n(s) - \bar{Q}_{i+1}^n(s))\,\mathrm{d}s : t \ge 0 \right\} \to \left\{ \int_0^t (\bar{Q}_i(s) - \bar{Q}_{i+1}(s))\,\mathrm{d}s : t \ge 0 \right\} \quad \text{as } n \to \infty.$$

It follows from (7)–(9), the functional strong law of large numbers for $\{(\bar{A}^n, \bar{S}_i^n, \bar{F}_i^n) : n \in \mathbb{N}\}$, and the random time-change theorem (see Theorem 5.3 in Chen and Yao 2001) that the limit satisfies (11). It also satisfies (12)–(13) and (15)–(16) in view of (2)–(4) and (6), respectively.

It remains to verify (14). It suffices to prove that for $0 \le t_1 < t_2$, $\bar{U}_i(t_2) - \bar{U}_i(t_1) = 0$ if $\bar{Q}_i(t) < 1$ for $t_1 \le t \le t_2$. This condition implies that $Q_i^n(t) < n$ for $t_1 \le t \le t_2$ when n is sufficiently large. By (5), $U_i^n(t_2) - U_i^n(t_1) = 0$, so that $\bar{U}_i(t_2) - \bar{U}_i(t_1) = 0$. □

Proof of Lemma EC.2. We follow the approach in Dong et al. (2015) to construct $\mathbf{Q}^n$ for the DQ–JSQ system and $(X^n_{\infty,1}, X^n_{\infty,2})$ for the associated auxiliary system. We will prove that on each sample path,

$$\sum_{i=1}^{\infty} Q^n_i(t) \le X^n_{\infty,1}(t) + X^n_{\infty,2}(t) \quad \text{and} \quad \sum_{i=M+1}^{\infty} Q^n_i(t) \le X^n_{\infty,2}(t) \quad \text{for all } t \ge 0. \tag{EC.9}$$

As a result, the stochastic order in Lemma EC.2 holds even if $\mathbf{Q}^n$ and $(X^n_{\infty,1}, X^n_{\infty,2})$ are defined on different probability spaces.

By the initial condition, we may assume that (EC.9) holds at time zero. Let $\{\tau_k : k \in \mathbb{N}\}$ be the sequence of event times—that is, there is a customer either arriving at both systems or departing (by service completion or abandonment) from one of the systems at time τ_k. We take $\tau_0 \coloneqq 0$ by convention. Note that $\mathbf{Q}^n$ and $(X^n_{\infty,1}, X^n_{\infty,2})$ are continuous-time Markov chains. Assume that we have obtained the sample paths of $\mathbf{Q}^n$ and $(X^n_{\infty,1}, X^n_{\infty,2})$ up to time τ_k for some $k \in \mathbb{N}_0$. With $\mathbf{Q}^n(\tau_k) = \mathbf{q}$ and $(X^n_{\infty,1}(\tau_k), X^n_{\infty,2}(\tau_k)) = (x_1, x_2)$, we put

$$\nu(\mathbf{q}, x_1, x_2) \coloneqq \lambda^n + (b_1(\mathbf{q}) + b_2(\mathbf{q})) \vee (\mu \wedge \theta)(x_1 + x_2),$$

where $b_1(\mathbf{q}) \coloneqq \sum_{i=1}^{M} (\mu + (i-1)\theta)(q_i - q_{i+1})$ and $b_2(\mathbf{q}) \coloneqq \sum_{i=M+1}^{\infty} (\mu + (i-1)\theta)(q_i - q_{i+1})$.

Let δ_{k+1} be an exponential random variable with mean $1/\nu(\mathbf{q}, x_1, x_2)$. Then, $\tau_{k+1} \coloneqq \tau_k + \delta_{k+1}$ is the next event time. We generate a standard uniform random variable U_k that is independent of $\{\mathbf{Q}^n(u) : 0 \le u \le \tau_k\}$ and $\{(X^n_{\infty,1}(u), X^n_{\infty,2}(u)) : 0 \le u \le \tau_k\}$ to determine the event at τ_{k+1} by the following procedure:

1. If $0 \le U_k \le \lambda^n / \nu(\mathbf{q}, x_1, x_2)$, there is an arrival at both systems at τ_{k+1}. By the JSQ policy,

$$Q^n_i(\tau_{k+1}) \coloneqq \begin{cases} q_i + 1, & i = \min\{j \in \mathbb{N} : q_j < n\}, \\ q_i, & \text{otherwise}. \end{cases}$$

   In addition, $X^n_{\infty,1}(\tau_{k+1}) \coloneqq x_1 + 1$ and $X^n_{\infty,2}(\tau_{k+1}) \coloneqq x_2$ if $x_1 < nM$, and $X^n_{\infty,1}(\tau_{k+1}) = x_1$ and $X^n_{\infty,2}(\tau_{k+1}) = x_2 + 1$ if $x_1 = nM$.

2. If $(\lambda^n + \sum_{i=M+1}^{j-1} (\mu + (i-1)\theta)(q_i - q_{i+1}))/\nu(\mathbf{q}, x_1, x_2) < U_k \le (\lambda^n + \sum_{i=M+1}^{j} (\mu + (i-1)\theta)(q_i - q_{i+1}))/\nu(\mathbf{q}, x_1, x_2)$ for some $j \ge M+1$, there is a customer either completing service or abandoning the system from a server having j customers in the DQ–JSQ system. Then,

$$Q^n_i(\tau_{k+1}) = \begin{cases} q_i - 1, & i = j, \\ q_i, & \text{otherwise}. \end{cases}$$

3. If $(\lambda^n + b_2(\mathbf{q}) + \sum_{i=1}^{j-1} (\mu + (i-1)\theta)(q_i - q_{i+1}))/\nu(\mathbf{q}, x_1, x_2) < U_k \le (\lambda^n + b_2(\mathbf{q}) + \sum_{i=1}^{j} (\mu + (i-1)\theta)(q_i - q_{i+1}))/\nu(\mathbf{q}, x_1, x_2)$ for some $1 \le j \le M$, there is a customer either completing service or abandoning the system from a server having j customers in the DQ–JSQ system. Then,

$$Q^n_i(\tau_{k+1}) = \begin{cases} q_i - 1, & i = j, \\ q_i, & \text{otherwise}. \end{cases}$$

4. If $\lambda^n/\nu(\mathbf{q},x_1,x_2) < U_k \le (\lambda^n + (\mu\wedge\theta)x_2)/\nu(\mathbf{q},x_1,x_2)$, there is a service completion from the second pool at time τ_{k+1}. Then, $X^n_{\infty,1}(\tau_{k+1}) \coloneqq x_1$ and $X^n_{\infty,2}(\tau_{k+1}) \coloneqq x_2 - 1$.
5. If $(\lambda^n + (\mu\wedge\theta)x_2)/\nu(\mathbf{q},x_1,x_2) < U_k \le (\lambda^n + (\mu\wedge\theta)(x_1+x_2))/\nu(\mathbf{q},x_1,x_2)$, there is a service completion from the first pool at time τ_{k+1}. Then, $X^n_{\infty,1}(\tau_{k+1}) \coloneqq x_1 - 1$ and $X^n_{\infty,2}(\tau_{k+1}) \coloneqq x_2$.

One can verify that the process $\mathbf{Q}^n$ constructed in this way has the same generator as the augmented queue length process in the nth DQ–JSQ system has. Therefore, these two processes have the same distribution. Similarly, $(X^n_{\infty,1}, X^n_{\infty,2})$ constructed in the above way has the same distribution as the corresponding pair of processes has in the nth auxiliary system.

Suppose that (EC.9) holds at τ_k for some $k\in\mathbb{N}_0$. Now we prove that it also holds at τ_{k+1}. Then, we may complete the proof by induction.

If $0 \le U_k \le \lambda^n/\nu(\mathbf{q},x_1,x_2)$,

$$\sum_{i=1}^{\infty} Q_i^n(\tau_{k+1}) = \sum_{i=1}^{\infty} Q_i^n(\tau_k) + 1 \le X^n_{\infty,1}(\tau_k) + X^n_{\infty,2}(\tau_k) + 1 = X^n_{\infty,1}(\tau_{k+1}) + X^n_{\infty,2}(\tau_{k+1}).$$

Suppose that $\sum_{i=M+1}^{\infty} Q_i^n(\tau_{k+1}) > X^n_{\infty,2}(\tau_{k+1})$. Then, we should have $\sum_{i=M+1}^{\infty} Q_i^n(\tau_k) = X^n_{\infty,2}(\tau_k)$ and thus $\sum_{i=1}^{M} Q_i^n(\tau_k) \le X^n_{\infty,1}(\tau_k)$. The hypothesis yields $\sum_{i=M+1}^{\infty} Q_i^n(\tau_{k+1}) = \sum_{i=M+1}^{\infty} Q_i^n(\tau_k) + 1$, and thus $Q_i^n(\tau_k) = n$ for all $i \le M$ under the JSQ policy. Because $\sum_{i=1}^{M} Q_i^n(t) = nM$, we deduce that $X^n_{\infty,1}(\tau_k) = nM$. This implies that $X^n_{\infty,2}(\tau_{k+1}) = X^n_{\infty,2}(\tau_k) + 1$. On the other hand, the hypothesis also yields $X^n_{\infty,2}(\tau_{k+1}) = X^n_{\infty,2}(\tau_k)$, which is a contradiction. Therefore, $\sum_{i=M+1}^{\infty} Q_i^n(\tau_{k+1}) \le X^n_{\infty,2}(\tau_{k+1})$.

If $\lambda^n/\nu(\mathbf{q},x_1,x_2) < U_k \le 1$, we first prove that $\sum_{i=M+1}^{\infty} Q_i^n(\tau_{k+1}) \le X^n_{\infty,2}(\tau_{k+1})$. Suppose on the contrary $\sum_{i=M+1}^{\infty} Q_i^n(\tau_{k+1}) > X^n_{\infty,2}(\tau_{k+1})$. Then, $\sum_{i=M+1}^{\infty} Q_i^n(\tau_k) = X^n_{\infty,2}(\tau_k)$, i.e., $\sum_{i=M+1}^{\infty} q_i = x_2$. We should also have $\sum_{i=M+1}^{\infty} Q_i^n(\tau_{k+1}) = \sum_{i=M+1}^{\infty} Q_i^n(\tau_k)$ and $X^n_{\infty,2}(\tau_{k+1}) = X^n_{\infty,2}(\tau_k) - 1$, which implies that $(\lambda^n + b_2(\mathbf{q}))/\nu(\mathbf{q},x_1,x_2) < U_k \le (\lambda^n + (\mu\wedge\theta)x_2)/\nu(\mathbf{q},x_1,x_2)$. Hence, $b_2(\mathbf{q}) < (\mu\wedge\theta)x_2$. On the other hand, $b_2(\mathbf{q}) = \sum_{i=M+1}^{\infty}(\mu + (i-1)\theta)(q_i - q_{i+1}) \ge (\mu\wedge\theta)\sum_{i=M+1}^{\infty} q_i = (\mu\wedge\theta)x_2$, a contradiction.

Finally, let us prove that $\sum_{i=1}^{\infty} Q_i^n(\tau_{k+1}) \le X^n_{\infty,1}(\tau_{k+1}) + X^n_{\infty,2}(\tau_{k+1})$ when $\lambda^n/\nu(\mathbf{q},x_1,x_2) < U_k \le 1$. If this is not true, $\sum_{i=1}^{\infty} Q_i^n(\tau_k) = X^n_{\infty,1}(\tau_k) + X^n_{\infty,2}(\tau_k)$, i.e., $\sum_{i=1}^{\infty} q_i = x_1 + x_2$. Since $(\lambda^n + b_1(\mathbf{q}) + b_2(\mathbf{q}))/\nu(\mathbf{q},x_1,x_2) < U_k \le (\lambda^n + (\mu\wedge\theta)(x_1+x_2))/\nu(\mathbf{q},x_1,x_2)$, we should have $b_1(\mathbf{q}) + b_2(\mathbf{q}) < (\mu\wedge\theta)(x_1+x_2)$. However, $b_1(\mathbf{q}) + b_2(\mathbf{q}) = \sum_{i=1}^{\infty}(\mu + (i-1)\theta)(q_i - q_{i+1}) \ge (\mu\wedge\theta)\sum_{i=1}^{\infty} q_i = (\mu\wedge\theta)(x_1+x_2)$, a contradiction. □

EC.4. Proof of Theorem 4

Put $I^n(\infty) \coloneqq \max\{i\in\mathbb{N}_0 : Q_i^n(\infty) = n\}$, which is the minimum number of customers that a server has in the steady state. Then, W^n has the same distribution as $T_a(I^n(\infty))$. Consider the number of servers having at least i customers in the steady state. This number will increase when an incoming customer joins a server with $i-1$ customers. According to the JSQ policy, the increasing rate is $\lambda^n \cdot \mathbb{P}(I^n(\infty) = i-1)$. The number will decrease when a customer leaves a server that has exactly i customers, either by service completion or by abandonment. The decreasing rate is $(\mu + (i-1)\theta) \cdot \mathbb{E}[Q_i^n(\infty) - Q_{i+1}^n(\infty)]$. Equalizing these two rates, we obtain the following balance equations:

$$\lambda^n \cdot \mathbb{P}(I^n(\infty) = i-1) = (\mu + \theta(i-1)) \cdot \mathbb{E}[Q_i^n(\infty) - Q_{i+1}^n(\infty)] \quad \text{for } i \in \mathbb{N},$$

which implies that

$$\lim_{n\to\infty} \mathbb{P}(I^n(\infty) = i-1) = \lim_{n\to\infty} \frac{n(\mu+\theta(i-1))}{\lambda^n} \cdot \mathbb{E}[\bar{Q}_i^n(\infty) - \bar{Q}_{i+1}^n(\infty)].$$

By part (ii) of Theorem 3 and the dominated convergence theorem,

$$\lim_{n\to\infty} \mathbb{P}(I^n(\infty) = i-1) = \begin{cases} 0, & i < \bar{q}, \\ 1-p, & i = \bar{q}, \\ p, & i = \bar{q}+1, \\ 0, & i > \bar{q}+1, \end{cases}$$

from which we deduce that $W^n \Rightarrow W$ as $n \to \infty$.

EC.5. Proof of Theorem 5 with More General Results

Theorem 5 follows from Propositions EC.1, EC.2, and Corollary EC.1, all of which hold for $\rho > 1$. Proposition EC.1 summarizes performance formulas for the DQ–JSQ system with $\rho > 1$.

PROPOSITION EC.1. *Assume that condition* (21) *holds. Then, the performance of the DQ–JSQ system satisfies:*

(i) *The mean fluid-scaled number of customers in the system*

$$\lim_{n\to\infty} \mathbb{E}[\bar{X}_{\mathrm{D}}^n(\infty)] = q+1.$$

(ii) *The probability of customer abandonment*

$$\lim_{n\to\infty} P_{\mathrm{D}}^n(\mathrm{Ab}) = \frac{\rho-1}{\rho}.$$

(iii) *The mean AWT*

$$\lim_{n\to\infty} \mathbb{E}[V_{\mathrm{D}}^n] = \frac{q}{\rho\mu}.$$

(iv) *The probability of delay*

$$\lim_{n\to\infty} P_{\mathrm{D}}^n(\mathrm{De}) = \begin{cases} r(\mu+\theta)/(\rho\mu), & 1 < \rho < 1 + \theta/\mu, \\ 1, & \rho \geq 1 + \theta/\mu. \end{cases}$$

(v) *The mean PWT of delayed customers*

$$\lim_{n\to\infty}\mathbb{E}[W_{\rm D}^n|W_{\rm D}^n>0]=\begin{cases}1/\mu, & 1<\rho<1+\theta/\mu,\\ \sum_{k=0}^{\lfloor q\rfloor}1/(\mu+k\theta)-(1-r)/(\rho\mu), & \rho\geq 1+\theta/\mu.\end{cases}$$

(vi) *The mean PWT*

$$\lim_{n\to\infty}\mathbb{E}[W_{\rm D}^n]=\sum_{k=0}^{\lfloor q\rfloor}\frac{1}{\mu+k\theta}-\frac{1-r}{\rho\mu}.$$

(vii) *The mean AWT of served customers*

$$\lim_{n\to\infty}\mathbb{E}[V_{\rm D}^n|W_{\rm D}^n\leq R]=\sum_{k=1}^{\lfloor q\rfloor}\frac{1}{\mu+k\theta}+\frac{r}{\mu+\bar{q}\theta}.$$

(viii) *The mean AWT of abandoning customers*

$$\lim_{n\to\infty}\mathbb{E}[V_{\rm D}^n|W_{\rm D}^n>R]=\frac{1}{q}\Big(\sum_{k=1}^{\lfloor q\rfloor}\frac{k}{\mu+k\theta}+\frac{r\bar{q}}{\mu+\bar{q}\theta}\Big).$$

(ix) *The variance of PWTs*

$$\lim_{n\to\infty}\mathrm{Var}(W_D^n)=\sum_{k=0}^{\lfloor q\rfloor-1}\Big(\frac{1}{\mu+k\theta}\Big)^2+\Big(\sum_{k=0}^{\lfloor q\rfloor-1}\frac{1}{\mu+k\theta}\Big)^2+\frac{2r(\mu+\bar{q}\theta)}{\rho\mu(\mu+\lfloor q\rfloor\theta)}\cdot\sum_{k=0}^{\lfloor q\rfloor}\frac{1}{\mu+k\theta}-\Big(\sum_{k=0}^{\lfloor q\rfloor}\frac{1}{\mu+k\theta}-\frac{1-r}{\rho\mu}\Big)^2.$$

When $1<\rho<1+\theta/\mu$, we have $\lfloor q\rfloor=0$. Then, the results in Proposition EC.1 are reduced to those in Theorem 5. The proof of Proposition EC.1 will be given later. The next proposition provides performance formulas for the PQ system when $\rho>1$.

Proposition EC.2. *Assume that condition* (21) *holds. Then, the performance of the* M/M/n+M *system satisfies:*

(i) *The mean fluid-scaled number of customers in the system*

$$\lim_{n\to\infty}\mathbb{E}[\bar{X}_{\rm P}^n(\infty)]=q+1.$$

(ii) *The probability of customer abandonment*

$$\lim_{n\to\infty}P_{\rm P}^n(\mathrm{Ab})=\frac{\rho-1}{\rho}.$$

(iii) *The mean AWT*

$$\lim_{n\to\infty}\mathbb{E}[V_{\rm P}^n]=\frac{q}{\rho\mu}.$$

(iv) *The probability of delay*

$$\lim_{n\to\infty}P_{\rm P}^n(\mathrm{De})=1.$$

(v) *The mean PWT, the mean PWT of delayed customers, and the mean AWT of served customers*

$$\lim_{n\to\infty}\mathbb{E}[W_{\mathrm{P}}^n]=\lim_{n\to\infty}\mathbb{E}[W_{\mathrm{P}}^n|W_{\mathrm{P}}^n>0]=\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{P}}^n|W_{\mathrm{P}}^n\le R]=w,$$

where $w:=\ln(\rho)/\theta$*.*

(vi) *The mean AWT of abandoning customers*

$$\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{P}}^n|W_{\mathrm{P}}^n>R]=-\frac{\mu w}{\theta q}+\frac{1}{\theta}.$$

(vii) *The variance of PWTs*

$$\lim_{n\to\infty}\mathrm{Var}(W_{\mathrm{P}}^n)=0.$$

The proof of Proposition EC.2 will also be given later. The performance formulas in the previous two propositions allow us to obtain comparison results for $\rho>1$.

Corollary EC.1. *Assume that condition* (21) *holds. Then, the performance of the nth DQ–JSQ system and that of the* M/M/n+M *system have the following asymptotic relationships:*

$$\lim_{n\to\infty}\mathbb{E}[\bar{X}_{\mathrm{D}}^n(\infty)]=\lim_{n\to\infty}\mathbb{E}[\bar{X}_{\mathrm{P}}^n(\infty)], \tag{EC.10}$$

$$\lim_{n\to\infty}P_{\mathrm{D}}^n(\mathrm{Ab})=\lim_{n\to\infty}P_{\mathrm{P}}^n(\mathrm{Ab}), \tag{EC.11}$$

$$\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{D}}^n]=\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{P}}^n], \tag{EC.12}$$

$$\lim_{n\to\infty}P_{\mathrm{D}}^n(\mathrm{De})<\lim_{n\to\infty}P_{\mathrm{P}}^n(\mathrm{De})\quad\textit{for } 1<\rho<1+\theta/\mu, \tag{EC.13}$$

$$\lim_{n\to\infty}P_{\mathrm{D}}^n(\mathrm{De})=\lim_{n\to\infty}P_{\mathrm{P}}^n(\mathrm{De})\quad\textit{for } \rho\ge 1+\theta/\mu, \tag{EC.14}$$

$$\lim_{n\to\infty}\mathbb{E}[W_{\mathrm{D}}^n]>\lim_{n\to\infty}\mathbb{E}[W_{\mathrm{P}}^n], \tag{EC.15}$$

$$\lim_{n\to\infty}\mathbb{E}[W_{\mathrm{D}}^n|W_{\mathrm{D}}^n>0]>\lim_{n\to\infty}\mathbb{E}[W_{\mathrm{P}}^n|W_{\mathrm{P}}^n>0], \tag{EC.16}$$

$$\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{D}}^n|W_{\mathrm{D}}^n\le R]<\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{P}}^n|W_{\mathrm{P}}^n\le R], \tag{EC.17}$$

$$\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{D}}^n|W_{\mathrm{D}}^n>R]>\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{P}}^n|W_{\mathrm{P}}^n>R], \tag{EC.18}$$

$$\lim_{n\to\infty}\mathrm{Var}(W_{\mathrm{D}}^n)>\lim_{n\to\infty}\mathrm{Var}(W_{\mathrm{P}}^n). \tag{EC.19}$$

Proof. The asymptotic relationships (EC.10)–(EC.14) follow from parts (i)–(iv) of Proposition EC.1 and the corresponding parts of Proposition EC.2. Inequality (EC.15) follows from

$$\begin{aligned} w&=\frac{1}{\theta}\ln\Big(\frac{\mu+\theta q}{\mu}\Big)=\int_0^q\frac{1}{\mu+\theta x}\,\mathrm{d}x=\sum_{k=0}^{\bar{q}-2}\int_k^{k+1}\frac{1}{\mu+\theta x}\,\mathrm{d}x+\int_{\bar{q}-1}^q\frac{1}{\mu+\theta x}\,\mathrm{d}x\\ &<\sum_{k=0}^{\bar{q}-2}\frac{1}{\mu+k\theta}+\frac{r}{\mu+\theta(\bar{q}-1)}=\sum_{k=0}^{\bar{q}-1}\frac{1}{\mu+k\theta}-\frac{1-r}{\mu+\theta(\bar{q}-1)}\\ &\le\sum_{k=0}^{\lfloor q\rfloor}\frac{1}{\mu+k\theta}-\frac{1-r}{\rho\mu},\end{aligned}$$

which also implies that (EC.16) holds for $\rho \geq 1+\theta/\mu$. If $1<\rho<1+\theta/\mu$, (EC.16) follows from

$$w=\frac{1}{\theta}\ln\Big(\frac{\mu+\theta r}{\mu}\Big)=\int_0^r \frac{1}{\mu+\theta x}\,\mathrm{d}x<\frac{r}{\mu}<\frac{1}{\mu}.$$

Inequality (EC.17) follows from

$$w=\sum_{k=0}^{\bar{q}-2}\int_k^{k+1}\frac{1}{\mu+\theta x}\,\mathrm{d}x+\int_{\bar{q}-1}^{q}\frac{1}{\mu+\theta x}\,\mathrm{d}x>\sum_{k=1}^{\lfloor q\rfloor}\frac{1}{\mu+k\theta}+\frac{r}{\mu+\bar{q}\theta}.$$

Since $\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{D}}^n]=\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{P}}^n]$ and $\lim_{n\to\infty}\mathbb{P}(W_{\mathrm{D}}^n>R)=\lim_{n\to\infty}\mathbb{P}(W_{\mathrm{P}}^n>R)$, we deduce (EC.18) from (EC.17). By Theorem 4 and Lemma EC.4 (see below), $\lim_{n\to\infty}\mathrm{Var}(W_{\mathrm{D}}^n)=\mathrm{Var}(W)>0$, where $W:=\chi\cdot T_a(\bar{q})+(1-\chi)\cdot T_a(\lfloor q\rfloor)$. Then, we obtain (EC.19). □

Corollary EC.1 provides comparison results between the two queueing designs for all $\rho>1$. By using the JSQ policy, the loss of capacity utilization induced by the DQ structure will vanish as n goes large. The fluid-scaled number of customers, the probability of customer abandonment, and the mean AWT will thus be approximately equal under the two designs. Although it is strictly less than one for $1<\rho<1+\theta/\mu$ under the DQ–JSQ design, the probability of delay approaches one for $\rho\geq 1+\theta/\mu$, getting close to that under the PQ design. When n is large, both the mean PWT and the mean PWT of delayed customers are longer under the DQ–JSQ design, while the mean AWT of served customers is shorter. Since the mean AWTs are approximately equal under the two designs, the mean AWT of abandoning customers would be longer in the DQ–JSQ system. As we discussed in Section 5.2, the steady-state PWT converges in distribution to the constant w under the PQ design, so that the variance of PWTs converges to zero. By contrast, the steady-state PWT in the DQ–JSQ system converges in distribution to a random variable with a positive variance (see Theorem 4 and part (ix) of Proposition EC.1), which implies that the DQ structure is intrinsically unfair as compared with the PQ structure.

Some preliminary results are required to prove Proposition EC.1. The following lemma summarizes some properties of $T_a(i)$ for $i\in\mathbb{N}$.

Lemma EC.3. *Put $T_a(i):=\sum_{k=0}^{i-1}\xi_{i,k}$ for $i\in\mathbb{N}$, where $\{\xi_{i,k}:k=0,\dots,i-1\}$ is a sequence of independent exponential random variables with $\mathbb{E}[\xi_{i,k}]=1/(\mu+k\theta)$. Then,*

$$\mathbb{E}[T_a(i)]=\sum_{k=0}^{i-1}\frac{1}{\mu+k\theta}, \tag{EC.20}$$

$$\mathbb{E}[T_a(i)^2]=\sum_{k=0}^{i-1}\frac{1}{(\mu+k\theta)^2}+\Big(\sum_{k=0}^{i-1}\frac{1}{\mu+k\theta}\Big)^2. \tag{EC.21}$$

$$\mathbb{E}[T_a(i)\wedge R]=\frac{i}{\mu+i\theta}, \tag{EC.22}$$

$$\mathbb{P}(T_a(i)\leq R)=\frac{\mu}{\mu+i\theta}, \tag{EC.23}$$

$$\mathbb{E}[T_a(i)\cdot \mathbb{1}_{\{T_a(i)\le R\}}] = \frac{\mu}{\mu+i\theta}\cdot\sum_{k=1}^{i}\frac{1}{\mu+k\theta}, \tag{EC.24}$$

$$\mathbb{E}[R\cdot \mathbb{1}_{\{T_a(i)> R\}}] = \frac{\theta}{\mu+i\theta}\cdot\sum_{k=1}^{i}\frac{k}{\mu+k\theta}, \tag{EC.25}$$

Proof. Equations (EC.20) and (EC.21) follow from the definition of $T_a(i)$. Write $F_a(x) \coloneqq \mathbb{P}(T_a(i)\le x)$ for $x\ge 0$. Then,

$$\begin{aligned}\mathbb{E}[T_a(i)\wedge R] &= \int_0^\infty \mathbb{P}(T_a(i)\wedge R > x)\,\mathrm{d}x = \int_0^\infty \mathbb{P}(T_a(i)>x)\cdot \mathrm{e}^{-\theta x}\,\mathrm{d}x = \frac{1}{\theta}\Big(1-\int_0^\infty \mathrm{e}^{-\theta x}\,\mathrm{d}F_a(x)\Big)\\ &= \frac{1}{\theta}\left(1-\mathbb{E}[\mathrm{e}^{-\theta T_a(i)}]\right) = \frac{1}{\theta}\Big(1-\prod_{k=0}^{i-1}\frac{\mu+k\theta}{\theta+\mu+k\theta}\Big)\\ &= \frac{i}{\mu+i\theta},\end{aligned}$$

where the third equality follows from integration by parts and the fifth equality follows from (22). By Fubini's theorem,

$$\mathbb{P}(T_a(i)\le R) = \int_0^\infty \theta\mathrm{e}^{-\theta x}\cdot F_a(x)\,\mathrm{d}x = \int_0^\infty \mathrm{e}^{-\theta y}\,\mathrm{d}F_a(y) = \mathbb{E}\left[\mathrm{e}^{-\theta T_a(i)}\right] = \prod_{k=0}^{i-1}\frac{\mu+k\theta}{\mu+(k+1)\theta} = \frac{\mu}{\mu+i\theta}.$$

Using Fubini's theorem again, we obtain

$$\begin{aligned}\mathbb{E}[T_a(i)\cdot \mathbb{1}_{\{T_a(i)\le R\}}] &= \int_0^\infty \theta\mathrm{e}^{-\theta x}\int_0^x y\,\mathrm{d}F_a(y)\,\mathrm{d}x = \int_0^\infty y\cdot\mathrm{e}^{-\theta y}\,\mathrm{d}F_a(y) = \mathbb{E}[T_a(i)\cdot \mathrm{e}^{-\theta T_a(i)}]\\ &= \prod_{k=0}^{i-1}\frac{\mu+k\theta}{\mu+(k+1)\theta}\cdot\sum_{k=0}^{i-1}\frac{1}{\mu+(k+1)\theta}\\ &= \frac{\mu}{\mu+i\theta}\cdot\sum_{k=1}^{i}\frac{1}{\mu+k\theta},\end{aligned}$$

where the fourth equality follows from

$$\mathbb{E}\left[T_a(i)\cdot\mathrm{e}^{-sT_a(i)}\right] = -\mathbb{E}\Big[\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{e}^{-sT_a(i)}\Big] = -\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}[\mathrm{e}^{-sT_a(i)}] = \prod_{k=0}^{i-1}\frac{\mu+k\theta}{s+\mu+k\theta}\cdot\sum_{k=0}^{i-1}\frac{1}{s+\mu+k\theta}.$$

Equation (EC.25) follows from

$$\begin{aligned}\mathbb{E}[R\cdot \mathbb{1}_{\{T_a(i)>R\}}] &= \int_0^\infty\int_0^x \theta\mathrm{e}^{-\theta y}\cdot y\,\mathrm{d}y\,\mathrm{d}F_a(x) = -\mathbb{E}[T_a(i)\cdot\mathrm{e}^{-\theta T_a(i)}] + \frac{1}{\theta}\big(1-\mathbb{E}[\mathrm{e}^{-\theta T_a(i)}]\big)\\ &= \frac{\theta}{\mu+i\theta}\cdot\sum_{k=1}^{i}\frac{k}{\mu+k\theta}.\end{aligned}$$

□

Then, we prove some uniform integrability results for the sequence of DQ–JSQ systems.

Lemma EC.4. *Assume that condition* (21) *holds. Then,* $\{\bar{X}_{\mathrm{D}}^n(\infty): n\in\mathbb{N}\}$, $\{W_{\mathrm{D}}^n : n\in\mathbb{N}\}$, *and* $\{(W_{\mathrm{D}}^n)^2 : n\in\mathbb{N}\}$ *are all uniformly integrable.*

Proof. Consider the M/M/∞ system that has arrival rate λ^n and mean service time $1/(\mu\wedge\theta)$. Let $X_\infty^n(\infty)$ be the steady-state number of customers in this system and $\bar{X}_\infty^n(\infty):=X_\infty^n(\infty)/n$. By (EC.7), $\bar{X}_{\mathrm{D}}^n(\infty)\le_{st}\bar{X}_\infty^n(\infty)$. Hence, it suffices to show that $\{\bar{X}_\infty^n(\infty):n\in\mathbb{N}\}$ is uniformly integrable. Note that $X_\infty^n(\infty)$ is a Poisson random variable with mean $\lambda^n/(\mu\wedge\theta)$, so that

$$\sup_{n\in\mathbb{N}}\mathbb{E}[\bar{X}_\infty^n(\infty)^2]=\sup_{n\in\mathbb{N}}\Big\{\Big(\frac{\lambda^n}{n(\mu\wedge\theta)}\Big)^2+\frac{\lambda^n}{n^2(\mu\wedge\theta)}\Big\}<\infty.$$

By Proposition A.2.2 in Ethier and Kurtz (1986), $\{\bar{X}_\infty^n(\infty):n\in\mathbb{N}\}$ is uniformly integrable.

Then, let us consider $\{W_{\mathrm{D}}^n:n\in\mathbb{N}\}$. Note that W_{D}^n has the same distribution as $T_a(I^n(\infty))$ where $I^n(\infty):=\max\{i\in\mathbb{N}_0:Q_i^n(\infty)=n\}$ is the minimum number of customers that a server has in the steady state. Taking $s=-\zeta$ with $0<\zeta<\mu\wedge\theta$ in (22) yields

$$\mathbb{E}[\mathrm{e}^{\zeta W_{\mathrm{D}}^n}]=\mathbb{E}\left[\prod_{i=0}^{I^n(\infty)-1}\frac{\mu+i\theta}{\mu+i\theta-\zeta}\right]<\frac{\mu+\theta\mathbb{E}[I^n(\infty)]}{\mu-\zeta}.$$

Since $nI^n(\infty)\le X_{\mathrm{D}}^n(\infty)$,

$$\mathbb{E}[\mathrm{e}^{\zeta W_{\mathrm{D}}^n}]\le\frac{\mu+\theta\mathbb{E}[\bar{X}_{\mathrm{D}}^n(\infty)]}{\mu-\zeta}\le\frac{\mu+\theta\mathbb{E}[\bar{X}_\infty^n(\infty)]}{\mu-\zeta}=\frac{1}{\mu-\zeta}\cdot\Big(\mu+\frac{\theta\lambda^n}{n(\mu\wedge\theta)}\Big),$$

from which we deduce that $\sup_{n\in\mathbb{N}}\mathbb{E}[\mathrm{e}^{\zeta W_{\mathrm{D}}^n}]<\infty$. By Proposition A.2.2 in Ethier and Kurtz (1986), both $\{W_{\mathrm{D}}^n:n\in\mathbb{N}\}$ and $\{(W_{\mathrm{D}}^n)^2:n\in\mathbb{N}\}$ are uniformly integrable. □

Now let us present the proof of Proposition EC.1.

Proof of Proposition EC.1. (i) It follows from (10), part (ii) of Theorem 3, and the fact that $\{\bar{X}_{\mathrm{D}}^n(\infty):n\in\mathbb{N}\}$ is uniformly integrable.

(ii) By Theorem 4,

$$\lim_{n\to\infty}P_{\mathrm{D}}^n(\mathrm{Ab})=\lim_{n\to\infty}\mathbb{P}(W_{\mathrm{D}}^n>R)=p\cdot\mathbb{P}(T_a(\bar{q})>R)+(1-p)\cdot\mathbb{P}(T_a(\lfloor q\rfloor)>R).$$

Then, the formula follows from (EC.23).

(iii) Since $V_{\mathrm{D}}^n:=W_{\mathrm{D}}^n\wedge R$ and $\{W_{\mathrm{D}}^n:n\in\mathbb{N}\}$ is uniformly integrable, $\{V_{\mathrm{D}}^n:n\in\mathbb{N}\}$ is also uniformly integrable. Then by Theorem 4,

$$\lim_{n\to\infty}\mathbb{E}[V_{\mathrm{D}}^n]=\lim_{n\to\infty}\mathbb{E}[W_{\mathrm{D}}^n\wedge R]=p\cdot\mathbb{E}[T_a(\bar{q})\wedge R]+(1-p)\cdot\mathbb{E}[T_a(\lfloor q\rfloor)\wedge R].$$

The formula follows from (EC.22).

(iv) If $1<\rho<1+\theta/\mu$, we have $0<q<1$ and by Theorem 4,

$$\lim_{n\to\infty}P_{\mathrm{D}}^n(\mathrm{De})=\lim_{n\to\infty}\mathbb{P}(W_{\mathrm{D}}^n>0)=p=\frac{r(\mu+\theta)}{\rho\mu}.$$

If $\rho\ge1+\theta/\mu$, we have $q\ge1$, so that

$$\lim_{n\to\infty}P_{\mathrm{D}}^n(\mathrm{De})=\lim_{n\to\infty}\mathbb{P}(W_{\mathrm{D}}^n>0)=p\cdot\mathbb{P}(T_a(\bar{q})>0)+(1-p)\cdot\mathbb{P}(T_a(\lfloor q\rfloor)>0)=1.$$

(v) Since $\{W_{\mathrm{D}}^n : n \in \mathbb{N}\}$ is uniformly integrable, so is $\{W_{\mathrm{D}}^n \cdot \mathbb{1}_{\{W_{\mathrm{D}}^n > 0\}} : n \in \mathbb{N}\}$. If $1 < \rho < 1 + \theta/\mu$, we have $0 < q < 1$ and by Theorem 4,

$$\lim_{n\to\infty} \mathbb{E}[W_{\mathrm{D}}^n | W_{\mathrm{D}}^n > 0] = \mathbb{E}[T_a(1)] = \frac{1}{\mu}.$$

If $\rho \geq 1 + \theta/\mu$, $\lim_{n\to\infty} \mathbb{P}(W_{\mathrm{D}}^n > 0) = 1$ by part (iv). Hence, $\lim_{n\to\infty} \mathbb{E}[W_{\mathrm{D}}^n | W_{\mathrm{D}}^n > 0] = \lim_{n\to\infty} \mathbb{E}[W_{\mathrm{D}}^n]$ (please refer to the proof of part (vi) below).

(vi) By Theorem 4 and the fact that $\{W_{\mathrm{D}}^n : n \in \mathbb{N}\}$ is uniformly integrable,

$$\lim_{n\to\infty} \mathbb{E}[W_{\mathrm{D}}^n] = p \cdot \mathbb{E}[T_a(\bar{q})] + (1-p) \cdot \mathbb{E}[T_a(\lfloor q \rfloor)].$$

Then, the formula follows from (EC.20).

(vii) Since $\{W_{\mathrm{D}}^n : n \in \mathbb{N}\}$ is uniformly integrable, so is $\{W_{\mathrm{D}}^n \cdot \mathbb{1}_{\{W_{\mathrm{D}}^n} \leq R\} : n \in \mathbb{N}\}$. By part (ii) of this proposition, $\lim_{n\to\infty} \mathbb{P}(W_{\mathrm{D}}^n \leq R) = 1/\rho$. Then by Theorem 4,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[V_{\mathrm{D}}^n | W_{\mathrm{D}}^n \leq R] &= \lim_{n\to\infty} \mathbb{E}[W_{\mathrm{D}}^n | W_{\mathrm{D}}^n \leq R] \\ &= \rho \cdot \big(p \cdot \mathbb{E}[T_a(\bar{q}) \cdot \mathbb{1}_{\{T_a(\bar{q}) \leq R\}}] + (1-p) \cdot \mathbb{E}[T_a(\lfloor q \rfloor) \cdot \mathbb{1}_{\{T_a(\lfloor q \rfloor) \leq R\}}]\big).\end{aligned}$$

The formula follows from (EC.24).

(viii) Note that $\{R \cdot \mathbb{1}_{\{W_{\mathrm{D}}^n > R\}} : n \in \mathbb{N}\}$ is uniformly integrable. By Theorem 4 and part (ii) of the present proposition,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[V_{\mathrm{D}}^n | W_{\mathrm{D}}^n > R] &= \lim_{n\to\infty} \mathbb{E}[R | W_{\mathrm{D}}^n > R] \\ &= \frac{\rho}{\rho - 1} \cdot \big(p \cdot \mathbb{E}[R \cdot \mathbb{1}_{\{T_a(\bar{q}) > R\}}] + (1-p) \cdot \mathbb{E}[R \cdot \mathbb{1}_{\{T_a(\lfloor q \rfloor) > R\}}]\big).\end{aligned}$$

Then, the formula follows from (EC.25).

(ix) By Theorem 4 and the uniform integrability of $\{(W_{\mathrm{D}}^n)^2 : n \in \mathbb{N}\}$,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[(W_{\mathrm{D}}^n)^2] &= p \cdot \mathbb{E}[T_a(\bar{q})^2] + (1-p) \cdot \mathbb{E}[T_a(\lfloor q \rfloor)^2] \\ &= \sum_{k=0}^{\lfloor q \rfloor - 1} \Big(\frac{1}{\mu + k\theta}\Big)^2 + \Big(\sum_{k=0}^{\lfloor q \rfloor - 1} \frac{1}{\mu + k\theta}\Big)^2 + \frac{2r(\mu + \theta\bar{q})}{\rho\mu(\mu + \theta\lfloor q \rfloor)} \sum_{k=0}^{\lfloor q \rfloor} \frac{1}{\mu + k\theta}.\end{aligned}$$

Then, we obtain $\lim_{n\to\infty} \mathrm{Var}(W_{\mathrm{D}}^n)$ using part (vi). □

To prove Proposition EC.2, we also require some uniform integrability results.

Lemma EC.5. *Assume that condition* (21) *holds. Then,* $\{\bar{X}_{\mathrm{P}}^n(\infty) : n \in \mathbb{N}\}$, $\{W_{\mathrm{P}}^n : n \in \mathbb{N}\}$, *and* $\{(W_{\mathrm{P}}^n)^2 : n \in \mathbb{N}\}$ *are all uniformly integrable.*

Proof. Using the fact that $\bar{X}^n_{\mathrm{P}}(\infty) \leq_{st} \bar{X}^n_{\infty}(\infty)$, we may follow a similar argument as in the proof of Lemma EC.4 to show the uniform integrability of $\{\bar{X}^n_{\mathrm{P}}(\infty) : n \in \mathbb{N}\}$.

Put $T^n_{\mathrm{P}}(0) := 0$ and $T^n_{\mathrm{P}}(i) := \sum_{k=1}^{i} \eta_k$ for $i \in \mathbb{N}$, where $\{\eta_k : k \in \mathbb{N}\}$ is a sequence of independent exponential random variables with $\mathbb{E}[\eta_k] = 1/(n\mu + (k-1)\theta)$. Then, W^n_{P} has the same distribution as $T^n_{\mathrm{P}}((X^n_{\mathrm{P}}(\infty) - n + 1)^+)$. For $i \in \mathbb{N}$, the Laplace transform of $T^n_{\mathrm{P}}(i)$ is

$$\mathbb{E}[\mathrm{e}^{-sT^n_{\mathrm{P}}(i)}] = \prod_{k=1}^{i} \frac{n\mu + (k-1)\theta}{s + n\mu + (k-1)\theta}.$$

Taking $s = -\zeta$ with $0 < \zeta < \mu \wedge \theta$ in this equation yields

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\zeta W^n_{\mathrm{P}}}] &= \mathbb{E}\Bigg[\prod_{k=1}^{(X^n_{\mathrm{P}}(\infty)-n+1)^+} \frac{n\mu + (k-1)\theta}{n\mu + (k-1)\theta - \zeta}\Bigg] < \frac{n\mu + \theta\mathbb{E}[X^n_{\mathrm{P}}(\infty)]}{n\mu - \zeta} \leq \frac{n\mu + \theta\mathbb{E}[X^n_{\infty}(\infty)]}{n\mu - \zeta} \\ &= \frac{1}{n\mu - \zeta} \cdot \Big(n\mu + \frac{\theta\lambda^n}{\mu \wedge \theta}\Big),\end{aligned}$$

from which we deduce that $\sup_{n\in\mathbb{N}} \mathbb{E}[\mathrm{e}^{\zeta W^n_{\mathrm{P}}}] < \infty$. By Proposition A.2.2 in Ethier and Kurtz (1986), both $\{W^n_{\mathrm{P}} : n \in \mathbb{N}\}$ and $\{(W^n_{\mathrm{P}})^2 : n \in \mathbb{N}\}$ are uniformly integrable. □

The proof of Proposition EC.2 is given below.

Proof of Proposition EC.2. Parts (i)–(v) and (vii) of the proposition follow from Theorem 2.3 in Whitt (2004) along with related uniform integrability results in Lemma EC.5. Part (vi) follows from the fact that $\lim_{n\to\infty} \mathbb{E}[V^n_{\mathrm{P}} | W^n_{\mathrm{P}} > R] = \mathbb{E}[R | R < w]$. □

EC.6. Proof of Theorem 6

By (22), $\lim_{i\to\infty} \mathbb{E}[\mathrm{e}^{-\theta T_a(i)}] = 0$, which implies that $\lim_{i\to\infty} T_a(i) = \infty$ almost surely. Since $T_a(0) = 0$ and $T_a(i)$ is stochastically strictly increasing in i, $\kappa(T, \alpha)$ is well defined for $T \geq 0$ and $0 < \alpha < 1$. Let us fix T and α in the rest of the proof.

We define a function $g : [0, \infty) \to \mathbb{R}$ by

$$g(x) := \frac{(\mu + \theta\underline{x})(\bar{x} - x)}{\mu + \theta x} \cdot \mathbb{P}(T_a(\underline{x}) > T) + \frac{(\mu + \theta\bar{x})(x - \underline{x})}{\mu + \theta x} \cdot \mathbb{P}(T_a(\bar{x}) > T),$$

where $\underline{x} := \lfloor x \rfloor$ and $\bar{x} := \underline{x} + 1$. Note that $g(q) = \mathbb{P}(W > T)$ where W is the steady-state PWT in Theorem 4. Clearly, g is continuous on $[n, n+1]$ for each $n \in \mathbb{N}_0$, thus continuous on $[0, \infty)$.

Now let us show that g is strictly increasing. We may write g as

$$g(x) = \mathbb{P}(T_a(\underline{x}) > T) + \frac{(\mu + \theta\bar{x})(x - \underline{x})}{\mu + \theta x} \cdot \big(\mathbb{P}(T_a(\bar{x}) > T) - \mathbb{P}(T_a(\underline{x}) > T)\big). \qquad \text{(EC.26)}$$

For $0 \leq x_1 < x_2$ with $\underline{x}_1 = \underline{x}_2$, we have $g(x_1) < g(x_2)$ because $(\mu + \theta\bar{x})(x - \underline{x})/(\mu + \theta x)$ is strictly increasing in x. If $\underline{x}_1 < \underline{x}_2$, we have $g(x_1) < \mathbb{P}(T_a(\bar{x}_1) > T) \leq \mathbb{P}(T_a(\underline{x}_2) > T) \leq g(x_2)$.

Since $g(0)=0$ and $\lim_{x\to\infty} g(x)=1$, there is a unique solution to $g(x)=\alpha$. Write $\hat{q}:=g^{-1}(\alpha)$. By (EC.26),

$$\hat{q}=\kappa(T,\alpha)+\frac{(\mu+\theta\kappa(T,\alpha))(1-r_0(T,\alpha))}{\mu+\theta(\kappa(T,\alpha)+r_0(T,\alpha))}.$$

Put $\hat{\lambda}:=\mu+\theta\hat{q}$. Then,

$$\hat{\lambda}=\frac{(\mu+\theta\kappa(T,\alpha))(\mu+\theta(\kappa(T,\alpha)+1))}{\mu+\theta(\kappa(T,\alpha)+r_0(T,\alpha))}.$$

For notational convenience, we write $\hat{n}(\lambda)$ for $\hat{n}_{\mathrm{D}}(\lambda,T,\alpha)$. We next prove $\lim_{\lambda\to\infty}\hat{n}(\lambda)/\lambda=1/\hat{\lambda}$. If this is not true, we may find a sequence of arrival rates $\{\lambda_k : k\in\mathbb{N}\}$ such that either $\hat{n}(\lambda_k)/\lambda_k > 1/(\hat{\lambda}-\varepsilon)$ or $\hat{n}(\lambda_k)/\lambda_k < 1/(\hat{\lambda}+\varepsilon)$ for some $\varepsilon\in(0,\theta\hat{q})$ and all $k\in\mathbb{N}$.

If $\hat{n}(\lambda_k)/\lambda_k > 1/(\hat{\lambda}-\varepsilon)$ for $k\in\mathbb{N}$, let us consider $\check{n}_1(\lambda):=\lfloor\lambda/(\hat{\lambda}-\varepsilon)\rfloor$. By Theorem 4,

$$\lim_{\lambda\to\infty}\mathbb{P}(W_{\lambda}^{\check{n}_1(\lambda)}>T)=g\Big(\frac{\hat{\lambda}-\varepsilon-\mu}{\theta}\Big)<g(\hat{q})=\alpha.$$

Then by (24), we should have $\hat{n}(\lambda_k)\le\check{n}_1(\lambda_k)$ for λ_k sufficiently large, which contradicts the fact that $\check{n}_1(\lambda_k)/\lambda_k\le 1/(\hat{\lambda}-\varepsilon)$.

If $\hat{n}(\lambda_k)/\lambda_k < 1/(\hat{\lambda}+\varepsilon)$ for $k\in\mathbb{N}$, we can find a further subsequence $\{\lambda_{k_j} : j\in\mathbb{N}\}$ and a constant $\hat{\lambda}_2\ge\hat{\lambda}+\varepsilon>\hat{\lambda}$ such that $\lim_{j\to\infty}\lambda_{k_j}/\hat{n}(\lambda_{k_j})=\hat{\lambda}_2$ and

$$\mathbb{P}(W_{\lambda_{k_j}}^{\hat{n}(\lambda_{k_j})}>T)\le\alpha \text{ for all } j\in\mathbb{N}. \tag{EC.27}$$

Using Theorem 4 again, we obtain

$$\lim_{j\to\infty}\mathbb{P}(W_{\lambda_{k_j}}^{\hat{n}(\lambda_{k_j})}>T)=g\Big(\frac{\hat{\lambda}_2-\mu}{\theta}\Big)>g\Big(\frac{\hat{\lambda}-\mu}{\theta}\Big)=g(\hat{q})=\alpha,$$

which contradicts (EC.27).

Equation (26) follows from the fact that $\kappa(0,\alpha)=0$ and $r_0(0,\alpha)=1-\alpha$.

EC.7. Proof of Theorem 7

Since $m(0)=0$, $\psi(0)=1$, and $T_a(0)=0$, we obtain $\hat{\alpha}(0)=0$. Put

$$u_k:=\frac{1}{\theta}\ln\Big(1+\frac{k\theta}{\mu}\Big)\quad\text{for } k\in\mathbb{N}_0. \tag{EC.28}$$

Then, $m(T)=k$ for $T\in[u_k,u_{k+1})$. Clearly, $\hat{\alpha}$ is continuous on $[u_k,u_{k+1})$. The continuity of $\hat{\alpha}$ on $[0,\infty)$ follows from the fact that $\hat{\alpha}(u_{k+1}-)=\hat{\alpha}(u_{k+1})=\mathbb{P}(T_a(k+1)>u_{k+1})$.

By (23), we may prove the next lemma, the proof of which is given later in this section. The monotonicity of $\hat{\alpha}$ on $[0,\infty)$ follows from this lemma along with the continuity of $\hat{\alpha}$.

Lemma EC.6. *The function $\hat{\alpha}$ defined by (29) satisfies $\hat{\alpha}'(T)>0$ for $T\in(u_k,u_{k+1})$ and $k\in\mathbb{N}_0$.*

For notational convenience, let us write $\varsigma := \mu/\theta$. We use the following lemma to prove $\hat{\alpha}(\infty) = \gamma(\varsigma,\varsigma)/\Gamma(\varsigma)$. The proof will also be given later in this section.

Lemma EC.7. *The tail probability $\mathbb{P}(T_a(k) > u_k)$ has the limit*

$$\lim_{k\to\infty} \mathbb{P}(T_a(k) > u_k) = \frac{\gamma(\varsigma,\varsigma)}{\Gamma(\varsigma)}.$$

Because $\hat{\alpha}$ is strictly increasing on $[0,\infty)$, $\hat{\alpha}(u_k) \le \hat{\alpha}(T) < \hat{\alpha}(u_{k+1})$ for $T \in [u_k, u_{k+1})$ and $k \in \mathbb{N}_0$. By (29), $\hat{\alpha}(u_k) = \mathbb{P}(T_a(i) > u_k)$. Then, it follows from Lemma EC.7 that $\hat{\alpha}(\infty) = \gamma(\varsigma,\varsigma)/\Gamma(\varsigma)$.

Next we prove that for a fixed $T \ge 0$, $\lim_{\lambda\to\infty} \hat{n}_{\mathrm{D}}(\lambda, T, \alpha)/\hat{n}_{\mathrm{P}}(\lambda, T, \alpha) \le 1$ if and only if $\alpha \ge \hat{\alpha}(T)$. Write $\hat{p}_k := \mathbb{P}(T_a(k) > T)$ for $k \in \mathbb{N}_0$ and

$$\phi(T,\alpha) := \frac{\mu + \theta(\kappa(T,\alpha) + r_0(T,\alpha))}{(\mu + \theta\kappa(T,\alpha))(\mu + \theta(\kappa(T,\alpha)+1))}. \tag{EC.29}$$

Since $\kappa(T,\alpha) = k$ for $\alpha \in [\hat{p}_k, \hat{p}_{k+1})$, $\phi(T,\alpha)$ is continuous and strictly decreasing in α on this interval. Then because $\phi(T, \hat{p}_{k+1}-) = \phi(T, \hat{p}_{k+1}) = 1/(\mu + \theta(k+1))$, $\phi(T,\alpha)$ is continuous and strictly decreasing in α on $[0,1)$. Note that $\phi(T,0) = 1/\mu$ and $\lim_{\alpha\uparrow 1} \phi(T,\alpha) = 0$, so that there is a unique $\check{\alpha}(T) \in [0,1)$ such that $\phi(T, \check{\alpha}(T)) = \mathrm{e}^{-\theta T}/\mu$. Moreover, $\phi(T,\alpha) \le \mathrm{e}^{-\theta T}/\mu$ if and only if $\alpha \ge \check{\alpha}(T)$. Then by (25) and (27), $\lim_{\lambda\to\infty} \hat{n}_{\mathrm{D}}(\lambda,T,\alpha)/\hat{n}_{\mathrm{P}}(\lambda,T,\alpha) \le 1$ if and only if $\alpha \ge \check{\alpha}(T)$.

Let us verify $\check{\alpha}(T) = \hat{\alpha}(T)$ for $T \ge 0$. Since $0 < r_0(T,\alpha) \le 1$, we obtain the following inequalities:

$$\frac{1}{\mu + \theta(\kappa(T,\alpha)+1)} < \phi(T,\alpha) \le \frac{1}{\mu + \theta\kappa(T,\alpha)}.$$

Because $\phi(T, \check{\alpha}(T)) = \mathrm{e}^{-\theta T}/\mu$,

$$\frac{1}{\mu + \theta(\kappa(T,\check{\alpha}(T))+1)} < \frac{\mathrm{e}^{-\theta T}}{\mu} \le \frac{1}{\mu + \theta\kappa(T,\check{\alpha}(T))},$$

which yields $\kappa(T,\check{\alpha}(T)) = \lfloor \mu(\mathrm{e}^{\theta T} - 1)/\theta \rfloor = m(T)$. Then using (EC.29), we have

$$r_0(T,\check{\alpha}(T)) = \frac{1}{\theta}\big(\mu + \theta m(T)\big)\big(\mu + \theta(m(T)+1)\big)\Big(\frac{\mathrm{e}^{-\theta T}}{\mu} - \frac{1}{\mu + \theta(m(T)+1)}\Big) = \psi(T).$$

On the other hand, the definition of r_0 leads to

$$r_0(T,\check{\alpha}(T)) = \frac{\mathbb{P}\big(T_a(m(T)+1) > T\big) - \check{\alpha}(T)}{\mathbb{P}\big(T_a(m(T)+1) > T\big) - \mathbb{P}\big(T_a(m(T)) > T\big)}.$$

Combining the above two equations, we obtain $\check{\alpha}(T) = \hat{\alpha}(T)$.

Now we present the proofs of Lemmas EC.6 and EC.7.

Proof of Lemma EC.6. By (23) and (29), through algebraic manipulation we obtain

$$\hat{\alpha}'(T) = \psi(T) \prod_{i=1}^{k} (\mu + (i-1)\theta) \sum_{j=1}^{k+1} \mathrm{e}^{-(\mu+\theta(j-1))T} (\mu + j\theta) \prod_{i=1, i\neq j}^{k+1} \frac{1}{(i-j)\theta}.$$

Then, it suffices to prove

$$H_k(T) := \sum_{j=1}^{k+1} \mathrm{e}^{-\theta(j-1)T}(\mu + j\theta) \prod_{i=1, i\neq j}^{k+1} \frac{k!}{(i-j)} > 0.$$

The expression of $H_k(T)$ can be simplified as

$$\begin{aligned} H_k(T) &= \sum_{j=0}^{k} \mathrm{e}^{-j\theta T}(\mu + (j+1)\theta)(-1)^j \binom{k}{j} \\ &= (\mu+\theta)\sum_{j=0}^{k} \mathrm{e}^{-j\theta T}(-1)^j \binom{k}{j} + k\theta \sum_{j=1}^{k} \mathrm{e}^{-j\theta T}(-1)^j \binom{k-1}{j-1} \\ &= (\mu+\theta)(1-\mathrm{e}^{-\theta T})^k - k\theta \mathrm{e}^{-\theta T}(1-\mathrm{e}^{-\theta T})^{k-1} \\ &= (1-\mathrm{e}^{-\theta T})^{k-1}\big(\mu + \theta - (\mu + (k+1)\theta)\mathrm{e}^{-\theta T}\big), \end{aligned}$$

where the third equality follows from the binomial theorem. Since $T > u_k$, we have

$$\mathrm{e}^{\theta T} > \frac{\mu + k\theta}{\mu} > \frac{\mu + (k+1)\theta}{\mu + \theta},$$

from which we deduce that $H_k(T) > 0$. □

Proof of Lemma EC.7. Write

$$G_k(s) := \sum_{j=1}^{k} \frac{s^{\varsigma+j-1}}{\varsigma + j - 1} \cdot (-1)^{j-1} \binom{k-1}{j-1} \quad \text{for } k \in \mathbb{N} \text{ and } s \geq 0.$$

Clearly, $G_k(0) = 0$. By the binomial theorem,

$$G_k'(s) = \sum_{j=1}^{k} s^{\varsigma+j-2}(-1)^{j-1}\binom{k-1}{j-1} = s^{\varsigma-1}\sum_{j=0}^{k-1}(-s)^j \binom{k-1}{j} = s^{\varsigma-1}(1-s)^{k-1},$$

from which we obtain $G_k(s) = \int_0^s t^{\varsigma-1}(1-t)^{k-1}\,\mathrm{d}t$. Then by (EC.28) and (23),

$$\begin{aligned} \mathbb{P}(T_a(k) > u_k) &= \sum_{j=1}^{k}\left(1 + \frac{k}{\varsigma}\right)^{-(\varsigma+j-1)} \prod_{i=1, i\neq j}^{k} \frac{\varsigma + i - 1}{i - j} \\ &= \frac{\prod_{i=1}^k (\varsigma + i - 1)}{(k-1)!} \sum_{j=1}^{k} \frac{(1+k/\varsigma)^{-(\varsigma+j-1)}}{\varsigma + j - 1} \cdot (-1)^{j-1}\binom{k-1}{j-1} \\ &= \frac{\prod_{i=1}^k (\varsigma + i - 1)}{(k-1)!} \int_0^{(1+k/\varsigma)^{-1}} t^{\varsigma-1}(1-t)^{k-1}\,\mathrm{d}t. \end{aligned} \tag{EC.30}$$

Because $\Gamma(k) = (k-1)!$ for $k \in \mathbb{N}$ and $\Gamma(z) = z \cdot \Gamma(z)$ for $z > 0$,

$$\frac{\prod_{i=1}^k (\varsigma + i - 1)}{k^\varsigma (k-1)!} = \frac{\Gamma(\varsigma + k)}{k^\varsigma \cdot \Gamma(\varsigma)\Gamma(k)}.$$

Then by Stirling's formula,

$$\lim_{k\to\infty}\frac{\prod_{i=1}^{k}(\varsigma+i-1)}{k^{\varsigma}(k-1)!}=\frac{1}{\Gamma(\varsigma)}\cdot\lim_{k\to\infty}\frac{\sqrt{2\pi(\varsigma+k-1)}\Big(\dfrac{\varsigma+k-1}{\mathrm{e}}\Big)^{\varsigma+k-1}}{k^{\varsigma}\sqrt{2\pi(k-1)}\Big(\dfrac{k-1}{\mathrm{e}}\Big)^{k-1}}=\frac{1}{\Gamma(\varsigma)}. \tag{EC.31}$$

Write $\ell:=\varsigma+k$. Then,

$$\int_0^{(1+k/\varsigma)^{-1}}t^{\varsigma-1}(1-t)^{k-1}\,\mathrm{d}t=\int_0^{\varsigma/\ell}t^{\varsigma-1}(1-t)^{\ell-\varsigma-1}\,\mathrm{d}t=\frac{1}{\ell^{\varsigma}}\int_0^{\varsigma}u^{\varsigma-1}\Big(1-\frac{u}{\ell}\Big)^{\ell-\varsigma-1}\,\mathrm{d}u,$$

and we obtain

$$\lim_{k\to\infty}k^{\varsigma}\int_0^{(1+k/\varsigma)^{-1}}t^{\varsigma-1}(1-t)^{k-1}\,\mathrm{d}t=\lim_{\ell\to\infty}\frac{(\ell-\varsigma)^{\varsigma}}{\ell^{\varsigma}}\int_0^{\varsigma}u^{\varsigma-1}\Big(1-\frac{u}{\ell}\Big)^{\ell-\varsigma-1}\,\mathrm{d}u=\gamma(\varsigma,\varsigma). \tag{EC.32}$$

The assertion of the lemma follows from (EC.30)–(EC.32). □

EC.8. Proof of Proposition 1

Following the argument in Theorem 2, we obtain $\bar{Q}_i(t)=1$ for $1\le i\le\bar{q}$ and $\bar{Q}_i(t)=0$ for $i\ge\bar{q}+2$. Then, $\bar{U}_i(t)=0$ for $i\ge\bar{q}+2$ and $t\ge 0$. By (11) and the fact that $\bar{Q}'_{\bar{q}+2}(t)=0$, we obtain $\bar{U}'_{\bar{q}+1}(t)=0$ and thus $\bar{U}_{\bar{q}+1}(t)=0$ for $t\ge 0$. By induction from $i=1$ to $\bar{q}-1$, we obtain $\bar{U}'_i(t)=\rho\mu$ for $1\le i\le\bar{q}-1$ using (11) and the fact that $\bar{Q}'_i(t)=0$. Hence, $\bar{U}_i(t)=\rho\mu t$ for $1\le i\le\bar{q}-1$ and $t\ge 0$.

Taking $i=\bar{q}$ and $\bar{q}+1$ in (11), we have the following two equations:

$$\begin{cases}0=\rho\mu t-\bar{U}_{\bar{q}}(t)-(\mu+\theta(\bar{q}-1))\displaystyle\int_0^t(1-\bar{Q}_{\bar{q}+1}(s))\,\mathrm{d}s,\\ \bar{Q}_{\bar{q}+1}(t)=\bar{Q}_{\bar{q}+1}(0)+\bar{U}_{\bar{q}}(t)-(\mu+\theta\bar{q})\displaystyle\int_0^t\bar{Q}_{\bar{q}+1}(s)\,\mathrm{d}s,\end{cases}$$

from which the expressions of $\bar{Q}_{\bar{q}+1}(t)$ and $\bar{U}_{\bar{q}}(t)$ follow.

EC.9. Proof of Proposition 2

The monotonicity of $\hat{\alpha}(\infty)$ follows from Theorem 1 in Chojnacki (2008). By Theorem 2 in Chojnacki (2008), we obtain $\lim_{\theta\downarrow 0}\hat{\alpha}(\infty)=1/2$. For $s>0$, $\mathrm{e}^{-s}s^s<s\gamma(s,s)<s^s$ and $\lim_{s\downarrow 0}s\Gamma(s)=\lim_{s\downarrow 0}\Gamma(s+1)=1$. It follows that $\lim_{\theta\to\infty}\hat{\alpha}(\infty)=\lim_{s\downarrow 0}s^s=1$.

EC.10. The DQ–JSQ Design When Patience Times are Long

In this section, we evaluate the fluid model for the DQ–JSQ system when customers' patience times are relatively long. A queueing system is considered with mean service time $1/\mu=1.0$ and mean patience time $1/\theta=5.0,10.0,20.0$, respectively (i.e., the abandonment rates are $\theta=0.2,0.1,0.05$).

We first set the number of servers to be $n=100$ and take $\rho=1+\theta/(2\mu)$, in order for condition (1) to hold. We summarize both simulation results (with 95% confidence intervals) and fluid approximations under the DQ–JSQ design in Table EC.1, where exact performance measures for the PQ

Table EC.1 Performance Comparison Between the DQ–JSQ and PQ Designs for $n = 100$ and $\rho = 1 + \theta/(2\mu)$

	$\theta = 0.2$ and $\rho = 1.1$			$\theta = 0.1$ and $\rho = 1.05$			$\theta = 0.05$ and $\rho = 1.025$		
	DQ–JSQ		PQ	DQ–JSQ		PQ	DQ–JSQ		PQ
	Sim.	App.	Exact	Sim.	App.	Exact	Sim.	App.	Exact
$P(\text{De})$	0.590 ± 0.004	*0.546*	0.989	0.589 ± 0.005	*0.524*	0.967	0.598 ± 0.006	*0.512*	0.947
$P(\text{Ab})$	0.099 ± 0.002	*0.091*	0.091	0.055 ± 0.002	*0.048*	0.050	0.031 ± 0.001	*0.024*	0.028
$\mathbb{E}[X(\infty)]$	153.2 ± 0.4	*150.0*	150.3	156.0 ± 0.6	*150.0*	152.0	159.6 ± 0.8	*150.0*	157.5
$\mathbb{E}[W]$	0.591 ± 0.006	*0.546*	0.485	0.601 ± 0.007	*0.524*	0.515	0.634 ± 0.009	*0.512*	0.576
$\mathbb{E}[W \mid W > 0]$	0.998 ± 0.006	*1.000*	0.490	1.010 ± 0.007	*1.000*	0.532	1.042 ± 0.007	*1.000*	0.609
$\mathbb{E}[V]$	0.492 ± 0.005	*0.455*	0.457	0.545 ± 0.007	*0.476*	0.497	0.601 ± 0.008	*0.488*	0.565
$\mathbb{E}[V \mid W < R]$	0.457 ± 0.005	*0.417*	0.475	0.527 ± 0.007	*0.455*	0.506	0.591 ± 0.008	*0.476*	0.569
$\mathbb{E}[V \mid W > R]$	0.836 ± 0.002	*0.833*	0.284	0.913 ± 0.002	*0.909*	0.334	0.999 ± 0.003	*0.952*	0.409
$\text{Var}(W)$	0.837 ± 0.017	*0.793*	0.048	0.864 ± 0.020	*0.773*	0.086	0.927 ± 0.024	*0.762*	0.145

Notes. The Markovian queueing system has $1/\mu = 1.0$, $1/\theta = 5.0, 10.0, 20.0$, $\rho = 1 + \theta/(2\mu)$, and $n = 100$ under the two queue structures. Both simulation results (with 95% confidence intervals) and fluid approximations (in italics) are provided for the DQ–JSQ design; exact results are provided for the PQ design.

Table EC.2 Performance Comparison Between the DQ–JSQ and PQ Designs for $n = 100$ and $\rho = 1.2$

	$\theta = 0.2$			$\theta = 0.1$			$\theta = 0.05$		
	DQ–JSQ		PQ	DQ–JSQ		PQ	DQ–JSQ		PQ
	Sim.	App.	Exact	Sim.	App.	Exact	Sim.	App.	Exact
$P(\text{De})$	0.906 ± 0.003	*1.000*	1.000	0.998 ± 0.001	*1.000*	1.000	1.000 ± 0.000	*1.000*	1.000
$P(\text{Ab})$	0.166 ± 0.002	*0.167*	0.167	0.167 ± 0.002	*0.167*	0.167	0.165 ± 0.002	*0.167*	0.167
$\mathbb{E}[X(\infty)]$	200.4 ± 0.5	*200.0*	200.0	299.8 ± 0.8	*300.0*	300.0	499.8 ± 1.1	*500.0*	500.0
$\mathbb{E}[W]$	1.022 ± 0.007	*1.000*	0.917	1.918 ± 0.009	*1.909*	1.828	3.739 ± 0.013	*3.731*	3.651
$\mathbb{E}[W \mid W > 0]$	1.123 ± 0.006	*1.000*	0.917	1.921 ± 0.009	*1.909*	1.828	3.739 ± 0.012	*3.731*	3.651
$\mathbb{E}[V]$	0.838 ± 0.005	*0.833*	0.833	1.667 ± 0.008	*1.667*	1.667	3.337 ± 0.012	*3.333*	3.333
$\mathbb{E}[V \mid W < R]$	0.827 ± 0.006	*0.833*	0.907	1.735 ± 0.009	*1.742*	1.818	3.565 ± 0.013	*3.564*	3.641
$\mathbb{E}[V \mid W > R]$	0.901 ± 0.011	*0.833*	0.467	1.331 ± 0.014	*1.288*	0.909	2.205 ± 0.021	*2.178*	1.793
$\text{Var}(W)$	1.195 ± 0.025	*1.000*	0.050	2.070 ± 0.050	*1.826*	0.100	3.791 ± 0.119	*3.490*	0.200

Notes. The Markovian queueing system has $1/\mu = 1.0$, $1/\theta = 5.0, 10.0, 20.0$, $\rho = 1.2$, and $n = 100$ under the two queue structures. Both simulation results (with 95% confidence intervals) and fluid approximations (in italics) are provided for the DQ–JSQ design; exact results are provided for the PQ design.

design are also provided for comparison. Although the approximate results are generally satisfactory in this table, the fluid approximations become less accurate when θ is smaller. One possible reason for this phenomenon is as follows: With $\rho = 1 + \theta/(2\mu)$, the traffic intensity approaches one as θ gets small. When θ is close to zero, the system will operate in a critically loaded regime rather than an overloaded regime—in this example, the traffic intensities are $\rho = 1.1, 1.05, 1.025$, respec-

Table EC.3 Performance Comparison Between the DQ–JSQ and PQ Designs for $n = 20$ and $\rho = 1 + \theta/(2\mu)$

	$\theta = 0.2$ and $\rho = 1.1$			$\theta = 0.1$ and $\rho = 1.05$			$\theta = 0.05$ and $\rho = 1.025$		
	DQ–JSQ		PQ	DQ–JSQ		PQ	DQ–JSQ		PQ
	Sim.	App.	Exact	Sim.	App.	Exact	Sim.	App.	Exact
$P(\mathrm{De})$	0.659 ± 0.004	*0.546*	0.899	0.687 ± 0.004	*0.524*	0.886	0.734 ± 0.004	*0.512*	0.891
$P(\mathrm{Ab})$	0.122 ± 0.002	*0.091*	0.106	0.077 ± 0.002	*0.048*	0.067	0.047 ± 0.001	*0.024*	0.044
$\mathbb{E}[X(\infty)]$	32.71 ± 0.10	*30.00*	31.32	35.33 ± 0.15	*30.00*	33.60	38.59 ± 0.21	*30.00*	37.54
$\mathbb{E}[W]$	0.743 ± 0.006	*0.546*	0.578	0.855 ± 0.008	*0.524*	0.705	1.006 ± 0.011	*0.512*	0.909
$\mathbb{E}[W \vert W > 0]$	1.106 ± 0.005	*1.000*	0.643	1.200 ± 0.007	*1.000*	0.796	1.319 ± 0.009	*1.000*	1.020
$\mathbb{E}[V]$	0.609 ± 0.005	*0.455*	0.530	0.763 ± 0.007	*0.476*	0.667	0.945 ± 0.006	*0.488*	0.875
$\mathbb{E}[V \vert W < R]$	0.576 ± 0.005	*0.417*	0.542	0.749 ± 0.008	*0.455*	0.674	0.943 ± 0.010	*0.476*	0.881
$\mathbb{E}[V \vert W > R]$	0.886 ± 0.010	*0.833*	0.425	1.028 ± 0.013	*0.909*	0.555	1.152 ± 0.016	*0.952*	0.738
$\mathrm{Var}(W)$	1.069 ± 0.019	*0.793*	0.185	1.294 ± 0.028	*0.773*	0.315	1.591 ± 0.060	*0.762*	0.549

Notes. The Markovian queueing system has $1/\mu = 1.0$, $1/\theta = 5.0, 10.0, 20.0$, $\rho = 1 + \theta/(2\mu)$, and $n = 20$ under the two queue structures. Both simulation results (with 95% confidence intervals) and fluid approximations (in italics) are provided for the DQ–JSQ design; exact results are provided for the PQ design.

Table EC.4 Performance Comparison Between the DQ–JSQ and PQ Designs for $n = 20$ and $\rho = 1.2$

	$\theta = 0.2$			$\theta = 0.1$			$\theta = 0.05$		
	DQ–JSQ		PQ	DQ–JSQ		PQ	DQ–JSQ		PQ
	Sim.	App.	Exact	Sim.	App.	Exact	Sim.	App.	Exact
$P(\mathrm{De})$	0.841 ± 0.003	*1.000*	0.980	0.968 ± 0.001	*1.000*	0.997	0.999 ± 0.001	*1.000*	1.000
$P(\mathrm{Ab})$	0.176 ± 0.002	*0.167*	0.169	0.168 ± 0.002	*0.167*	0.167	0.167 ± 0.002	*0.167*	0.167
$\mathbb{E}[X(\infty)]$	40.97 ± 0.12	*40.00*	40.23	60.30 ± 0.22	*60.00*	60.06	99.89 ± 0.34	*100.0*	100.0
$\mathbb{E}[W]$	1.099 ± 0.007	*1.000*	0.845	1.957 ± 0.011	*1.909*	1.851	3.753 ± 0.017	*3.731*	3.672
$\mathbb{E}[W \vert W > 0]$	1.288 ± 0.006	*1.000*	0.969	2.006 ± 0.010	*1.909*	1.856	3.755 ± 0.017	*3.731*	3.672
$\mathbb{E}[V]$	0.885 ± 0.006	*0.833*	0.949	1.684 ± 0.009	*1.667*	1.670	3.330 ± 0.014	*3.333*	3.333
$\mathbb{E}[V \vert W < R]$	0.873 ± 0.006	*0.833*	0.903	1.758 ± 0.010	*1.742*	1.802	3.573 ± 0.016	*3.564*	3.622
$\mathbb{E}[V \vert W > R]$	0.973 ± 0.008	*0.833*	0.562	1.384 ± 0.011	*1.288*	1.008	2.208 ± 0.016	*2.178*	1.893
$\mathrm{Var}(W)$	1.443 ± 0.026	*1.000*	0.235	2.465 ± 0.059	*1.826*	0.495	4.462 ± 0.150	*3.490*	1.001

Notes. The Markovian queueing system has $1/\mu = 1.0$, $1/\theta = 5.0, 10.0, 20.0$, $\rho = 1.2$, and $n = 20$ under the two queue structures. Both simulation results (with 95% confidence intervals) and fluid approximations (in italics) are provided for the DQ–JSQ design; exact results are provided for the PQ design.

tively. The fluid approximations, which are substantiated by asymptotic analysis for overloaded systems, may not be as accurate in a critically loaded regime.

When customers' patience times are long, the fluid model may still provide accurate approximations for overloaded systems. If we fix the traffic intensity at $\rho = 1.2$ in the above example, the corresponding fluid approximations become accurate again—such numerical results are summarized in Table EC.2. Because condition (1) no longer holds for $\theta = 0.2, 0.1, 0.05$, we use the formulas proposed in Proposition EC.1 to produce fluid approximations in this table. Indeed, we expect that as the mean patience time goes large, the augmented queue length process, being properly scaled, will also converge to the fluid limit specified in Theorem 3 in the overloaded regime. (One may refer to He 2016 for a joint scaling approach where both the number of servers and the mean patience time are used as scaling factors.) Such a fluid limit, however, may not well capture the dynamics of the DQ–JSQ system in a critically loaded regime, in which case a diffusion limit may serve as a more refined approximate model. We would leave such topics for future research.

We also change the number of servers to $n = 20$ and repeat the above numerical experiments. The numerical results are summarized in Tables EC.3 and EC.4. The fluid approximations appear to be less accurate when n is not large, and the approximation errors are much greater when the traffic intensity is close to one. Such observations are consistent with the previous numerical examples.

References

Billingsley P (1999) *Convergence of Probability Measures* (New York: Wiley), 2nd edition.

Chen H, Yao DD (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization* (New York: Springer).

Chojnacki W (2008) Some monotonicity and limit results for the regularised incomplete gamma function. *Annales Polonici Mathematici* 94(3):283–291.

Dong J, Feldman P, Yom-Tov GB (2015) Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research* 63(2):305–324.

Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence* (New York: Wiley).

He S (2016) Diffusion approximation for efficiency-driven queues when customers are patient. *Operations Research* To appear.

Mukherjee D, Borst SC, van Leeuwaarden JSH, Whiting PA (2019) Asymptotic optimality of power-of-d load balancing in large-scale systems. *Mathematics of Operations Research* To appear.

Reed J, Ward AR (2004) A diffusion approximation for a generalized Jackson network with reneging. *Proceedings of the 42nd Allerton Conference on Communication, Control, and Computing.*

Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50(10):1449–1461.