Proofs for Theorems

In this electronic companion, we present the proofs of the technical results presented in our manuscript "Many-Server Queueing Systems with Heterogeneous Strategic Servers in Heavy Traffic". The proofs are organized following the sections in the main document.

EC.1. Additional Notation

In addition to the notation described in Section 1.2, we define the short-hand notation $x \wedge y := \min\{x, y\}$. Set d_S to be the metric for Skorokhod-J_1 topology and lastly, we denote the optional quadratic variation of a stochastic process $Y(t)$ as $[Y]_t$.

EC.2. Proofs for Results Presented in Section 3.2

LEMMA 1. *For $1/2 \leq \alpha \leq 1$ and $t \geq 0$, the sequence of processes $\{(\hat{X}^n_\alpha(t))^+\}_{n\in\mathbb{N}}$ and $\{\hat{I}^n_\alpha(t)\}_{n\in\mathbb{N}}$ are stochastically bounded. Moreover, for $1/2 < \alpha \leq 1$ there exists a time $t_M > 0$ such that $\sup_{t_M \leq t \leq T}(\hat{X}^n_\alpha(t))^+ \xrightarrow{p} 0$ for all $T > t_M$ as $n \to \infty$.*

Proof. Let $T > 0$, $n \in \mathbb{N}$, and set $\theta^{n,0} := \inf\{t \geq 0 : (\hat{X}^n_\alpha(t))^+ \leq 0\}$. Then, for all $0 \leq t \leq \theta^{n,0}$,

$$\begin{aligned}
(\hat{X}^n_\alpha(t))^+ &\leq (\hat{X}^n_\alpha(0))^+ + n^{-\alpha}A^n(t) - n^{-\alpha}\sum_{k=1}^{N^n_\alpha} S_k(\tilde{\mu}^n_k t), \\
&\leq (\hat{X}^n_\alpha(0))^+ + n^{-\alpha}\left(A^n(t) - \lambda^n t\right) - n^{-\alpha}\sum_{k=1}^{N^n_\alpha}\left(S_k(\tilde{\mu}^n_k t) - \tilde{\mu}^n_k t\right) \\
&\quad + n^{-\alpha}\left(\lambda^n - N^n_\alpha \bar{\mu}_F\right)t - n^{-\alpha}\sum_{k=1}^{N^n_\alpha}\left(\tilde{\mu}^n_k - \bar{\mu}_F\right)t \\
&\xrightarrow{p} \xi_0 - \beta\bar{\lambda}^\alpha \bar{\mu}_F^{1-\alpha} t. \qquad \text{(EC.1)}
\end{aligned}$$

Defining $\theta^{n,1}_\epsilon := \inf\left\{s > \theta^{n,0} : \left(\hat{X}^n_\alpha(s)\right)^+ > \epsilon\right\}$ and $\tilde{\theta}^{n,1}_\epsilon := \sup\left\{\theta^{n,0} \leq s < \theta^{n,1}_\epsilon : \left(\hat{X}^n_\alpha(s)\right)^+ < \epsilon/2\right\}$,

$$\begin{aligned}
\mathbb{P}\left(\sup_{\theta^{n,0}\leq s\leq T}\left(\hat{X}^n_\alpha(s)\right)^+ > \epsilon\right) &\leq \mathbb{P}\left(\theta^{n,0} < \tilde{\theta}^{n,1}_\epsilon \leq \theta^{n,1}_\epsilon \leq T\right) \\
&\leq \mathbb{P}\left(\sup_{\theta^{n,0}\leq s_1 < s_2 \leq T}\left\{\left|\frac{A^n(s_2) - A^n(s_1) - \lambda^n(s_2 - s_1)}{n^\alpha}\right|\right.\right.
\end{aligned}$$

$$+\left|\frac{\sum_{k=1}^{N^n}\left(S_k(\tilde{\mu}_k s_2)-S_k(\tilde{\mu}_k s_1)-\tilde{\mu}_k(s_2-s_1)\right)}{n^\alpha}\right| + \left|\frac{(\lambda^n-\sum_{k=1}^{N^n_\alpha}\tilde{\mu}_k)(s_2-s_1)}{n^\alpha}\right|\Bigg\} > \epsilon/2\Bigg)$$

$$\leq 2\mathbb{P}\left(n^{-\alpha}|A^n(s)-\lambda^n s|^*_T > \epsilon/16\right) + 2\mathbb{P}\left(n^{-\alpha}\left|\sum_{k=1}^{n_\alpha} S^n_k(s)-\tilde{\mu}^n_k s\right|^*_T > \epsilon/16\right)$$

$$+\mathbb{P}\left(n^{-\alpha}\left(\lambda^n-\sum_{k=1}^{N^n_\alpha}\tilde{\mu}_k\right)>0\right). \tag{EC.2}$$

The right-hand side of (EC.2) converges to 0, and combining (EC.1) and (EC.2) we obtain the stochastic boundedness of $\{(\hat{X}^n_\alpha(t))^+\}_{n\in\mathbb{N}}$. Now, setting $t_M=(M_0+1)/\bar{\lambda}^\alpha\bar{\mu}_F^{1-\alpha}$, we have

$$\mathbb{P}\left(\sup_{t_M\leq s\leq T}\left(\hat{X}^n_\alpha(s)\right)^+>\epsilon\right)\leq\mathbb{P}\left(\sup_{\theta^{n,0}\leq s\leq T}\left(\hat{X}^n_\alpha(s)\right)^+>\epsilon,\theta^{n,0}\leq t_M\right)+\mathbb{P}\left(\theta^{n,0}>t_M\right).$$

Equation (EC.2) implies that the first probability term on the right-hand side converges to 0. For the second term, $\theta^{n,0}>t_M$ implies that $\left(\hat{X}^n_\alpha(t_M)\right)^+>0$. Equation (EC.1) then implies that the probability of this event goes to 0, which proves the second part of the lemma.

We now prove the stochastic boundedness of $\{\hat{I}^n_\alpha(t)\}_{n\in\mathbb{N}}$. The case when $\alpha=1/2$ was established in Lemma 1 in Büke and Qin (2023). The proof for $1/2<\alpha\leq 1$ is simpler. We have

$$\left|\hat{I}^n_\alpha(t)\right|^*_T\leq\left|\hat{X}^n_\alpha(0)\right|+\left|\frac{A^n(t)-\lambda^n t}{n^\alpha}\right|^*_T+\left|\frac{\sum_{k=1}^{N^n_\alpha}S^n_k(t)-\mu^n_k t}{n^\alpha}\right|^*_T+\left|\frac{\sum_{k=1}^{N^n_\alpha}\mu^n_k t-\lambda^n t}{n^\alpha}\right|^*_T$$
$$+\left|\frac{R^n\left(\gamma\int_0^t(X^n(s)-N^n_\alpha)^+ds\right)-\gamma\int_0^t(X^n(s)-N^n_\alpha)^+ds}{n^\alpha}\right|^*_T+\left|\gamma\int_0^t(\hat{X}^n_\alpha(s))^+ds\right|^*_T.$$

The stochastic boundedness of $\{(\hat{X}^n_\alpha(t))^+\}_{n\in\mathbb{N}}$ and the martingale central limit theorem (c.f., Pang et al. (2007)) implies that the last two terms converge to 0 in probability. The second and third terms converge to 0 in probability due to law of large numbers for Poisson processes. Finally, the first and the fourth terms are stochastically bounded due to Assumption 2 and the central limit theorem. This proves the lemma. □

EC.3. Proofs for Results Presented in Section 4

THEOREM 2. *Suppose that the limiting fairness process under the adopted routing policy is $\eta_{\alpha,t}$. Then, the scaled system length processes $\{\hat{X}^n_\alpha\}_{n\in\mathbb{N}}\Rightarrow\xi_\alpha$ as $n\to\infty$, where*

1. *if* $\alpha = 1/2$, *then the process* ξ_α *is the strong solution to the stochastic differential equation*

$$\xi(t) = \xi_0 + \sqrt{2\bar{\lambda}}W(t) + \left(\beta_F(\bar{\lambda}\bar{\mu}_F)^{1/2} + \zeta_1\right)t + \langle \iota, \eta_{1/2,t}\rangle \int_0^t \xi^-(s)ds - \gamma\int_0^t \xi^+(s)ds,$$

for all $t \geq 0$ *where* $\zeta_1 \sim$ *Normal*$(0,\sigma_F^2\bar{\lambda}^\alpha\bar{\mu}_F^{-\alpha})$ *and* W *is a standard Brownian motion.*

2. *if* $1/2 < \alpha \leq 1$, *then the process* ξ_α *is the solution to the ordinary differential equation*

$$\xi(t) = \xi_0 + \beta\bar{\lambda}^\alpha\bar{\mu}_F^{1-\alpha}t + \langle \iota, \eta_{\alpha,t}\rangle \int_0^t \xi^-(s)ds, \text{ for all } t \geq 0.$$

Proof. Our proof follows the same steps as Theorem 5 in Büke and Qin (2023) with the additional modification on the number of servers. Adding and subtracting appropriate terms to (4) and normalizing with n^α, we get

$$\hat{X}_\alpha^n(t) = \hat{X}_\alpha^n(0) + \hat{M}_{\alpha,1}^n(t) - \hat{M}_{\alpha,2}^n(t) - \hat{M}_{\alpha,3}^n(t) + n^{-\alpha}(\lambda^n - N_\alpha^n\bar{\mu}_F)t - n^{-\alpha}\left(\sum_{k=1}^{N_\alpha^n}\tilde{\mu}_k^n - N_\alpha^n\bar{\mu}_F\right)t$$
$$-\sum_{k=1}^{N_\alpha^n}\tilde{\mu}_k^n\int_0^t \hat{I}_{k,\alpha}^n(s)ds - \gamma\int_0^t(\hat{X}_\alpha^n(s))^+ds, \tag{EC.3}$$

where

$$\hat{M}_{\alpha,1}^n(t) := \frac{A^n(t) - \lambda^n t}{n^\alpha},$$
$$\hat{M}_{\alpha,2}^n(t) := \frac{S^n\left(\sum_{k=1}^{N_\alpha^n}\tilde{\mu}_k^n\left(t - \int_0^t I_k^n(s)ds\right)\right) - \sum_{k=1}^{N_\alpha^n}\tilde{\mu}_k^n\left(t - \int_0^t I_k^n(s)ds\right)}{n^\alpha},$$
$$\hat{M}_{\alpha,3}^n(t) := \frac{R^n\left(\gamma\int_0^t(X^n(s) - N_\alpha^n)^+ds\right) - \gamma\int_0^t(X^n(s) - N_\alpha^n)^+ds}{n^\alpha}.$$

Using the martingale central limit theorem, both $\hat{M}_{\alpha,1}^n(t)$ and $\hat{M}_{\alpha,2}^n(t)$ converge weakly to 0 when $\alpha > 1/2$ and to $\sqrt{\bar{\lambda}}W(t)$, where $W(t)$ is a standard Brownian motion when $\alpha = 1/2$. To see this, we focus on $\hat{M}_{\alpha,2}^n$. The process $M_{\alpha,2}^n(t)$ is a compensated Poisson process and is a martingale with predictable quadratic variation

$$\frac{\sum_{k=1}^{N_\alpha^n}\tilde{\mu}_k^n\left(t - \int_0^t I_k^n(s)ds\right)}{n^{2\alpha}} = \frac{N_\alpha^n}{n^{2\alpha}}\frac{\sum_{k=1}^{N_\alpha^n}\tilde{\mu}_k^n t}{N_\alpha^n} - \frac{\int_0^t \hat{I}_\alpha^n(s)ds}{n^\alpha}.$$

The stochastic boundedness of $\hat{I}_\alpha^n$ (Lemma 1) implies that the second term on the right-hand side converges 0 for all $1/2 \leq \alpha \leq 1$. Using the law of large numbers, the first term converges to 0 if

$\alpha > 1/2$ and to $\bar{\lambda}$ if $\alpha = 1/2$. The jumps of $M^n_{\alpha,2}(t)$ are bounded by $1/n^\alpha$, hence the martingale central limit theorem implies that $M^n_{\alpha,2}(t) \Rightarrow \sqrt{\bar{\lambda}} W(t)$. The proof for $M^n_{\alpha,1}(t)$ follows the same steps. Similarly, $\hat{M}^n_{\alpha,3}(t) \Rightarrow 0$ for all $\alpha \geq 1/2$ as a result of the stochastic boundedness of Lemma 1. Plugging in (1), we get $n^{-\alpha}(\lambda^n - N^n_\alpha \bar{\mu}_F)t \to -\beta \bar{\lambda}^\alpha \bar{\mu}_F^{1-\alpha} t$. Finally, using the central limit theorem, we have $n^{-\alpha}\left(\sum_{k=1}^{N^n_\alpha} \tilde{\mu}^n_k - N^n_\alpha \bar{\mu}_F\right) t \xrightarrow{p} 0$ for $\alpha > 1/2$ and $n^{-\alpha}\left(\sum_{k=1}^{N^n_\alpha} \tilde{\mu}^n_k - N^n_\alpha \bar{\mu}_F\right) t \Rightarrow \zeta \bar{\lambda}^\alpha \bar{\mu}_F^{-\alpha} t$ for $\alpha = 1/2$, where ζ is a normal random variable with mean 0 and variance σ_F^2. As $\mathcal{S}_\epsilon \eta^n_{\alpha,t} \Rightarrow \mathcal{S}_\epsilon \eta_{\alpha,t}$ for any $\epsilon > 0$, using Lemma 2 in Büke and Qin (2023), a modification of the Skorokhod representation theorem, we can assume that all the above processes converge with probability 1 and to prove the theorem, we need only to prove that for any $\rho > 0$, $\mathbb{P}(d_S(\hat{X}^n_\alpha, \xi_\alpha) > \rho) \to 0$ as $n \to \infty$, where $d_S(\cdot,\cdot)$ is the metric for Skorokhod-J_1 topology.

For any $\varpi > 0$, we can find a sequence of homeomorphisms $\Lambda^n(t) : [0,T] \to [0,T]$ with derivative $\dot{\Lambda}^n(t)$ and N_ϖ such that for any $n > N_\varpi$,

$$\left|\hat{M}^n_{\alpha,1}(t) + \hat{M}^n_{\alpha,2}(t) + \hat{M}^n_{\alpha,3}(t) - \sqrt{2\bar{\lambda}} W(\Lambda^n t)\right|^*_T \vee \left|\langle \iota, \mathcal{S}_\epsilon \eta^n_{\alpha,t}\rangle - \langle \iota, \mathcal{S}_\epsilon \eta_{\alpha,\Lambda^n(t)}\rangle\right|^*_T \vee \left|\dot{\Lambda}^n(t) - 1\right|^*_T < \varpi,$$

if $\alpha = 1/2$ and

$$\left|\hat{M}^n_{\alpha,1}(t) + \hat{M}^n_{\alpha,2}(t) + \hat{M}^n_{\alpha,3}(t)\right|^*_T \vee \left|\langle \iota, \mathcal{S}_\epsilon \eta^n_{\alpha,t}\rangle - \langle \iota, \mathcal{S}_\epsilon \eta_{\alpha,\Lambda^n(t)}\rangle\right|^*_T \vee \left|\dot{\Lambda}^n(t) - 1\right|^*_T < \varpi,$$

if $1/2 < \alpha \leq 1$. We will prove our result by showing

$$\sup_{0 \leq t \leq T} \left|\hat{X}^n_\alpha(t) - \xi_\alpha(\Lambda^n(t))\right| \to 0, \text{w.p. } 1. \tag{EC.4}$$

Using the tightness of the processes, without loss of generality, we can assume that

$$\sup_{n \in \mathbb{N}} \left\{ \left|\langle \iota, \eta^n_{\alpha,t}\rangle\right|^*_T \vee \left|\hat{X}^n_\alpha(t)\right|^*_T \vee \left|\xi_\alpha(t)\right|^*_T \right\} < K.$$

Now, taking $\xi_\alpha(t)$ to be the solution of the appropriate equation in the statement of the theorem, plugging in the definition of fairness process for the seventh term on the right-hand side of (EC.3), for any ϖ we have an N_ϖ such that $n > N_\varpi$ implies

$$\begin{aligned} |\hat{X}^n_\alpha(t) - \xi(\Lambda^n(t))| \leq \varpi + \gamma &\left| \int_0^t (\hat{X}^n_\alpha(s))^+ ds - \int_0^{\Lambda^n(t)} (\xi(s))^+ ds \right| \\ &+ \left| \langle \iota, \eta^n_{\alpha,t}\rangle \int_0^t (\hat{X}^n_\alpha(s))^- ds - \langle \iota, \eta^n_{\alpha,\Lambda^n(t)}\rangle \int_0^{\Lambda^n(t)} (\xi(s))^- ds \right|. \end{aligned} \tag{EC.5}$$

We can bound the second term on the right-hand side of (EC.5) as

$$\left|\int_0^t (\hat{X}^n_\alpha(s))^+ ds - \int_0^{\Lambda^n(t)} (\xi(s))^+ ds\right| \le \int_0^t \left|\hat{X}^n_\alpha(s) - \xi_\alpha(\Lambda^n(s))ds\right| + \int_0^t \left|(1-\dot{\Lambda}^n(t))\xi_\alpha(\Lambda^n(s))\right| ds$$
$$\le \int_0^t \left|\hat{X}^n_\alpha(s) - \xi_\alpha(\Lambda^n(s))ds\right| + \varpi K t.$$

To bound the third term on the right-hand side of (EC.5),

$$\left|\langle \iota, \eta^n_{\alpha,t}\rangle \int_0^t (\hat{X}^n_\alpha(s))^- ds - \langle \iota, \eta_{\alpha,\Lambda^n(t)}\rangle \int_0^{\Lambda(t)} (\xi(s))^- ds\right|$$
$$\le \left|\left(\langle \iota, \eta^n_{\alpha,t}\rangle - \langle \iota, \mathcal{S}_\epsilon \eta^n_{\alpha,t}\rangle\right) \int_0^t (\hat{X}^n_\alpha(s))^- ds - \left(\langle \iota, \eta_{\alpha,\Lambda^n(t)}\rangle - \langle \iota, \mathcal{S}_\epsilon \eta_{\alpha,\Lambda^n(t)}\rangle\right) \int_0^{\Lambda(t)} (\xi_\alpha(s))^- ds\right|$$
$$+ \left|\left(\langle \iota, \mathcal{S}_\epsilon \eta^n_{\alpha,t}\rangle - \langle \iota, \mathcal{S}_\epsilon \eta_{\alpha,\Lambda^n(t)}\rangle\right) \int_0^t (\hat{X}^n_\alpha(s))^- ds - \langle \iota, \mathcal{S}_\epsilon \eta_{\alpha,\Lambda^n(t)}\rangle \int_0^t \left(\left(\hat{X}^n_\alpha(s)\right)^- - \left(\xi_\alpha(\Lambda(s))\right)^-\right) ds\right|$$
$$+ \left|\langle \iota, \mathcal{S}_\epsilon \eta_{\alpha,\Lambda^n(t)}\rangle \int_0^t (1-\dot{\Lambda}(s))\left(\xi_\alpha(\Lambda(s))\right)^- ds\right|$$
$$\le (2\epsilon + \varpi(1+K))Kt + K\int_0^t \left|\hat{X}^n(s) - \hat{\xi}(\Lambda(s))\right| ds.$$

Plugging these bounds in (EC.5), using Grönwall's inequality, and choosing ϵ and ϖ appropriately, our result follows. □

Now, we prove the interchangeability of many-server limit and limit as $t \to \infty$. To do so, we need the following uniform integrability result.

LEMMA EC.1. *For any $1/2 \le \alpha \le 1$, the stationary scaled system lengths $\{\hat{X}^n(\infty)\}_{n\in\mathbb{N}}$ are uniformly integrable and hence tight.*

Proof. For any $n \in \mathbb{N}$, we decompose the scaled system length process into its negative and positive parts as

$$\hat{X}^n_\alpha(\infty) = \left(\hat{X}^n_\alpha(\infty)\right)^+ - \left(\hat{X}^n_\alpha(\infty)\right)^-,$$

and prove the uniform integrability of each part separately. We consider $\left(\hat{X}^n_\alpha(t)\right)^-$ and use a coupling argument. Now, suppose we fix $n > 0$ and let $\theta^n_{A,i}$ be the occurrence time of the ith event for Poisson process $A^n(t)$. Assume that we know $(\tilde{\mu}^n_1, \ldots, \tilde{\mu}^n_{N^n_\alpha})$ and define $\tilde{S}^n(t)$ to be a Poisson process with rate $\sum_{k=1}^{N^n} \tilde{\mu}^n_k$ and let $\theta^n_{S,i}$ be the occurrence time for the ith event. We also define a

sequence $\{U_i^n\}_{i\in\mathbb{N}}$ of independent uniform(0,1) random variables. For any given time t, we know that the system should be serving with a rate equal to the sum of service rates of busy servers, i.e., $\sum_{k=1}^{N^n}\tilde{\mu}_k^n(1-I_k^n(t))$, hence we use the thinning property of Poisson process where event i of $\tilde{S}^n(t)$ is accepted as an actual departure with probability $p_i^n(\theta_{S_i}^n-)=\frac{\sum_{k=1}^{N^n}\tilde{\mu}_k^n(1-I_k^n(\theta_{S,i}^n-))}{\sum_{k=1}^{N^n}\tilde{\mu}_k^n}$ by checking whether $U_i^n\le p_i^n(\theta_{S_i}^n-)$ or not. The processes $I_k^n(t)$ can be rigorously defined by using U_i^ns and a routing process based on our routing policy. As this does not play a major role in our proof, we refer the reader to Büke and Qin (2023) for the detailed construction of idleness processes. Now, we can write

$$\begin{aligned}\left(\hat{X}_\alpha^n(t)\right)^- = \left(\hat{X}_\alpha^n(0)\right)^- + n^{-\alpha}\sum_{i=1}^{\tilde{S}^n(t)}\mathbb{I}\left(\left(\hat{X}_\alpha^n(\theta_{S,i}^n-)\right)^+=0\right)\mathbb{I}\left(U_i^n\le p_i^n\left(\theta_{S_i}^n-\right)\right)\\ - n^{-\alpha}\sum_{i=1}^{A^n(t)}\mathbb{I}\left(\left(\hat{X}_\alpha^n(\theta_{A,i}^n-)\right)^->0\right).\end{aligned}$$

We define a birth-death process $\{Y_1^n(t)\}_{n\in\mathbb{N}}$ with $Y_1^n(0)=(X^n(0))^-$ w.p.1, whose birth rate at $Y^n(t)=i$ is $\sum_{k=1}\tilde{\mu}_k^n-\mu_{\min}i$ and death rate is λ^n if $Y_1^n(t)>0$. Then, we can couple the scaled process $\hat{Y}_1^n(t)=n^{-\alpha}Y_1^n(t)$ with the system length process by writing it as

$$\hat{Y}_1^n(t)=\hat{Y}_1^n(0)+n^{-\alpha}\sum_{i=1}^{\tilde{S}^n(t)}\mathbb{I}\left(U_i^n\le\tilde{p}_i^n\left(\theta_{S,i}^n-\right)\right)-n^{-\alpha}\sum_{i=1}^{A^n(t)}\mathbb{I}\left(\hat{Y}_1^n(\theta_{A,i}^n-)>0\right),$$

where $\tilde{p}_i^n\left(\theta_{S_i}^n-\right):=\frac{\sum_{k=1}^{N^n}\tilde{\mu}_k^n-\mu_{\min}\hat{Y}_1\left(\theta_{S_i}^n-\right)}{\sum_{k=1}^{N^n}\tilde{\mu}_k^n}$. To see that $\left(\hat{X}_\alpha^n(t)\right)^-\le\hat{Y}_1^n(t)$ for all $t\ge0$ with probability 1, define $\vartheta^n=\left\{t:\left(\hat{X}_\alpha^n(t)\right)^->\hat{Y}_1^n(t)\right\}$. As at most one event occurs at any given time t with probability 1, we have $\left(\hat{X}_\alpha^n(\vartheta^n-)\right)^-=\hat{Y}_1^n(\vartheta^n-)$. By definition we have

$$\tilde{p}_i^n\left(\vartheta^n-\right)\ge p_i^n(\vartheta^n-).$$

Hence, if ϑ^n is an event epoch for $\tilde{S}^n(t)$, we have $\left(\hat{X}_\alpha^n(\vartheta^n)\right)^-\le\hat{Y}_1^n(\vartheta^n)$ and if ϑ^n is an event epoch for $A^n(t)$, we have $\left(\hat{X}_\alpha^n(\vartheta^n)\right)^-=\hat{Y}_1^n(\vartheta^n)$, both contradicts with the definition of ϑ^n. Hence, we conclude that $\left(\hat{X}_\alpha^n(t)\right)^-\le\hat{Y}_1^n(t)$ for all $t\ge0$ with probability 1.

Now, define $\zeta^n:=n^{-\alpha}\left(\sum_{k=1}^{N^n}\tilde{\mu}_k^n-N^n\bar{\mu}\right)$ and by re-arranging the terms we have

$$\sum_{k=1}^{N^n}\tilde{\mu}_k^n=\lambda^n+\beta\left(\lambda^n\right)^\alpha\bar{\mu}^{1-\alpha}+n^\alpha\zeta^n.$$

Take $M_1 := (\beta (\lambda^n)^\alpha \bar{\mu}^{1-\alpha} + n^\alpha \zeta^n)^+ + 2n^\alpha$, and define a new birth-death process $Y_2^n(t)$ with $Y_2^n(0) = Y_1^n(0)$ with probability 1, whose birth rate at $Y_2^n(t) = i$ is $\lambda^n - \mu_{\min} \min\{i, M_1\}$ and death rate is λ^n at $Y_2^n(t) \neq 0$. By definition, $Y_2(t)$ is stochastically greater than $Y_1^n(t)$. Using a similar argument as above, we can couple $Y_2(t)$ with a simple birth death process $Y_3(t)$ where $Y_3^n(0) = (Y_2^n(0) - M)^+$ with birth rate $\lambda^n - \mu_{\min} M_1$, death rate λ^n and $Y_2^n(t) \leq M_1 + Y_3^n(t)$ for all $t \geq 0$. As birth and death rates are constants, $Y_3(t)$ is equivalent to an $M/M/1$ queue. Hence, we can prove that $\left(\hat{X}^n(\infty)\right)^-$ is uniformly integrable by showing that $\sup \mathbb{E}\left[(n^{-\alpha}(M_1 + Y_3^n(\infty)))^2\right] < \infty$. For $n \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\left(n^{-\alpha}(M_1 + Y_3^n(\infty))\right)^2\right] &= n^{-2\alpha}\left(\mathbb{E}[M_1^2] + 2\mathbb{E}[M_1 Y_3^n(\infty)] + \mathbb{E}\left[(Y_3^n(\infty))^2\right]\right) \\
&= n^{-2\alpha}\left(\mathbb{E}\left[M_1^2\right] + 2\mathbb{E}\left[M_1 \mathbb{E}[Y_3^n(\infty)|\vec{\mu}]\right] + \mathbb{E}\left[\mathbb{E}\left[(Y_3^n(\infty))^2 |\vec{\mu}\right]\right]\right) \\
&= n^{-2\alpha}\left(\mathbb{E}\left[M_1^2\right] + 2\mathbb{E}\left[\frac{\lambda^n - \mu_{\min} M_1}{\mu_{\min}}\right] + \mathbb{E}\left[\frac{(\lambda^n - \mu_{\min} M_1)(2\lambda^n - \mu_{\min} M_1)}{\mu_{\min}^2 M_1^2}\right]\right) \\
&= n^{-2\alpha}\left(\mathbb{E}\left[M_1^2\right] + \frac{2\lambda^n}{\mu_{\min}} - 2\mathbb{E}\left[M_1\right] + 2\mathbb{E}\left[\frac{(\lambda^n)^2}{\mu_{\min}^2 M_1^2}\right] - 3\mathbb{E}\left[\frac{\lambda^n}{\mu_{\min} M_1}\right] + 1\right) \\
&\leq n^{-2\alpha}\left(\mathbb{E}\left[M_1^2\right] + \frac{2\lambda^n}{\mu_{\min}} + 2\mathbb{E}\left[M_1\right] + \frac{2(\lambda^n)^2}{\mu_{\min}^2 n^{2\alpha}} + \frac{3\lambda^n}{\mu_{\min} n^\alpha} + 1\right).
\end{aligned}$$

Assumption 1 implies that the second, fourth and fifth terms converge to a finite number for any $\alpha \geq 1/2$ and hence can be bounded uniformly for any n. Also, using the identity $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$n^{-2\alpha}\mathbb{E}\left[M_1^2\right] \leq 4n^{-2\alpha}\beta^2 (\lambda^n)^{2\alpha} \bar{\mu}^{2-2\alpha} + 4\mathbb{E}[(\zeta^n)^2] + 2.$$

Again we can use Assumption 1 to show that the first term converges to a finite number. Using independence we have

$$\mathbb{E}[(\zeta^n)^2] = n^{-2\alpha}\mathbb{E}\left[\left(\sum_{k=1}^{N^n} \tilde{\mu}_k^n - N^n \bar{\mu}\right)^2\right] = n^{-2\alpha}\mathbb{E}\left[\sum_{k=1}^{N^n} (\tilde{\mu}_k^n - \bar{\mu})^2\right] \leq n^{-2\alpha} N^n (\mu_{\max} - \mu_{\min})^2,$$

which also converges and hence can be uniformly bounded for all n. Similarly,

$$n^{-2\alpha}\mathbb{E}[M_1] \leq n^{-2\alpha}\mathbb{E}[\beta (\lambda^n)^\alpha \bar{\mu}^{1-\alpha} |\mu_{\max} - \mu_{\min}| + 2n^{-\alpha}] - n^{-2\alpha} N^n,$$

which also converges and proves the uniform integrability of $\left\{\left(\hat{X}_\alpha^n(\infty)\right)^-\right\}_{n\in\mathbb{N}}$. The proof of the uniform integrability of $\left\{\left(\hat{X}_\alpha^n(\infty)\right)^+\right\}_{n\in\mathbb{N}}$ follows the same lines and hence is omitted. □

THEOREM 3. *For many-server systems with random and heterogeneous service rates, for any* $1/2 \leq \alpha \leq 1$ *the following convergence results hold as* $n \to \infty$*:*

1. $\hat{X}^n_\alpha(\infty) \Rightarrow \xi_\alpha(\infty)$,
2. $\mathbb{E}[\hat{X}^n_\alpha(\infty)] \to \mathbb{E}[\xi_\alpha(\infty)]$,
3. $\eta^n_\infty \Rightarrow \eta_\infty$,

where $\xi_\alpha(t) \Rightarrow \xi_\alpha(\infty)$ *and* $\eta_t \Rightarrow \eta_\infty$ *as* $t \to \infty$.

Proof. Suppose $\hat{X}^n_\alpha(0)$, $(\tilde{\mu}^n_1, \ldots, \tilde{\mu}^n_{N^n_\alpha})$ are distributed according to the stationary measure of the nth system π^n. Then, $\hat{X}^n_\alpha(t)$, $(\tilde{\mu}^n_1, \ldots, \tilde{\mu}^n_{N^n_\alpha})$ are also distributed according to π^n. As we know that the $\{\pi^n\}$ are tight and $\hat{X}^n_\alpha(t) \Rightarrow \xi_\alpha(t)$, the first convergence holds and the second convergence holds as a result of the uniform integrability. To see the third one, we need to see that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{N^n} \delta_{\tilde{\mu}^n_k}(\mathbb{A}) \int_0^T \hat{I}^n_k(s) ds = \sum_{k=1}^{N^n} \delta_{\tilde{\mu}^n_k}(\mathbb{A}) \mathbb{E}[I^n_k(\infty)|\vec{\mu}].$$

Again, starting with a stationary system and using the uniform integrability the result follows.

□

EC.4. Proofs for Results Presented in Section 5

LEMMA 2. *The set of scaled instantaneous allocations,* $\{\bar{\psi}^n_t\}_{n\in\mathbb{N}}$*, is tight.*

Proof. We define the discrete-time measure-valued stochastic process $\mathcal{U}^n_{A,i}(\mathbb{A}) = \delta_{\mu_k}(\mathbb{A})\mathbb{I}(X^n(\theta^n_{A,i}-) \leq 0)$ for all $\mathbb{A} \in \mathcal{B}(\mathbb{R}_+)$ if the ith incoming arrival in the nth system is immediately routed to server k and 0 (thought as a measure) otherwise. Similarly, we define $\mathcal{U}^n_{S,i}(\mathbb{A}) = \delta_{\mu_k}(\mathbb{A})\mathbb{I}(X^n(\theta^n_{S,i}-) \leq 0)$ for all $\mathbb{A} \in \mathcal{B}(\mathbb{R}_+)$, if the ith potential service completion is from kth server, if it is not an actual service completion $\mathcal{U}^n_{S,i} = 0$. Then, for any $f \in C^b$ we have the following balance

$$\langle f, \bar{\psi}^n_t \rangle = \langle f, \bar{\psi}^n_0 \rangle + n^{-1} \sum_{i=1}^{S^n(t)} \langle f, \mathcal{U}^n_{S,i} \rangle - n^{-1} \sum_{i=1}^{A^n(t)} \langle f, \mathcal{U}^n_{A,i} \rangle. \tag{EC.6}$$

To prove tightness, we use the conditions introduced by Jakubowski (1986) that we recall next.

THEOREM EC.1 (**Jakubowski (1986)**). *A sequence of stochastic processes* $\{\eta_t^n\}_{n\in\mathbb{N}}$ *taking values in* $\mathbb{D}_{\mathcal{P}}[0,T]$ *is tight if and only if:*

J1. (Compact Containment Condition) For each $\rho, T>0$*, there exists a compact set* $\mathcal{K}_\rho \subset \mathcal{P}$ *such that*

$$\liminf_{n\to\infty} \mathbb{P}(\bar{\psi}_t^n \in \mathcal{K}_\rho, \text{ for all } t\in[0,T]) > 1-\rho.$$

J2. There exists a family of functions $\mathbb{F}$ *such that*

i. $H\in\mathbb{F}:\mathcal{P}\to\mathbb{R}$*,* $\mathbb{F}$ *separates points in* $\mathcal{P}$ *and* $\mathbb{F}$ *is closed under addition.*

ii. For any fixed $H\in\mathbb{F}$*, the sequence of functions* $\{h^n(t):=H(\eta_t^n), \text{ for all } t\in\mathbb{R}\}_{n\in\mathbb{N}}$ *is tight in* $\mathbb{D}_{\mathbb{R}}[0,\infty)$ *endowed with Skorokhod-*J_1 *topology.*

We now return to our proof. Using Lemma 1, we know that, for all $\epsilon>0$, there exists a K_ϵ such that $\mathbb{P}(\sup_{0\le t\le T}|\hat{I}_\alpha^n(t)|>K_\epsilon)<\epsilon$ for all $n\in\mathbb{N}$. Define $\mathcal{K}_\epsilon$ as the set of measures bounded by K_ϵ on the support $[\mu_{\min},\mu_{\max}]$. The set $\mathcal{K}_\epsilon$ is compact and Lemma 1 implies J1.

To show that J2 holds, we define

$$\mathbb{F}=\{H:\mathcal{M}_F[\mu_{\min},\mu_{\max}]\to\mathbb{R}:\exists f\in\mathbb{C}^b_{\mathbb{R}_+}[0,\infty) \text{ such that } H(\psi)=\langle f,\psi\rangle \text{ for all } \psi\in\mathcal{M}_F\},$$

where $\mathcal{M}_F[\mu_{\min},\mu_{\max}]$ is the set of finite measures on $[\mu_{\min},\mu_{\max}]$. The set $\mathbb{F}$ separates the points in $\mathcal{M}_F$ and is closed under addition. Take K_f such that $H(\psi)=\langle f,\psi\rangle$ with $f(\mu)\le K_f$ for all $\mu\in[\mu_{\min},\mu_{\max}]$. To show tightness of $\{\langle f,\bar{\psi}_t^n\rangle\}_{n\in\mathbb{N}}$, we need to show that for all $\epsilon,\rho>0$

1. there exists an $M_{f,\epsilon}$ such that $\mathbb{P}\left(\sup_{0\le t\le T}|\langle f,\bar{\psi}_t^n\rangle|>M_{f,\epsilon}\right)<\epsilon$ and
2. there exists a ρ and an N_ρ such that for all $n>N_\rho$, $\mathbb{P}\left(w(\langle f,\bar{\psi}_t^n\rangle,\rho)\ge\epsilon\right)<\epsilon$, where

$$w(\langle f,\bar{\psi}_t^n\rangle,\rho)=\inf_{\{t_i\}}\max_i \sup_{t_i\le t,s\le t_{i+1}} |\langle f,\bar{\psi}_t^n\rangle-\langle f,\bar{\psi}_s^n\rangle|$$

and $\{t_i\}_{0\le i\le\nu}$ is any ρ-sparse set, i.e., $0=t_0<t_1<\cdots<t_\nu=T$ with $\min_i|t_{i+1}-t_i|>\rho$.

Again taking $M_{f,\rho}=K_\rho K_f$, we have

$$\mathbb{P}\left(\sup_{0\le t\le T}|\langle f,\bar{\psi}_t^n\rangle|>M_{f,\rho}\right)\le\mathbb{P}\left(K_f\sup_{0\le t\le T}\hat{I}^n(t)>M_{f,\rho}\right)\le\rho,$$

which implies the first condition. Now, we prove the second condition. Using (EC.6), for any $0 \le s < t \le T$ we have

$$
\begin{aligned}
|\langle f, \bar{\psi}_t^n\rangle - \langle f, \bar{\psi}_s^n\rangle| &= \left| n^{-1} \sum_{i=1}^{S^n(t)} \langle f, \mathcal{U}_{S,i}^n\rangle - n^{-1} \sum_{i=1}^{A^n(t)} \langle f, \mathcal{U}_{A,i}^n\rangle - n^{-1} \sum_{i=1}^{S^n(s)} \langle f, \mathcal{U}_{S,i}^n\rangle - n^{-1} \sum_{i=1}^{A^n(s)} \langle f, \mathcal{U}_{A,i}^n\rangle \right| \\
&\le n^{-1} K_f \left| S^n(t) - S^n(s) - \sum_{k=1}^{N^n} \mu_k t + \sum_{k=1}^{N^n} \mu_k s \right| + n^{-1} K_f \sum_{k=1}^{N^n} \mu_k \, |t-s| \\
&\quad + n^{-1} K_f \, |A^n(t) - A^n(s) - \lambda^n t + \lambda^n s| + n^{-1} K_f \lambda^n \, |t-s| \\
&\le n^{-1} K_f \left| S^n(t) - \sum_{k=1}^{N^n} \mu_k t \right| + n^{-1} K_f \left| S^n(s) - \sum_{k=1}^{N^n} \mu_k s \right| \\
&\quad + K_f \left| n^{-1} \sum_{k=1}^{N^n} \mu_k - \bar{\lambda}(1+\beta) \right| |t-s| + n^{-1} K_f \, |A^n(t) - \lambda^n t| + |A^n(s) - \lambda^n s| \\
&\quad + n^{-1} K_f \lambda^n \, |t-s| + \bar{\lambda}(1+\beta) \, |t-s| \\
&\le n^{-1} 2 K_f \sup_{0 \le t \le T} \left| S^n(t) - \sum_{k=1}^{N^n} \mu_k t \right| + n^{-1} 2 K_f \sup_{0 \le t \le T} |A^n(t) - \lambda^n t| \\
&\quad + K_f \left| n^{-1} \sum_{k=1}^{N^n} \mu_k - \bar{\lambda}(1+\beta) \right| |t-s| + n^{-1} K_f \lambda^n \, |t-s| + \bar{\lambda}(1+\beta) \, |t-s| \, .
\end{aligned}
\tag{EC.7}
$$

Using the martingale central limit theorem , the first and second terms on the right-hand side converges to 0 in probability. Similarly, using the law of large numbers, we can show that the third term also converges to 0 in probability. Finally, from Assumption 1, we know that $n^{-1}\lambda^n \to \bar{\lambda}$. Hence, choosing $\rho < \epsilon/2(K_f \bar{\lambda}(2+\beta))$ and N large enough the second condition follows. Moreover, by examining (EC.7), one can conclude that any limit is continuous. □

LEMMA 3. *Assume the subsequence $\bar{\psi}_t^{n_k} \Rightarrow \bar{\psi}_t$ as $n_k \to \infty$. Then, $\bar{\psi}_t$ satisfies*

$$
\begin{aligned}
\langle f, \bar{\psi}_{1,t}\rangle = \langle f, \bar{\psi}_{1,0}\rangle + \frac{\bar{\lambda}}{\bar{\mu}}(1+\beta)\langle f \times \iota, F\rangle \int_0^t \mathbb{I}(\xi_{1,s} \le 0) ds - \int_0^t \langle f \times \iota, \bar{\psi}_{1,s-}\rangle \mathbb{I}(\xi_{1,s} \le 0) ds \\
- \bar{\lambda} \int_0^t \frac{\langle f \times h, \bar{\psi}_{1,s-}\rangle}{\langle h, \bar{\psi}_{s-}\rangle} \mathbb{I}(\xi_{1,s} \le 0) ds \text{ for all } t \ge 0.
\end{aligned}
\tag{EC.8}
$$

Proof. The martingale central limit theorem (c.f., Pang et al. (2007)) implies $\sup_{0 \le t \le T} n^{-1} |S^n(t) - \sum_{k=1}^{N^n} \mu_k t| \xrightarrow{p} 0$ and $\sup_{0 \le t \le T} n^{-1} |A^n(t) - \lambda^n t| \xrightarrow{p} 0$ Using the Skorokhod

representation theorem (see, e.g., Theorem 6.7 in Billingsley (1999)), we can assume that these and the subsequence in the statement of the lemma converge almost surely. Also, as mentioned in the proof of Lemma 2, for any $f \in C_{\mathbb{R}_+}[0,T]$, the limit $\langle f, \bar{\psi}_t \rangle$ is continuous and hence, the convergence holds in the supremum norm as well as the Skorokhod d_S metric. We define the processes

$$M_1^n(t) := \sum_{i=1}^{S^n(t)} \langle f, \mathcal{U}_{S,i}^n \rangle - \sum_{i=1}^{S^n(t)} \left\langle f, \frac{\sum_{k=1}^{N_\alpha^n} \mu_k (1 - I_k^n(\theta_{S,i}^n-))\delta_{\mu_k} \mathbb{I}(X^n(\theta_{S,i}^n-) \le 0)}{\sum_{k=1}^{N_\alpha^n} \mu_k} \right\rangle,$$
$$M_2^n(t) := \sum_{i=1}^{A^n(t)} \langle f, \mathcal{U}_{A,i}^n \rangle - \sum_{i=1}^{A^n(t)} \left\langle f, \frac{\sum_{k=1}^{N_\alpha^n} h(\mu_k) I_k^n(\theta_{A,i}^n-)\delta_{\mu_k} \mathbb{I}(X^n(\theta_{A,i}^n-) \le 0)}{\sum_{k=1}^{N_\alpha^n} h(\mu_k) I_k^n(\theta_{A,i}^n-)} \right\rangle,$$

where $0/0$ is assumed to be 0. It is easy to see that both M_1^n and M_2^n are $\mathcal{F}_t$ martingales. After some algebraic manipulations, Equation (EC.6) becomes

$$\begin{aligned}
\langle f, \bar{\psi}_t^n \rangle &= \langle f, \bar{\psi}_0^n \rangle + n^{-1} M_1^n(t) + n^{-1} \sum_{i=1}^{S^n(t)} \left\langle f, \frac{\sum_{k=1}^{N_\alpha^n} \mu_k (1 - I_k^n(\theta_{S,i}^n-))\delta_{\mu_k} \mathbb{I}(X^n(\theta_{S,i}^n-) \le 0)}{\sum_{k=1}^{N_\alpha^n} \mu_k} \right\rangle \\
&\quad - n^{-1} M_2^n(t) - n^{-1} \sum_{i=1}^{A^n(t)} \left\langle f, \frac{\sum_{k=1}^{N_\alpha^n} h(\mu_k) I_k^n(\theta_{A,i}^n-)\delta_{\mu_k} \mathbb{I}(X^n(\theta_{A,i}^n-) \le 0)}{\sum_{k=1}^{N_\alpha^n} h(\mu_k) I_k^n(\theta_{S,i}^n-)} \right\rangle \\
&= \langle f, \bar{\psi}_0^n \rangle + n^{-1} M_1^n(t) + n^{-1} \frac{\langle f \times \iota, \sum_{k=1}^{N^n} \delta_{\mu_k} \rangle}{\sum_{k=1}^{N^n} \mu_k} \int_0^t \mathbb{I}(X^n(s-) \le 0) dS^n(s) \\
&\quad - n^{-1} \int_0^t \frac{\langle f \times \iota, \bar{\psi}_{s-}^n \rangle}{n^{-1} \sum_{k=1}^{N^n} \mu_k} \mathbb{I}(X^n(s-) \le 0) dS^n(s) - n^{-1} M_2^n(t) \\
&\quad - n^{-1} \int_0^t \frac{\langle f \times h, \bar{\psi}_{s-}^n \rangle}{\langle h, \bar{\psi}_{s-}^n \rangle} \mathbb{I}(X^n(s-) \le 0) dA^n(s) \\
&= \langle f, \bar{\psi}_0^n \rangle + \frac{M_1^n(t)}{n} + \frac{\langle f \times \iota, n^{-1} \sum_{k=1}^{N^n} \delta_{\mu_k} \rangle}{n^{-1} \sum_{k=1}^{N^n} \mu_k} \int_0^t \mathbb{I}\left(\frac{X^n(s-)}{n} \le 0 \right) d\left(\frac{S^n(s) - \sum_{k=1}^{N^n} \mu_k s}{n} \right) \\
&\quad + \langle f \times \iota, n^{-1} \sum_{k=1}^{N^n} \delta_{\mu_k} \rangle \int_0^t \mathbb{I}\left(\frac{X^n(s-)}{n} \le 0 \right) ds \\
&\quad - \int_0^t \frac{\langle f \times \iota, \bar{\psi}_{s-}^n \rangle}{n^{-1} \sum_{k=1}^{N^n} \mu_k} \mathbb{I}\left(\frac{X^n(s-)}{n} \le 0 \right) d\left(\frac{S^n(s) - \sum_{k=1}^{N^n} \mu_k s}{n} \right) - \int_0^t \langle f \times \iota, \bar{\psi}_{s-}^n \rangle \mathbb{I}\left(\frac{X^n(s-)}{n} \le 0 \right) ds \\
&\quad - \frac{M_2^n(t)}{n} - \int_0^t \frac{\langle f \times h, \bar{\psi}_{s-}^n \rangle}{\langle h, \bar{\psi}_{s-}^n \rangle} \mathbb{I}\left(\frac{X^n(s-)}{n} \le 0 \right) d\left(\frac{A^n(s) - \lambda^n s}{n} \right) \\
&\quad - \int_0^t \frac{\langle f \times h, \bar{\psi}_{s-}^n \rangle}{\langle h, \bar{\psi}_{s-}^n \rangle} \frac{\lambda^n}{n} \mathbb{I}\left(\frac{X^n(s-)}{n} \le 0 \right) ds. \qquad \text{(EC.9)}
\end{aligned}$$

Since $f \in C^b_{[\mu_{\min},\mu_{\max}]}[0,\infty)$, we assume that $f(\mu) \le K_f$ for all $\mu \in [\mu_{\min},\mu_{\max}]$. By our assumption, we know that $\sup_{0\le t\le T} |\langle f, \bar{\psi}^{n_k}_t\rangle - \langle f, \bar{\psi}_t\rangle| \to 0$ almost surely, along the subsequence $\{\bar{\psi}^{n_k}_t\}_{k=1}^{\infty}$. The martingales $M^n_1(t)$ and $M^n_2(t)$ can be written as

$$n^{-1}M^n_1(t) = n^{-1}\sum_{i=1}^{S^n(t)} (\langle f, \mathcal{U}^n_{S,i}\rangle - \mathbb{E}[\langle f, \mathcal{U}^n_{S,i}\rangle | \mathcal{F}_{v^n_i-}]),$$
$$\text{and} \quad n^{-1}M^n_2(t) = n^{-1}\sum_{i=1}^{A^n(t)} (\langle f, \mathcal{U}^n_{A,i}\rangle - \mathbb{E}[\langle f, \mathcal{U}^n_{A,i}\rangle | \mathcal{F}_{u^n_i-}]).$$

Since both are pure jump martingales, we can write the optional quadratic variation of these martingales as

$$[n^{-1}M^n_1(t)] = n^{-2}\sum_{i=1}^{S^n(t)} (\langle f, \mathcal{U}^n_{S,i}\rangle - \mathbb{E}[\langle f, \mathcal{U}^n_{S,i}\rangle | \mathcal{F}_{v^n_i-}])^2 \le \frac{K_f^2}{n^2}\sup_{0\le t\le T} S^n(t)$$
$$[n^{-1}M^n_2(t)] = n^{-2}\sum_{i=1}^{A^n(t)} (\langle f, \mathcal{U}^n_{A,i}\rangle - \mathbb{E}[\langle f, \mathcal{U}^n_{A,i}\rangle | \mathcal{F}_{u^n_i-}])^2 \le \frac{K_f^2}{n^2}\sup_{0\le t\le T} A^n(t)$$

which converges to 0 almost surely. We know that $h(\mu) > 0$ and it is continuous on the closed interval $[\mu_{\min},\mu_{\max}]$ and hence, there exists ϵ_h and K_h such that $0 < \epsilon_h \le h(\mu) \le K_h$ for all $\mu \in [\mu_{\min},\mu_{\max}]$. Also, Assumption 1 implies the existence of K_N and k_n such that $0 < \epsilon_N \le n^{-1}N^n \le K_N < \infty$. Hence, we have the following bounds:

$$\langle f\times\iota, \bar{\psi}^n_t\rangle \le \langle f\times\iota, n^{-1}\sum_{k=1}^{N^n_\alpha}\delta_{\mu_k}\rangle \le K_f K_N \mu_{\max}, \langle f\times h, \bar{\psi}^n_{s-}\rangle \le \langle f\times h, n^{-1}\sum_{k=1}^{N^n_\alpha}\delta_{\mu_k}\rangle \le K_f K_h K_N$$
$$\frac{\langle f\times\iota, \bar{\psi}^n_t\rangle}{n^{-1}\sum_{k=1}^{N^n_\alpha}\mu_k} \le \frac{\langle f\times\iota, n^{-1}\sum_{k=1}^{N^n_\alpha}\delta_{\mu_k}\rangle}{n^{-1}\sum_{k=1}^{N^n}\mu_k} \le \frac{K_f K_N \mu_{\max}}{\epsilon_N} \text{ and } \frac{\langle f\times h, \bar{\psi}^n_{s-}\rangle}{\langle h, \bar{\psi}^n_{s-}\rangle} \le \frac{K_f K_h}{\epsilon_h}.$$

These bounds along with $\sup_{0\le t\le T} n^{-1}|S^n(t) - \sum_{k=1}^{N^n}\mu_k t| \to 0$ and $\sup_{0\le t\le T} n^{-1}|A^n(t) - \lambda^n t| \to 0$ almost surely as $n\to\infty$, the third, fifth and the eighth terms on the right-hand side of (EC.9) converges to 0 almost surely. Using the dominated convergence theorem, we deduce that

$$\int_0^t \langle f\times\iota, \bar{\psi}^n_{s-}\rangle \mathbb{I}\left(\frac{X^n(s-)}{n} \le 0\right) ds \to \int_0^t \langle f\times\iota, \bar{\psi}_{s-}\rangle \mathbb{I}\left(\xi(s-) \le 0\right) ds$$
$$\int_0^t \frac{\langle f\times h, \bar{\psi}^n_{s-}\rangle}{\langle h, \bar{\psi}^n_{s-}\rangle}\frac{\lambda^n}{n}\mathbb{I}\left(\frac{X^n(s-)}{n} \le 0\right) ds \to \int_0^t \frac{\langle f\times h, \bar{\psi}_{s-}\rangle}{\langle h, \bar{\psi}_{s-}\rangle}\bar{\lambda}\mathbb{I}\left(\xi(s-) \le 0\right) ds.$$

Hence, our lemma follows. □

THEOREM 5. *When $\alpha=1$, the stationary limiting fairness measure $\eta_{1,\infty}$ under an h-random policy can be written as*

$$\eta_{1,\infty}(\mathbb{A})=\frac{\int_{\mathbb{A}}(1+L_F\tilde{h}(\mu))^{-1}dF(\mu)}{\int_{\mu_{\min}}^{\mu_{\max}}(1+L_F\tilde{h}(\mu))^{-1}dF(\mu)},\ \textit{for all } \mathbb{A}\in\mathcal{B}[\mu_{\min},\mu_{\max}] \tag{EC.10}$$

where $\tilde{h}(\mu)=h(\mu)/\mu$ and L_F is the unique solution of

$$\int_{\mu_{\min}}^{\mu_{\max}}\mu\frac{1+\beta}{\bar{\mu}_F(1+L_F\tilde{h}(\mu))}dF(\mu)=\beta. \tag{EC.11}$$

Proof. We know that $\xi_{1\infty}>0$, and hence any fixed point of (14) satisfies

$$\frac{\bar{\lambda}}{\bar{\mu}}(1+\beta)\langle f\times\iota,F\rangle\mathbb{I}(\xi_{1,t}\le 0)=\langle f\times\iota,\bar{\psi}_{1,t}\rangle\mathbb{I}(\xi_{1,t}\le 0)+\bar{\lambda}\frac{\langle f\times h,\bar{\psi}_{1,t}\rangle}{\langle h,\bar{\psi}_{1,t}\rangle}\mathbb{I}(\xi_{1,t}\le 0). \tag{EC.12}$$

As (12) implies that $\bar{\psi}_{1,\infty}$ is absolutely continuous, it possesses a Radon-Nikodym derivative $\bar{g}(\mu)$ with respect to F. Setting $f(\mu)=1$ for all μ, we have

$$\int_{\mu_{\min}}^{\mu_{\max}}\mu\bar{g}(\mu)dF(\mu)=\bar{\lambda}\beta. \tag{EC.13}$$

As the equation holds for any $f\in\mathbb{C}^b_{[\mu_{\min},\mu_{\max}]}[0,\infty)$, defining $c_g=\int_{\mu_{\min}}^{\mu_{\max}}h(\mu)\bar{g}(\mu)dF(\mu)$ to simplify the notation, for F-almost all μ

$$\mu\bar{g}(\mu)+\frac{\bar{\lambda}h(\mu)\bar{g}(\mu)}{c_g}=\frac{\bar{\lambda}}{\bar{\mu}_F}(1+\beta)\mu.$$

Defining $L_F=\bar{\lambda}/c_g$ and re-organizing terms, we obtain

$$\bar{g}(\mu)=\frac{\bar{\lambda}}{\bar{\mu}_F}(1+\beta)(1+L_F\tilde{h}(\mu))^{-1}. \tag{EC.14}$$

Plugging in (EC.13) we obtain (16), and since the integrand is decreasing in L_F then the L_F that satisfies this equation must be unique. This proves that for any limit point $\bar{\psi}_{1,t}$, we have $\langle f,\bar{\psi}_{1,t}\rangle\to\int_{\mu_{\min}}^{\mu_{\max}}f(\mu)\bar{g}(\mu)dF(\mu)$ as $t\to\infty$ for all $f\in\mathbb{C}^b_{[\mu_{\min},\mu_{\max}]}[0,\infty)$. Hence, using the fact that the indicator function of any Borel set $\mathbb{A}$ can be approximated by functions in $\mathbb{C}^b_{[\mu_{\min},\mu_{\max}]}[0,\infty)$, we have $\bar{\psi}_{1,t}(\mathbb{A})\to\int_{\mathbb{A}}g(\mu)dF(\mu)$. Using Theorem 3, this implies that $\bar{\psi}^n_\infty(\mathbb{A})\Rightarrow\int_{\mathbb{A}}g(\mu)dF(\mu)$ as $n\to\infty$. Plugging $\bar{\psi}_{1,t}$ into the definition of the fairness process, the result follows. □

LEMMA 4. *Define* $h_{\min} = \min_{\mu_{\min}\le\mu\le\mu_{\max}} h(\mu)$ *and* $h_{\max} = \max_{\mu_{\min}\le\mu\le\mu_{\max}} h(\mu)$ *and let* L_F *be the solution of* (16) *for some service rate distribution* F*. Then,*

$$\frac{\beta}{h_{\max}} \le L_F \le \frac{\beta}{h_{\min}}. \tag{EC.15}$$

Proof. Dividing both sides of (EC.13) by $\bar{\lambda}\beta$, this equation defines a probability measure $\bar{F}$ on $[\mu_{\min}, \mu_{\max}]$ and c_g in the proof of Theorem 5 can be written as $c_g = \bar{\lambda}\beta \int_{\mu_{\min}}^{\mu_{\max}} \tilde{h}(\mu)\frac{\mu\bar{g}(\mu)}{\bar{\lambda}\beta} dF(\mu)$. The integral is the expectation of $h(\tilde{\mu})$ where $\tilde{\mu}$ has distribution $\bar{F}$, and hence, is bounded from below by $h_{\min}$ and from above by $h_{\max}$. The result follows as $L_F = \bar{\lambda}/c_g$. □

COROLLARY 1. *Suppose that* $g(\mu)$ *satisfies* (17) *strictly. Then,* $g(\mu)$ *is the density function of* $\eta_{1,\infty}$ *with respect to* F *under the h-random policy where*

$$h(\mu) = \frac{\Big((1+\beta)(\bar{\mu}_F)^{-1} - \beta\langle \iota, \eta_{1,\infty}\rangle^{-1} g(\mu)\Big)\mu}{\beta\bar{\lambda}\langle \iota, \eta_{1,\infty}\rangle^{-1} g(\mu)}.$$

Proof. Let $\bar{g}(\mu)$ be the density function of $\bar{\psi}_{1,\infty}$ with respect to F. The definition of the limiting fairness measure and (10) implies that $\bar{g}(\mu) = \beta\bar{\lambda}\langle \iota, \eta_{1,\infty}\rangle^{-1} g(\mu)$. Plugging into the definition of c_g, we can see that for the proposed $h(\mu)$ we have

$$\begin{aligned} c_g &= \int_{\mu_{\min}}^{\mu_{\max}} \left(\frac{1+\beta}{\bar{\mu}_F} - \beta\langle \iota, \eta_{1,\infty}\rangle^{-1} g(\mu)\right)\mu dF(\mu) \\ &= \frac{1+\beta}{\bar{\mu}_F}\int_{\mu_{\min}}^{\mu_{\max}} \mu dF(\mu) - \beta\langle \iota, \eta_{1,\infty}\rangle^{-1}\int_{\mu_{\min}}^{\mu_{\max}} \mu g(\mu) dF(\mu) \\ &= 1. \end{aligned}$$

Hence, $L_F = \bar{\lambda}$. Now, plugging this in (EC.14), we get the desired result. □

EC.5. Proofs for Results Presented in Section 6

THEOREM 6. *If* $\sup_{\mu_{\min}\le\mu\le\mu_{\max}} \left|n^{1-\alpha}\mathbb{E}[I_k^n(\infty)|\tilde{\mu}_k^n = \mu] - \beta g(\mu)\bar{\lambda}^{\alpha-1}\bar{\mu}_F^{2-\alpha}\langle \iota, \eta_{\alpha,\infty}\rangle^{-1}\right| \to 0$*, then,*

1. *if* $n^{\alpha-1}f'(n^{\alpha-1}x) \to 0$ *for all* $x > 0$ *as* $n \to \infty$*, then for any* $\epsilon > 0$ *there exists an* N_ϵ *such that* $n > N_\epsilon$ *implies that the optimal strategy for any server* k *in the* n*th system is* $\tilde{\mu}_k^{n,*} \in [\tilde{\mu}_{\min,k}^n, \tilde{\mu}_{\min,k}^n + \epsilon)$*,*

2. *if the limiting stationary fairness measure is deterministic with a decreasing density $g(\mu)$, then for any $\epsilon > 0$ there exists an N_ϵ such that $n > N_\epsilon$ implies that the optimal strategy for any server k in the nth system is $\tilde{\mu}_k^{n,*} \in [\tilde{\mu}_{\min,k}^n, \tilde{\mu}_{\min,k}^n + \epsilon)$,*
3. *if $n^{\alpha-1} f'(n^{\alpha-1}x) \to \infty$ for all $x > 0$ as $n \to \infty$ and $g(\mu)$ is strictly increasing and concave, then for any $\epsilon > 0$ there exists an N_ϵ such that $n > N_\epsilon$ implies that the optimal strategy for any server k in the nth system is $\tilde{\mu}_k^{n,*} \in (\tilde{\mu}_{\max,k}^n - \epsilon, \tilde{\mu}_{\min,k}^n]$ with probability 1.*

Proof. To simplify notation, define $C = \beta \bar{\lambda}^{\alpha-1} \bar{\mu}_F^{2-\alpha} \langle \iota, \eta_\alpha \rangle^{-1}$. The uniform convergence assumption implies that for any $\rho > 0$, there exists an N_ρ such that for all $\mu_{\min} \le \mu \le \mu_{\max}$

$$\left| \mathbb{E}[I_k^n(\infty) | \tilde{\mu}_k^n = \mu] - n^{\alpha-1} C g(\mu) \right| \le n^{\alpha-1} \rho. \tag{EC.16}$$

First, we analyze the best response for server k for all large n for the approximating problem

$$\max_{\tilde{\mu}_{\min,k}^n \le \mu \le \tilde{\mu}_{\max,k}^n} f(n^{\alpha-1} C g(\mu)) - a_k^n c(\mu). \tag{EC.17}$$

Taking the derivative of the objective, we get

$$n^{\alpha-1} g'(\mu) f'(n^{\alpha-1} C g(\mu)) - \tilde{a}_k^n c'(\mu). \tag{EC.18}$$

First, we analyze part 1, i.e., the situation when $n^{\alpha-1} f'(n^{\alpha-1}x) \to 0$ for all $x > 0$. We know that $f(\mu)$ is concave and thus its derivative is decreasing. Since $g(\mu)$ is continuous, it attains its minimum at $\mu_g^* \in [\mu_{\min} \mu_{\max}]$. Then, we have $f'(n^{\alpha-1} C g(\mu_g^*)) \ge f'(n^{\alpha-1} C g(\mu))$ for all $\mu_{\min} \le \mu \le \mu_{\max}$. Hence, there exists an N_1 such that $n^{\alpha-1} f'(n^{\alpha-1} C g(\mu_g^*)) < \frac{a_{\min} c_{\min}}{2 C g'_{\max}}$ for all $n > N_1$ and hence, the minimizer of (EC.17) is $\mu_k^{n,*} = \tilde{\mu}_{\min,k}^n$ for all k and $n > N_1$. Using the gradient inequality for concave functions on $f(n^{\alpha-1}x)$, we get for any $\mu_{\min} \le \mu \le \mu_{\max}$ that

$$\begin{aligned} f(n^{\alpha-1} C g(\mu)) - \tilde{a}_k^n c(\mu) \le f(n^{\alpha-1} C g(\mu_{\min,k}^n)) + n^{\alpha-1} C f'(n^{\alpha-1} C g(\mu_g^*))(g(\mu) - g(\mu_{\min,k}^*)) \\ - \tilde{a}_k^n c(\mu_{\min,k}^n) - a_{\min} c_{\min} (\mu - \mu_{\min,k}^*). \end{aligned}$$

Hence, for any $n > N_1$, we have

$$f(n^{\alpha-1} C g(\mu)) - \tilde{a}_k^n c(\mu) \le f(n^{\alpha-1} C (g(\mu_{\min,k}^n)) - \tilde{a}_k^n c(\mu_{\min,k}^n) - \frac{a_{\min} c_{\min}}{2} (\mu - \mu_{\min,k}^*),$$

and for any $\mu > \mu^n_{\min,k} + 6\epsilon(a_{\min}c_{\min})^{-1}$,

$$f(n^{\alpha-1}C(g(\mu))) - \tilde{a}^n_k c(\mu) \leq f(n^{\alpha-1}C(g(\mu^n_{\min,k})) - \tilde{a}^n_k c(\mu^n_{\min,k}) - 3\epsilon. \tag{EC.19}$$

Again using (EC.16), there exists an N_2 such that $n > N_2$ implies

$$\mathbb{E}[I^n_k(\infty)|\tilde{\mu}^n_k = \mu] \geq n^{\alpha-1}\frac{Cg(\mu^*_g)}{2}.$$

Using the mean-value theorem and the concavity of $f(\cdot)$, for any $\mu_{\min} \leq \mu \leq \mu_{\max}$ and $n > N_2 \vee N_\rho$

$$|f(\mathbb{E}[I^n_k(\infty)|\tilde{\mu}^n_k = \mu]) - f(n^{\alpha-1}Cg(\mu))| \leq n^{\alpha-1}f'\left(\frac{n^{\alpha-1}Cg(\mu^*_g)}{2}\right)\rho.$$

Now, choosing N_3 such that $n > N_3$ implies

$$n^{\alpha-1}f'\left(\frac{n^{\alpha-1}Cg(\mu^*_g)}{2}\right) \leq \frac{\epsilon}{\rho},$$

and using (EC.19), we have

$$f(\mathbb{E}[I^n_k(\infty)|\tilde{\mu}^n_k = \mu]) - \tilde{a}^n_k c(\mu) \leq f(\mathbb{E}[I^n_k(\infty)|\tilde{\mu}^n_k = \tilde{\mu}^n_{\min,k}]) - \tilde{a}^n_k c(\tilde{\mu}^n_{\min,k}) - \epsilon,$$

for all $\mu > \mu^n_{\min,k} + 6\epsilon(a_{\min}c_{\min})^{-1}$ and $n > (N_1 \vee N_2 \vee N_3 \vee N_\rho)$.

Now, we consider the second case where $g(\mu)$ is decreasing. As (EC.18) is negative for all μ, the minimizer of (EC.17) is $\mu^{n,*}_k = \tilde{\mu}^n_{\min,k}$ for all n. Following the same steps as in part 1, part 2 follows.

Now, we analyze the part 3. As $f(\cdot)$ is concave, $f'(n^{\alpha-1}Cg(\mu)) \geq f'(n^{\alpha-1}Cg(\mu_{\max}))$ for all $\mu_{\min} \leq \mu \leq \mu_{\max}$. Using this and the convexity of $c(\mu)$, there exists an N_4, such that for $n > N_4$ the derivative in (EC.18) is positive for all $\mu_{\min} \leq \mu \leq \mu_{\max}$, which in turn implies the maximizer of (EC.17) is $\mu^{n,*}_k = \tilde{\mu}^n_{\max,k}$. Now, choosing N_1 such that $n > N_1$ implies $n^{\alpha-1}f'(n^{\alpha-1}Cg(\mu_{\max})) > \frac{2a_{\min}c_{\min}}{Cg'_{\min}}$ and following the same steps as in part 1, the theorem follows. □

To prove Theorem 7, we needd the following uniform integrability result:

LEMMA EC.2. *For any $\alpha > 1/2$, the collection of random variables $\{(\hat{I}^n_\alpha(\infty))^{-1}\mathbb{I}(I^n(\infty) > 0)\}_{n\in\mathbb{N}}$ is uniformly integrable.*

Proof. We need to prove that for any $\rho > 0$, there exists an $M > 0$ such that

$$\sup_n \mathbb{E}\left[\frac{n^\alpha}{I^n(\infty)}\mathbb{I}(I^n(\infty)>0)\mathbb{I}\left(\frac{n^\alpha}{I^n(\infty)} > M\right)\right] < \rho. \tag{EC.20}$$

For any $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{n^\alpha}{I^n(\infty)}\mathbb{I}(I^n(\infty)>0)\mathbb{I}\left(\frac{n^\alpha}{I^n(\infty)} > M\right)\right] \\
&\quad= \mathbb{E}\left[\frac{n^\alpha}{I^n(\infty)}\mathbb{I}(I^n(\infty)>0)\mathbb{I}\left(\frac{n^\alpha}{I^n(\infty)} > M\right)\mathbb{I}\left(\left|\sum_{k=1}^{N^n}\mu_k^n - N^n\bar{\mu}_F\right| \le \frac{\beta}{2}(\lambda^n)^\alpha(\bar{\mu}_F)^{1-\alpha}\right)\right] \\
&\qquad+ \mathbb{E}\left[\frac{n^\alpha}{I^n(\infty)}\mathbb{I}(I^n(\infty)>0)\mathbb{I}\left(\frac{n^\alpha}{I^n(\infty)} > M\right)\mathbb{I}\left(\left|\sum_{k=1}^{N^n}\mu_k^n - N^n\bar{\mu}_F\right| > \frac{\beta}{2}(\lambda^n)^\alpha(\bar{\mu}_F)^{1-\alpha}\right)\right] \\
&\quad\le \mathbb{E}\left[\frac{n^\alpha}{I^n(\infty)}\mathbb{I}(I^n(\infty)>0)\mathbb{I}\left(\frac{n^\alpha}{I^n(\infty)} > M\right) \mid \left|\sum_{k=1}^{N^n}\mu_k^n - N^n\bar{\mu}_F\right| \le \frac{\beta}{2}(\lambda^n)^\alpha(\bar{\mu}_F)^{1-\alpha}\right] \\
&\qquad+ n^\alpha\mathbb{P}\left(\left|\sum_{k=1}^{N^n}\mu_k^n - N^n\bar{\mu}_F\right| > \frac{\beta}{2}(\lambda^n)^\alpha(\bar{\mu}_F)^{1-\alpha}\right).
\end{aligned} \tag{EC.21}$$

First, we concentrate on the first term of the right-hand side. From Assumption 1, for any $\epsilon_1 > 0$, there exists an N_1 such that for any $n > N_1$, $n^{-1}\lambda^n \le \bar{\lambda} + \epsilon_1$. Hence, with $n > N_1$, we define

$$K_\alpha = \left\lceil \frac{\beta(\bar{\lambda}+\epsilon_1)^\alpha \bar{\mu}_F^{1-\alpha}}{4\mu_{\max}} \right\rceil.$$

Now, we consider a sequence of birth-death processes $\{Y^n(t)\}_{n\in\mathbb{N}}$. The birth rate in the nth system is uniformly equal to λ^n for any state. When the system is at state $Y^n(t) = i$, the death rate is given by

$$\nu_i = \begin{cases} i\mu_{\min} & \text{if } i \le N^n - n^\alpha K_\alpha \\ \lambda^n + \frac{\beta}{4}(\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha} & \text{if } i > N^n - n^\alpha K_\alpha \end{cases}.$$

Now, we show that the process $N^n - Y^n(t)$ is stochastically smaller than $\hat{I}^n(t)$. Remembering that due to non-idling property, we have $\hat{I}^n(t) = (\hat{X}^n_\alpha(t))^-$, we can write

$$\begin{aligned}
\hat{I}^n_\alpha = \left(\hat{X}^n(0)\right)^- &+ n^{-\alpha}\sum_{i=1}^{\tilde{S}^n(t)}\mathbb{I}\left(\left(\hat{X}^n(\theta^n_{S,i}-)\right)^+ = 0\right)\mathbb{I}\left(U^n_i \le p^n_i\left(\theta^n_{S_i}-\right)\right) \\
&- n^{-\alpha}\sum_{i=1}^{A^n(t)}\mathbb{I}\left(\left(\hat{X}^n(\theta^n_{A,i}-)\right)^- > 0\right).
\end{aligned}$$

Similarly, we can couple the birth-death process with the idleness process as

$$Y^n(t) = Y^n(0) - n^{-\alpha} \sum_{i=1}^{\tilde{S}^n(t)} \mathbb{I}\left(U_i^n \le \tilde{p}_i^n\left(\theta_{S_i}^n -\right)\right) + n^{-\alpha} \sum_{i=1}^{A^n(t)} \mathbb{I}\left(\left(\hat{Y}_1^n(\theta_{A,i}^n -)\right)^- > 0\right),$$

where $\tilde{p}_i^n\left(\theta_{S_i}^n -\right) = \frac{\nu_i}{\sum_{k=1}^{N^n} \tilde{\mu}_k^n}$. Suppose $N^n - Y^n(0) = I^n(0)$ and define $\vartheta^n = \inf\{t : N^n - Y^n(t) > I^n(t)\}$. Then, with probability 1, $N^n - Y^n(\vartheta-) = I^n(\vartheta-)$. As for any state where $I^n(t) = i$, $\frac{\nu_i}{\sum_{k=1}^{N^n} \tilde{\mu}_k^n} \le \frac{\sum_{k=1}^{N^n} \mu_k (1 - I_k^n(t))}{\sum_{k=1}^{N^n} \mu_k}$, if ϑ is an event epoch for $S^n(t)$, we have $N^n - Y^n(\vartheta-) \le N^n - Y^n(\vartheta-)$. If ϑ is an event epoch for A^n, then $N^n - Y^n(\vartheta-) \le N^n - Y^n(\vartheta-)$. This leads to a contradiction and we conclude that $N^n - Y^n(t)$ is stochastically less than $I^n(t)$. Now,

$$\begin{aligned}\mathbb{E}\left[\frac{n^\alpha}{I^n(\infty)} \mathbb{I}(I^n(\infty) > 0) \mathbb{I}\left(\frac{n^\alpha}{I^n(\infty)} > \frac{2}{K_\alpha}\right)\right] &\le n^\alpha \mathbb{P}\left(I^n(\infty) < \frac{K_\alpha n^\alpha}{2}\right) \\ &\le n^\alpha \mathbb{P}\left(Y^n(\infty) \ge N^n - \frac{K_\alpha n^\alpha}{2}\right).\end{aligned}$$

Now, as $\nu_i > \lambda^n$ for all $i \ge N^n - n^\alpha K_\alpha$, for each n, the birth-death process $Y^n(t)$ is positive recurrent and for all $i \ge N^n - n^\alpha K_\alpha$, we have

$$\begin{aligned}\mathbb{P}(Y^n(\infty) = i) &= \left(\frac{\lambda^n}{\lambda^n + \frac{\beta}{4}(\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha}}\right)^{i - N^n + n^\alpha K_\alpha} \times \mathbb{P}(Y^n(\infty) = N^n - n^\alpha K_\alpha) \\ &\le \left(\frac{\lambda^n}{\lambda^n + \frac{\beta}{4}(\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha}}\right)^{i - N^n + n^\alpha K_\alpha}.\end{aligned}$$

Hence,

$$n^\alpha \mathbb{P}\left(Y^n(\infty) \ge N^n - \frac{n^\alpha K}{2}\right) \le \left(\frac{\lambda^n}{\lambda^n + \frac{\beta}{4}(\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha}}\right)^{n^\alpha K_\alpha/2} \frac{(\lambda^n + \frac{\beta}{4}(\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha}) n^\alpha}{\frac{\beta}{4}(\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha}}.$$

When $\alpha > 1/2$, we see that the first-term on the right-hand side approaches 0 exponentially fast, whereas the second term increases linearly as $n \to \infty$. This implies that there exists an N_2 where the first term on the right-hand side of (EC.21) is less than $\rho/2$ for $n > N_2$.

To address the second term on the right-hand side of (EC.21), choose p to be the smallest even number such that $p\alpha > p/2 + \alpha$. Then, using Markov's inequality

$$n^\alpha \mathbb{P}\left(\left|\sum_{k=1}^{N^n} \mu_k^n - N^n \bar{\mu}_F\right| > \frac{\beta}{2}(\lambda^n)^\alpha (\bar{\mu}_F)^{1-\alpha}\right) \le \frac{n^\alpha \sum_{j=1}^{p/2} \binom{N^n}{2j} (\mu_{\max} - \mu_{\min})^p}{\left(\frac{\beta}{2}(\lambda^n)^\alpha (\bar{\mu}_F)^{1-\alpha}\right)^p}.$$

It is now easy to see that the numerator scales with $p/2+\alpha$ where as the denominator scales with $p\alpha$. Hence, the right-hand side approaches 0 as $n\to\infty$ and there exists an N_3 such that $n>N_3$ implies

$$n^{\alpha}\mathbb{P}\left(\left|\sum_{k=1}^{N^n}\mu_k^n - N^n\bar{\mu}_F\right| > \frac{\beta}{2}(\lambda^n)^{\alpha}(\bar{\mu}_F)^{1-\alpha}\right) < \rho/2.$$

Choosing $M=\max\{2/K_\alpha, N_1^\alpha, N_2^\alpha, N_3^\alpha\}$, (EC.20) holds and our result follows. □

THEOREM 7. *As $n\to\infty$, we have $\mathbb{E}[I_k^n(\infty)|\mu_k=m_k]\to\frac{m_k\langle h,\bar{\psi}_\infty\rangle}{m_k\langle h,\psi_\infty\rangle+\bar{\lambda}h(m_k)}$.*

Proof. We define $U_{A,k,i}^n=1$ if the ith event epoch of $A^n(t)$ corresponds to an arrival directly route to server k and 0 otherwise. Similarly, we define $U_{S,k,i}^n=1$ if the event epoch of $S^n(t)$ correspondes to an actual service completion at server k. Then, we have the following balance equation:

$$\begin{aligned}
I_k^n(t) &= I_k^n(0)+\sum_{i=1}^{S^n(t)}U_{S,k,i}^n-\sum_{i=1}^{A^n(t)}U_{A,k,i}^n\\
&= I_k^n(0)+\left(\sum_{i=1}^{S^n(t)}U_{S,k,i}^n-\mu_k\int_0^t(1-I_k^n(s-))ds\right)+\mu_k\int_0^t(1-I_k^n(s-))ds\\
&\quad-\left(\sum_{i=1}^{A^n(t)}U_{A,k,i}^n-\frac{\lambda^n}{n}\int_0^t\frac{h(\tilde{\mu}_k^n)I_k^n(s-)}{\langle h,\bar{\psi}_{1,s-}^n\rangle}ds\right)-\frac{\lambda^n}{n}\int_0^t\frac{h(\tilde{\mu}_k^n)I_k^n(s-)}{\langle h,\bar{\psi}_{1,s-}^n\rangle}ds.
\end{aligned}$$

The second and fourth terms on the right-hand side are Poisson martingales with initial value 0. Assuming that the system is in stationarity, taking the expectation of both sides and using Fubini's theorem, we have

$$\frac{\lambda^n h(\mu)}{n}\mathbb{E}\left[\frac{I_k^n(\infty)}{\langle h,\bar{\psi}_{1,\infty}^n\rangle}|\mu_k=\mu\right]+\mu\mathbb{E}[I_k^n(\infty)|\mu_k=\mu])=\mu.$$

Theorem 5 implies that $\langle h,\bar{\psi}_\infty^n\rangle\xrightarrow{p}\langle h,\bar{\psi}_\infty\rangle$ where the limit is deterministic. Using Lemma EC.2 and the notation $L_F=\bar{\lambda}\langle h,\bar{\psi}_\infty\rangle^{-1}$, we can find an N_ϵ such that $n>N_\epsilon$ implies

$$|L_F h(\mu)\mathbb{E}[I_k^n(\infty)|\mu_k=\mu]+\mu\mathbb{E}[I_k^n(\infty)|\mu_k=\mu])-\mu|=\epsilon.$$

Re-arranging the equation, the desired uniform convergence result holds. □

THEOREM 8. *For $1/2 \le \alpha \le 1$, under any idle-time order based policy, (20) holds and we have*

$$n^{1-\alpha}\mathbb{E}[I_k^n(\infty)|\tilde{\mu}_k^n = \mu] \to \mu\beta\bar{\lambda}^{\alpha-1}(\bar{\mu}_F^{-\alpha}\sigma_F^2 + \bar{\mu}_F^{2-\alpha}), \quad \textit{for all } 1/2 \le \alpha < 1, \quad \textit{and}$$
$$\mathbb{E}[I_k^n(\infty)|\tilde{\mu}_k^n = \mu] \to \frac{\beta\langle \iota, \eta_{1,\infty}\rangle^{-1}}{\mu\beta\langle \iota, \eta_{1,\infty}\rangle^{-1} + 1}, \quad \textit{for } \alpha = 1,$$

as $n \to \infty$, where $\eta_{1,\infty}$ is the limiting fairness measure for h-random policy with $h(\mu) = 1$ as given in Theorem 5.

Proof. Theorem 9 in Gopalakrishnan et al. (2016) ensures that all the idle-time order based policies have the same stationary distribution. Hence, it is enough to prove the theorem only for one idle-time order based policy for each α. Hence, the case when $\alpha = 1$ follows from our Theorem 7 when taking $h(\mu) = 1$ for all $\mu_{\min} \le \mu \le \mu_{\max}$. To prove the result for $1/2 \le \alpha < 1$, we concentrate on the longest-idle-server-first policy and follow a similar approach to the proofs of Lemma 4 and Theorem 6 in Büke and Qin (2023).

As in the proof of Lemma 1, we define the ith event epoch of the arrival process $A^n(t)$ as $\theta_{A,i}^n$ and the inter-arrival time between arrival $i-1$ and i as $u_i^n = \theta_{A,i}^n - \theta_{A,i-1}^n$. As the arrival process is a Poisson process, u_i^n are independent exponential random variables with rate λ^n. Similarly, we define the ith epoch of the potential service completion process, $S_k^n(t)$, of server k as $\theta_{S,k,i}^n$ and the inter-event time between $(i-1)$st and ith epoch as $v_{k,i}^n = \theta_{S,k,i}^n - \theta_{S,k,i-1}^n$ for all $i \in \mathbb{N}$, where $v_{k,i}^n$ are independent exponential random variables with rate $\tilde{\mu}_k^n$. We also denote the ith epoch of the actual service completion process, $D_k^n(t)$, of server k as $\bar{\theta}_{S,k,i}^n$. We define $\phi_{k,i}^n$ the idling time of server k after the ith server completion as $\bar{\phi}_{k,i}^n := \inf\{t - \bar{\theta}_{S,k,i}^n : I_k^n(t) = 0, t > \bar{\theta}_{S,k,i}^n\}$. For longest-idle-server-first, we have

$$\bar{\phi}_{k,i}^n = \sum_{j=2}^{I^n(\bar{\theta}_{S,k,i}^n)} u_{A(\bar{\theta}_{S,k,i}^n)+j}^n + (u_{A(\bar{\theta}_{S,k,i}^n)+1} - \bar{\theta}_{S,k,i}^n).$$

Using the expression as the start point, for the ith event epoch of the potential service completion process $S_k^n(t)$, we associate the potential idling time of server k, $\phi_{k,i}^n$, defined as

$$\phi_{k,i}^n = \sum_{j=2}^{I^n(\theta_{S,k,i}^n)} u_{A(\theta_{S,k,i}^n)+j}^n + (u_{A(\theta_{S,k,i}^n)+1} - \theta_{S,k,i}^n).$$

We also define $\phi^n_{-k} := \inf\{t \geq 0 : I^n_k(t) = 0\}$ as the first time server k is busy and

$$\phi^n_{-k} \leq \sum_{j=1}^{I^n(0)} u^n_j.$$

Similar to equation (14) in Büke and Qin (2023), we can write

$$\begin{aligned}\int_0^t I^n_k(s)ds &= (\phi^n_{-k} \wedge t) + \sum_{i=1}^{D^n_k(t)} (\bar{\phi}^n_{k,i} \wedge (t - \bar{\theta}^n_{S,k,i})) \\ &= (\phi^n_{-k} \wedge t) + \sum_{i=1}^{D^n_k(t)} \bar{\phi}^n_{k,i} - \sum_{i=1}^{D^n_k(t)} (\bar{\phi}^n_{k,i} - t + \bar{\theta}^n_{S,k,i})^+. \end{aligned} \tag{EC.22}$$

For the proofs in this section, we assume that the system starts in stationarity. Hence, for all $t \geq 0$, we have

$$\mathbb{E}[I^n_k(\infty)|\tilde{\mu}^n_k = \mu] = \mathbb{E}[I^n_k(t)|\tilde{\mu}^n_k = \mu].$$

We need the following lemma to prove our theorem.

LEMMA EC.3. *For any k, $I^n_k(\infty) \xrightarrow{p} 0$ and $\mathbb{E}[I^n_k(\infty)|\mu^n_k = \mu] \to 0$ as $n \to \infty$.*

Proof. We have

$$\begin{aligned}\mathbb{E}\Big[\int_0^t I^n_k(s)ds|\tilde{\mu}^n_k = \mu\Big] &= \int_0^t \mathbb{E}[I^n_k(s)|\tilde{\mu}^n_k = \mu]ds \\ &= \mathbb{E}[(\phi^n_{-k} \wedge t)|\tilde{\mu}^n_k = \mu] + \mathbb{E}\left[\sum_{i=1}^{D^n_k(t)} \bar{\phi}^n_{k,i} \wedge (t - \bar{\theta}^n_{S,k,i})|\tilde{\mu}^n_k = \mu\right] \\ &\leq \mathbb{E}[\phi^n_{-k}|\tilde{\mu}^n_k = \mu] + \mathbb{E}\left[\sum_{i=1}^{S^n_k(t)} \phi^n_{k,i}|\tilde{\mu}^n_k = \mu\right] \\ &\leq \frac{\mathbb{E}[I^n(\infty)]}{\lambda^n} \\ &\quad + \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{I}(\theta^n_{S,k,i} \leq t)\left(\sum_{j=2}^{I^n(\theta^n_{S,k,i})} u^n_{A(\theta^n_{S,k,i})+j} + (u_{A(\theta^n_{S,k,i})+1} - \theta^n_{S,k,i})\right)|\tilde{\mu}^n_k = \mu\right] \\ &\leq \frac{\mathbb{E}[I^n(\infty)]}{\lambda^n} \\ &\quad + \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbb{I}(\theta^n_{S,k,i} \leq t)\mathbb{E}\left[\left(\sum_{j=2}^{I^n(\theta^n_{S,k,i})} u^n_{A(\theta^n_{S,k,i})+j} + (u_{A(\theta^n_{S,k,i})+1} - \theta^n_{S,k,i})\right)|\mathcal{F}_{\theta^n_{S,k,i}-}\right]|\tilde{\mu}^n_k = \mu\right] \\ &\leq \frac{\mathbb{E}[I^n(\infty)]}{\lambda^n} + \mathbb{E}\left[\sum_{i=1}^{S^n_k(t)} \frac{(I^n(\theta^n_{S,k,i}-)+1)}{\lambda^n}|\tilde{\mu}^n_k = \mu\right] \\ &\leq \frac{\mathbb{E}[I^n(\infty)]}{\lambda^n} + \frac{\mu t\mathbb{E}[I^n(\infty)+1|\tilde{\mu}^n_k = \mu]}{\lambda^n} \to 0.\end{aligned}$$

As $\mu \leq \mu_{\max}$, the convergence is uniform. The convergence in probability follows using Markov's inequality, which concludes the proof. $\square$

Our stationarity assumption combined with (EC.22) implies that

$$\begin{aligned} n^{1-\alpha}\mathbb{E}[I_k^n(\infty)|\tilde{\mu}_k^n = \mu] &= \frac{n^{1-\alpha}}{t}\int_0^t \mathbb{E}[I_k^n(s)|\tilde{\mu}_k^n = \mu]ds \\ &= \frac{n^{1-\alpha}}{t}\mathbb{E}[(\phi_{-k}^n \wedge t)|\tilde{\mu}_k^n = \mu] + \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{D_k^n(t)} \bar{\phi}_{k,i}^n|\tilde{\mu}_k^n = \mu\right] \\ &\quad - \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{D_k^n(t)} (\bar{\phi}_{k,i}^n - t + \bar{\theta}_{S,k,i}^n)^+|\tilde{\mu}_k^n = \mu\right]. \end{aligned}$$

Hence, for all $t \geq 0$, we can bound $n^{1-\alpha}\mathbb{E}[I_k^n(\infty)|\tilde{\mu}_k^n = \mu]$ as

$$n^{1-\alpha}\mathbb{E}[I_k^n(\infty)|\tilde{\mu}_k^n = \mu] \geq \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{D_k^n(t)} \phi_i^n|\tilde{\mu}_k^n = \mu\right] - \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{D_k^n(t)} (\phi_i^n - (t - \bar{\theta}_{S,k,i}^n))^+|\tilde{\mu}_k^n = \mu\right] \tag{EC.23}$$

$$n^{1-\alpha}\mathbb{E}[I_k^n(\infty)|\tilde{\mu}_k^n = \mu] \leq \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{D_k^n(t)} \phi_i^n|\tilde{\mu}_k^n = \mu\right] + \frac{n^{1-\alpha}}{t}\mathbb{E}[\phi_{-k}^n|\tilde{\mu}_k^n = \mu]. \tag{EC.24}$$

The first terms on the right-hand sides of (EC.23) and (EC.24) are the same and we first concentrate on this. We have then

$$\begin{aligned} \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{D_k^n(t)} \phi_i^n|\tilde{\mu}_k^n = \mu\right] &= \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{S_k^n(t)}\sum_{j=1}^{I^n(\bar{\theta}_{S,k,i}^n)} u^n_{A(\bar{\theta}_{S,k,i}^n)+j}|\tilde{\mu}_k^n = \mu\right] \\ &= \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{S_k^n(t)}(1 - I_k^n(\theta_{S,k,i}^n-))\sum_{j=1}^{I^n(\theta_{S,k,i}^n)} u^n_{A(\theta_{S,k,i}^n)+j}|\tilde{\mu}_k^n = \mu\right] \\ &= \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{S_k^n(t)}\sum_{j=1}^{I^n(\theta_{S,k,i}^n)} u^n_{A(\theta_{S,k,i}^n)+j}|\tilde{\mu}_k^n = \mu\right] \\ &\quad - \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{S_k^n(t)} I_k^n(\theta_{S,k,i}^n-)\sum_{j=1}^{I^n(\theta_{S,k,i}^n)} u^n_{A(\theta_{S,k,i}^n)+j}|\tilde{\mu}_k^n = \mu\right] \\ &= \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\int_0^t \sum_{j=1}^{I^n(s-)+1} u^n_{A(s-)+j} dS_k^n(s)|\tilde{\mu}_k^n = \mu\right] \\ &\quad - \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\int_0^t \left(I_k^n(s-)\sum_{j=1}^{I^n(s-)} u^n_{A(s-)+j}\right) dS_k^n(s)|\tilde{\mu}_k^n = \mu\right] \end{aligned}$$

$$= \frac{n\mu}{\lambda^n t}\mathbb{E}\left[\int_0^t (\hat{I}^n_\alpha(s-) + n^{-\alpha})ds|\tilde{\mu}^n_k = \mu\right] - \frac{n\mu}{\lambda^n t}\mathbb{E}\left[\int_0^t I^n_k(s-)(\hat{I}^n_\alpha(s-) + n^{-\alpha})ds|\tilde{\mu}^n_k = \mu\right].$$

Using the stochastic boundedness of $\hat{I}^n_\alpha$ and Lemma EC.3, the second term converges to 0. Using the stationarity assumption,

$$\sup_{\mu_{\min}\le\mu\le\mu_{\max}}\left|\frac{n\mu}{\lambda^n t}\mathbb{E}\left[\int_0^t (\hat{I}^n_\alpha(s-) + n^{-\alpha})ds|\tilde{\mu}^n_k = \mu\right] - \frac{\mu}{\lambda}\mathbb{E}[\hat{I}^n_\alpha\infty)|\tilde{\mu}^n_k = \mu]\right| \to 0.$$

Next, we concentrate on the second term on the right-hand side of (EC.23). Each summand corresponds to the remaining idling time from the service completion at $\bar{\theta}^n_{S,k,i}$. This implies only the summand corresponding to the last service completion is positive and

$$(\phi^n_i - (t - \bar{\theta}^n_{S,k,i}))^+ \le \sum_{j=1}^{I^n(t)} u_{A^n(t)+j}.$$

Hence,

$$\begin{aligned}\frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{i=1}^{D^n_k(t)}(\phi^n_i - (t-\bar{\theta}^n_{S,k,i}))^+|\mu^n_k = \mu\right] &\le \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{j=1}^{I^n(t)} u_{A^n(t)+j}|\mu^n_k = \mu\right]\\ &= \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{I^n(t)} u_{A^n(t)+j}|\mathcal{F}_t\right]|\mu^n_k = \mu\right]\\ &= \frac{n}{\lambda^n t}\mathbb{E}\left[\hat{I}^n_\alpha(t)|\mu^n_k = \mu\right]\\ &= \frac{n}{\lambda^n t}\mathbb{E}\left[\hat{I}^n_\alpha(\infty)|\mu^n_k = \mu\right].\end{aligned}$$

Similarly, we can bound the second term on the right-hand side of (EC.24) as

$$\frac{n^{1-\alpha}}{t}\mathbb{E}[\phi^n_{-k}|\tilde{\mu}^n_k = \mu] \le \frac{n^{1-\alpha}}{t}\mathbb{E}\left[\sum_{j=1}^{I^n(\infty)} u^n_j|\tilde{\mu}^n_k = \mu\right] = \frac{n}{\lambda^n t}\mathbb{E}\left[\hat{I}^n_\alpha(\infty)|\mu^n_k = \mu\right].$$

Taking t large enough, both terms can be made arbitrarily small and hence (EC.23) and (EC.24) implies

$$\sup_{\mu_{\min}\le\mu\le\mu_{\max}}\left|n^{1-\alpha}\mathbb{E}[I^n_k(\infty)|\tilde{\mu}^n_k = \mu] - \frac{\mu}{\lambda}\mathbb{E}[\hat{I}_\alpha(\infty)|\tilde{\mu}^n_k = \mu]\right| \to 0.$$

Now, plugging in the expected value to be $\beta\bar{\lambda}^\alpha\bar{\mu}_F^{1-\alpha}\langle\iota,\eta_{\alpha,\infty}\rangle$ with $\langle\iota,\eta_{\alpha,\infty}\rangle = \bar{\mu}_F^{-1}(\sigma_F^2 + \bar{\mu}_F^2)$, the result follows. □

LEMMA 5. *Under Assumption 4, the limiting utility function $U_k(\mu)$ is concave.*

Proof. The utility of idleness function is increasing and concave. Hence, to prove the concavity of the first part, it is enough to prove that the idleness experienced by a server with rate μ, $(1+L_{F^{(0)}}\tilde{h}(\mu))^{-1}$, is concave (see, e.g., page 86 in Boyd and Vandenberghe, 2004). Taking the second derivative yields

$$\frac{d^2(1+L_{F^{(0)}}\tilde{h}(\mu))^{-1}}{d\mu^2} = \frac{L_{F^{(0)}}(L_{F^{(0)}}(2\tilde{h}'(\mu)^2 - \tilde{h}(\mu)\tilde{h}''(\mu)) - \tilde{h}''(\mu))}{(1+L_{F^{(0)}}\tilde{h}(\mu))^3}.$$

Using the convexity of $\tilde{h}(\mu)$, we can conclude that the second derivative is negative for the idleness experienced by a server with rate μ is concave under (24) and hence, the result follows. □

THEOREM 9. *If F is an equilibrium service rate distribution, the distribution function F has the form in (27), where L_F is the solution of*

$$\int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1-\beta L_F \tilde{h}(\mu)}{1+L_F\tilde{h}(\mu)} dF(\mu|L_F) = 0.$$

Proof. As L_F uniquely characterizes the distribution of idleness among servers, our result will follow if the distribution of service rates resulting from L_F also yields the same L_F for the h-random policy under consideration. Plugging in (16), this can be written as

$$\int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1+\beta}{\bar{\mu}_F(1+L_F\tilde{h}(\mu))} dF(\mu|L_F) = \beta.$$

Taking $\bar{\mu}_F$ to the right-hand side and writing it as an integral, we get

$$\int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1+\beta}{1+L_F\tilde{h}(\mu)} dF(\mu) = \beta \int_{\mu_{\min}}^{\mu_{\max}} \mu dF(\mu|L_F).$$

Re-arranging the terms, we get the desired result. □

PROPOSITION 2. *Let the distribution of $\tilde{a}_k^n$, F_a, be continuous. Then, Equation (28) has a solution in the interval $[1/(\beta\tilde{h}(\mu_{\min})), 1/(\beta\tilde{h}(\mu_{\max}))]$ and has no solution outside of this interval.*

Proof. As a_k^n is a continuous random variable, using the continuity of $C(\mu, L_{F^{(0)}})$ we can conclude that if $L_i \to L_{F^{(0)}}$, we have $F(\mu|L_i) \to F(\mu|L_{F^{(0)}})$ for all $\mu \in [\mu_{\min}, \mu_{\max}]$. Hence,

$$\begin{aligned}
&\left| \int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1-\beta L_i \tilde{h}(\mu)}{1+L_i \tilde{h}(\mu)} dF(\mu|L_i) - \int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1-\beta L_{F^{(0)}} \tilde{h}(\mu)}{1+L_{F^{(0)}} \tilde{h}(\mu)} dF(\mu|L_{F^{(0)}}) \right| \\
&\qquad \leq \left| \int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1-\beta L_i \tilde{h}(\mu)}{1+L_i \tilde{h}(\mu)} dF(\mu|L_i) - \int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1-\beta L_{F^{(0)}} \tilde{h}(\mu)}{1+L_{F^{(0)}} \tilde{h}(\mu)} dF(\mu|L_{F_i}) \right| \\
&\qquad\quad + \left| \int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1-\beta L_{F^{(0)}} \tilde{h}(\mu)}{1+L_{F^{(0)}} \tilde{h}(\mu)} dF(\mu|L_i) - \int_{\mu_{\min}}^{\mu_{\max}} \mu \frac{1-\beta L_{F^{(0)}} \tilde{h}(\mu)}{1+L_{F^{(0)}} \tilde{h}(\mu)} dF(\mu|L_{F^{(0)}}) \right|.
\end{aligned}$$

As $i \to \infty$, the first term on the right-hand side converges due to continuity of the integrand, and the second term converges using the definition of weak convergence, which implies the left-hand side of (28) is continuous with respect to L_F. Now, if $L_F = 1/(\beta \tilde{h}(\mu_{\min}))$, the integrand of (28) is non-negative for all $\mu \in [\mu_{\min}, \mu_{\max}]$ and $\tilde{h}(\mu)$ being strictly decreasing, implies that the integral is positive for all $L_F \leq 1/(\beta \tilde{h}(\mu_{\min}))$. Similarly, if $L_F = 1/(\beta \tilde{h}(\mu_{\max}))$, the integrand is non-positive and the integral is negative. Using the intermediate value theorem, the result follows. □