Utility maximizing load balancing policies
Online companion

Diego Goldsztajn, Sem C. Borst
Eindhoven University of Technology, 5612 AZ Eindhoven, Netherlands, d.e.goldsztajn@tue.nl, s.c.borst@tue.nl
Johan S.H. van Leeuwaarden
Tilburg University, 5037 AB Tilburg, Netherlands, j.s.h.vanleeuwaarden@uvt.nl

This online supplement contains the proofs of some of the technical results stated in the body of the paper. Specifically, various auxiliary results are proved in Appendix A, a relative compactness result is established in Appendix B and several results pertaining to the limiting behavior of SLTA are proved in Appendix C.

Appendix A: Auxiliary results

Construction of Section 5.1.2. First we define the functions η_k. For this purpose, suppose that $q \in Q_n$ and $r \geq 1$ are the current values of $\boldsymbol{q}_n$ and $\boldsymbol{r}_n$, respectively. Then

$$\mathcal{G}(q,r) := \begin{cases} \{(i,j) \in \mathcal{I}_+ : (i,j) \rhd (i_r, j_r)\} & \text{if} \quad r > 1, \\ \emptyset & \text{if} \quad r = 1, \end{cases}$$

is the set of the coordinates that could belong to a server pool with a green token; i.e., a server pool of class i with exactly $j-1$ tasks has a green token if and only if $(i,j) \in \mathcal{G}(q,r)$. We partition the latter set of coordinates as follows:

$$\mathcal{G}_1(q,r) := \mathcal{G}(q,r) \cap \{(i,j) \in \mathcal{I}_+ : i \neq i_{r-1}\},$$
$$\mathcal{G}_2(q,r) := \mathcal{G}(q,r) \cap \{(i,j) \in \mathcal{I}_+ : i = i_{r-1}\}.$$

Let $\delta_q(i,j) := q(i,j-1) - q(i,j)$ denote the fraction of server pools which are of class i and have exactly $j-1$ tasks when the occupancy state is q, and define

$$I(q,i,j) := \left[1 - \sum_{(k,l) \trianglelefteq (i,j)} \delta_q(k,l), 1 - \sum_{(k,l) \triangleleft (i,j)} \delta_q(k,l) \right) \quad \text{for all} \quad (i,j) \in \mathcal{I}_+.$$

These intervals form a partition of $[0,1)$ such that the length of $I(q,i,j)$ is the fraction of server pools which are of class i and have exactly $j-1$ tasks. The length of $I(q,i,j)$ is equal to the probability of picking a server pool with coordinates (i,j) uniformly at random.

Note that $G_1(q,r) := \sum_{(k,l) \in \mathcal{G}_1(q,r)} \delta_q(k,l)$ is the fraction of server pools of class $i \neq i_{r-1}$ with a green token. If the latter quantity is positive, then we let

$$G_1(q,r,i,j) := \left[\frac{1 - \sum_{(k,l) \trianglelefteq (i,j), (k,l) \in \mathcal{G}_1(q,r)} \delta_q(k,l)}{G_1(q,r)}, \frac{1 - \sum_{(k,l) \triangleleft (i,j), (k,l) \in \mathcal{G}_1(q,r)} \delta_q(k,l)}{G_1(q,r)} \right).$$

These intervals yield another partition of $[0,1)$, where the length of $G_1(q,r,i,j)$ is the fraction of server pools which are of class i and have exactly $j-1$ tasks, but only among those server pools which are of class $i \neq i_{r-1}$ and have a green token. If $G_1(q,r)=0$, then we define $G_1(q,r,i,j) := \emptyset$ for all $(i,j) \in \mathcal{I}_+$. Sets $G_2(q,r,i,j)$ are defined similarly.

The functions η_k are defined as follows:

$$\eta_k(q,r,i,j) := \begin{cases} \mathbb{1}_{\{U_k \in G_1(q,r,i,j)\}} & \text{if} \quad (i,j) \in \mathcal{G}_1(q,r), \\ \mathbb{1}_{\{G_1(q,r)=0, U_k \in G_2(q,r,i,j)\}} & \text{if} \quad (i,j) \in \mathcal{G}_2(q,r), \\ \mathbb{1}_{\{G_1(q,r)=G_2(q,r)=0, I(q,i_r,j_r) \neq \emptyset\}} & \text{if} \quad (i,j) = (i_r,j_r), \\ \mathbb{1}_{\{G_1(q,r)=G_2(q,r)=0, I(q,i_r,j_r)=\emptyset, U_k \in I(q,i,j)\}} & \text{otherwise.} \end{cases}$$

Observe that $\eta_k(q,r,i,j) \in \{0,1\}$ for all k and $(i,j) \in \mathcal{I}_+$, and that for a fixed k there exists a unique $(i,j) \in \mathcal{I}_+$ such that $\eta_k(q,r,i,j)=1$. Also, if (q,r) is the value of $(\boldsymbol{q}_n, \boldsymbol{r}_n)$ when the k^{th} task arrives to the system, then $\eta_k(q,r,i,j)=1$ if and only if this task has to be sent to a server pool of class i with exactly $j-1$ tasks.

It only remains to define the sets $I_{n,k}$ and $D_{n,k}$ that appear in (20b). The former is the set of those $\omega \in \Omega$ that satisfy:

$$\begin{aligned} &\boldsymbol{q}\left(\tau^-_{n,k}(\omega), i, j\right) = \alpha_n(i) \quad \text{for all} \quad (i,j) \rhd \left(i_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)}, j_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)}\right), \\ &n\boldsymbol{q}\left(\tau^-_{n,k}(\omega), i_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)}, j_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)}\right) \geq n\alpha_n\left(i_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)}\right) - 1. \end{aligned}$$

If the state of the system is $(\boldsymbol{q}, \boldsymbol{r})$, then the first condition means that there are no green tokens right before the k^{th} task arrives, and the second condition means that the number of yellow tokens is at most one. Finally, $D_{n,k}$ is the set of those $\omega \in \Omega$ such that

$$\begin{aligned} &\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right) > 1, \\ &n - \sum_{i=1}^{m} n\boldsymbol{q}\left(\tau^-_{n,k}(\omega), i, \ell_i\left(\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)\right)\right) \geq n\beta_n, \\ &\boldsymbol{q}\left(\tau^-_{n,k}(\omega), i_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)-1}, j_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)-1}\right) < \alpha_n\left(i_{\boldsymbol{r}\left(\tau^-_{n,k}(\omega)\right)-1}\right). \end{aligned}$$

The second condition means that the number of green tokens is larger than or equal to $n\beta_n$ right before the k^{th} task arrives. The last condition means that at least one of these tokens is of class $i_{\boldsymbol{r}-1}$. □

Proof of Lemma 2. Fix $T \geq 0$ and let $M_T := \lambda T(\mathrm{e}-1)+1$. The set

$$E_T := \left\{\omega \in \Omega : \mathcal{N}^{\lambda}_n(\omega, T) \leq nM_T \text{ for all large enough } n\right\}.$$

has probability one as a consequence of the Borel-Cantelli lemma and the Chernoff bound

$$\mathbb{P}\left(\mathcal{N}^{\lambda}_n(T) > nM_T\right) \leq \frac{\mathrm{e}^{\lambda T(\mathrm{e}-1)n}}{\mathrm{e}^{M_T n}} = \mathrm{e}^{-n}.$$

As a result, the following set also has probability one:

$$\Gamma := \bigcap_{T \in \mathbb{N}} \left(E_T \cap \Gamma_\infty\right).$$

Note that for each $\omega \in \Gamma$ and $T \geq 0$ there exists $n^1_T(\omega)$ such that $\mathcal{N}^{\lambda}_n(\omega, T) \leq nM_T$ for all $n \geq n^1_T(\omega)$. Fix $T \geq 0$, $\omega \in \Gamma$ and $0 < \varepsilon < \min\{\alpha(i) : 1 \leq i \leq m\}$. Next we establish that there exist $j_T(\omega)$ and $n_T(\omega)$ such that

$$\boldsymbol{q}_n\left(\omega, t, i, j_T(\omega)\right) \leq \varepsilon < \alpha_n(i) \quad \text{for all} \quad t \in [0,T], \quad 1 \leq i \leq m \quad \text{and} \quad n \geq n_T(\omega). \tag{51}$$

If the load balancing policy is JLMU, then (8) and the above statement imply that $\sigma_j\left(\boldsymbol{q}_n(\omega,t)\right) \le j_T(\omega)$ for all $t\in[0,T]$ and $n \ge n_T(\omega)$, which in turn implies that the claim of the lemma holds. Suppose instead that SLTA is used and let

$$r_T(\omega) := \min\left\{k \ge 1 : (i_k, j_k) \trianglelefteq (i, j_T(\omega)) \text{ for all } i\right\}.$$

It follows from (23) that $\boldsymbol{r}_n(\omega,t) \le \max\{r_T(\omega), R(\omega)\} =: R_T(\omega)$ for all $t\in[0,T]$ and $n \ge n_T(\omega)$. Furthermore, the latter property and Remark 1 imply that $\mathcal{A}_n(\omega,t,i,j) = 0$ for all $t\in[0,T]$, $n \ge n_T(\omega)$ and $(i,j) \lhd (i_{R_T(\omega)+1}, j_{R_T(\omega)+1})$. Therefore, the claim of the lemma also holds for SLTA.

The following arguments apply both for JLMU and SLTA. Below we omit ω from the notation for brevity. Since $q_0 \in \ell_1$, there exists j_0 such that $q_0(i,j_0) \le \varepsilon/4$ for all i, and by (9) and (22a), there exists $n_T \ge n_T^1$ such that

$$\boldsymbol{q}_n(0,i,j_0) \le \frac{\varepsilon}{2} < \varepsilon < \alpha_n(i) \quad \text{for all} \quad 1 \le i \le m \quad \text{and} \quad n \ge n_T.$$

Define $j_T := j_0 + \lceil 2M_T/\varepsilon \rceil - 1$. Also, fix $t\in[0,T]$, $1\le i\le m$ and $n \ge n_T$. We have

$$\sum_{j=j_0}^{j_T} n\left[\boldsymbol{q}_n(t,i,j) - \frac{\varepsilon}{2}\right] \le \sum_{j=j_0}^{j_T} n\left[\boldsymbol{q}_n(t,i,j) - \boldsymbol{q}_n(0,i,j)\right] \le \sum_{j=j_0}^{j_T} n\mathcal{A}_n(t,i,j) \le nM_T.$$

For the first inequality, observe that $\boldsymbol{q}_n(0,i,j) \le \boldsymbol{q}_n(0,i,j_0) \le \varepsilon/2$ for all $j \ge j_0$, and for the last inequality, note that $\mathcal{N}_n^\lambda(t) \le \mathcal{N}_n^\lambda(T) \le nM_T$ because $n \ge n_T^1$. Since $\boldsymbol{q}_n(t,i,j)$ is non-increasing in j, we conclude that

$$\boldsymbol{q}_n(t,i,j_T) \le \frac{1}{j_T - j_0 + 1}\sum_{j=j_0}^{j_T} \boldsymbol{q}_n(t,i,j) \le \frac{M_T + \frac{\varepsilon}{2}(j_T - j_0 + 1)}{j_T - j_0 + 1} < \varepsilon < \alpha_n(i).$$

This completes the proof. □

Proof of Lemma 3. Consider a function $\boldsymbol{x} \in D[0,\infty)$ such that $\boldsymbol{x}(0) \ge 0$. Then there exist unique $\boldsymbol{y}, \boldsymbol{z} \in D[0,\infty)$ such that the following statements hold.

(a') $\boldsymbol{z}(t) = \boldsymbol{x}(t) + \boldsymbol{y}(t) \ge 0$ for all $t \ge 0$.

(b') $\boldsymbol{y}$ is non-decreasing and $\boldsymbol{y}(0) = 0$.

(c') $\boldsymbol{y}$ is flat off $\{t \ge 0 : \boldsymbol{z}(t) = 0\}$, i.e., $\dot{\boldsymbol{y}}(t)\mathbb{1}_{\{\boldsymbol{z}(t)>0\}} = 0$ almost everywhere.

The map (Ψ, Φ) such that $\Psi(\boldsymbol{x}) = \boldsymbol{y}$ and $\Phi(\boldsymbol{x}) = \boldsymbol{z}$ is called the one-dimensional Skorokhod mapping with lower reflecting barrier at zero and satisfies

$$\Psi(\boldsymbol{x})(t) = \sup_{s\in[0,t]}\left[-\boldsymbol{x}(s)\right]^+ \quad \text{and} \quad \Phi(\boldsymbol{x})(t) = \boldsymbol{x}(t) + \Psi(\boldsymbol{x})(t).$$

Also, if $\boldsymbol{x}, \boldsymbol{y} \in D[0,\infty)$ satisfy $\boldsymbol{x}(0), \boldsymbol{y}(0) \ge 0$, then for each $T \ge 0$, we have

$$||\Psi(\boldsymbol{x}) - \Psi(\boldsymbol{y})||_T \le ||\boldsymbol{x} - \boldsymbol{y}||_T \quad \text{and} \quad ||\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y})||_T \le 2\,||\boldsymbol{x} - \boldsymbol{y}||_T.$$

The latter definition and properties of the Skorokhod mapping with lower reflecting barrier at zero can be found in (Chen and Yao 2013, Theorem 6.1) and (Karatzas and Shreve 2014, Lemma 6.14).

Suppose $\boldsymbol{x}$ is as in the statement of the lemma. Define $\boldsymbol{y}$ and $\boldsymbol{z}$ as in (24); i.e.,

$$\boldsymbol{y}(t) := \sup_{s\in[0,T]}\left[\boldsymbol{x}(s) - \alpha\right]^+ = \Psi(\alpha - \boldsymbol{x})(t) \quad \text{and} \quad \boldsymbol{z}(t) := \boldsymbol{x}(t) - \boldsymbol{y}(t) = \alpha - \Phi(\alpha - \boldsymbol{x})(t).$$

Properties (b') and (c') imply (b) and (c). Also, (a) holds since $\boldsymbol{z} = \alpha - \Phi(\alpha - \boldsymbol{x}) \leq \alpha$ by (a') and $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y}$ by definition.

Conversely, suppose that $\boldsymbol{y}$ and $\boldsymbol{z}$ satisfy (a)-(c). Then $\alpha - \boldsymbol{z} = \alpha - \boldsymbol{x} + \boldsymbol{y} \geq 0$ and $\boldsymbol{y}$ satisfies (b') and (c') with $\boldsymbol{z}$ replaced by $\alpha - \boldsymbol{z}$, hence we conclude that $\boldsymbol{y} = \Psi(\alpha - \boldsymbol{x})$ and $\alpha - \boldsymbol{z} = \Phi(\alpha - \boldsymbol{x})$. Therefore, $\boldsymbol{y}$ and $\boldsymbol{z}$ are as in (24).

The Lipschitz properties of Ψ_α and Φ_α follow from the Lipschitz properties of Ψ and Φ and the identities $\Psi_\alpha(\boldsymbol{x}) = \Psi(\alpha - \boldsymbol{x})$ and $\Phi_\alpha(\boldsymbol{x}) = \alpha - \Phi(\alpha - \boldsymbol{x})$. □

Proof of Lemma 8. By assumption and Tonelli's theorem,

$$\begin{aligned}\sum_{x,y\in S_n} \pi_n(x)\,|A_n(x,y)|\,f(y) &= \sum_{x\in S_n} \pi_n(x) \sum_{y\in S_n} |A_n(x,y)|\,f(y)\\ &= E\left[\sum_{y\in S_n} |A_n(x_n,y)|\,f(y)\right] < \infty.\end{aligned}$$

Therefore, we may use Fubini's theorem, which yields

$$0 = \sum_{y\in S_n} f(y) \sum_{x\in S_n} \pi_n(x) A_n(x,y) = \sum_{x\in S_n} \pi_n(x) \sum_{y\in S_n} A_n(x,y) f(y) = E\left[A_n f(x_n)\right].$$

The first equality follows from the stationarity of π_n. □

Appendix B: Relative compactness

Consider the metric

$$d(x,y) := \sum_{(i,j)\in\mathcal{I}} \frac{\min\{|x(i,j) - y(i,j)|, 1\}}{2^j} \quad \text{for all} \quad x, y \in \mathbb{R}^{\mathcal{I}},$$

which induces the product topology on $\mathbb{R}^{\mathcal{I}}$. For each $T \geq 0$, let $D_{\mathbb{R}^{\mathcal{I}}}[0,T]$ denote the space of all càdlàg functions on $[0,T]$ with values in $\mathbb{R}^{\mathcal{I}}$ and equip this space with the uniform metric, defined by

$$\varrho_T(\boldsymbol{x},\boldsymbol{y}) := \sup_{t\in[0,T]} d\left(\boldsymbol{x}(t), \boldsymbol{y}(t)\right) \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{y} \in D_{\mathbb{R}^{\mathcal{I}}}[0,T].$$

The topology of uniform convergence over compact sets on $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ is compatible with the metric

$$\varrho_\infty(\boldsymbol{x},\boldsymbol{y}) := \sum_{k=0}^{\infty} \frac{\min\{\varrho_k(\boldsymbol{x},\boldsymbol{y}), 1\}}{2^k} \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{y} \in D_{\mathbb{R}^{\mathcal{I}}}[0,\infty).$$

Since $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ is metrizable, a subset of $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ is relatively compact if and only if every sequence within the subset has a convergent subsequence. The same holds for the metric spaces $D_{\mathbb{R}^{\mathcal{I}}}[0,T]$ defined above.

The following arguments apply both for JLMU and SLTA. Fix some arbitrary time horizon $T \geq 0$. In order to prove Proposition 2, we first demonstrate that the sequences $\{\mathcal{A}_n : n \geq 1\}$, $\{\mathcal{D}_n : n \geq 1\}$ and $\{\boldsymbol{q}_n : n \geq 1\}$ are relatively compact in the space $D_{\mathbb{R}^{\mathcal{I}}}[0,T]$ with probability one, and that every convergent subsequence has a componentwise Lipschitz limit. Here we are actually referring to the restrictions to the interval $[0,T]$ of the processes $\mathcal{A}_n$, $\mathcal{D}_n$ and $\boldsymbol{q}_n$, but the symbols $|_{[0,T]}$ have been omitted in order to maintain a lighter notation; this is also done in the sequel when there is no risk of confusion. The latter properties are established using the methodological framework Bramson (1998). The following result defines a set of probability one where the relative compactness holds.

LEMMA 1. *There exists a set of probability one* $\Gamma_T \subset \Gamma_0$ *where:*

$$\lim_{n\to\infty} ||\boldsymbol{q}_n(0) - q_0||_1 = 0, \tag{52a}$$

$$\lim_{n\to\infty} \sup_{t\in[0,T]} n^\gamma \left| \frac{1}{n}\mathcal{N}_n^\lambda(t) - \lambda t \right| = 0, \tag{52b}$$

$$\lim_{n\to\infty} \sup_{t\in[0,\mu_j T]} n^\gamma \left| \frac{1}{n}\mathcal{N}_{(i,j)}(nt) - t \right| = 0 \quad \textit{for all} \quad (i,j)\in\mathcal{I}_+, \tag{52c}$$

for all $\gamma \in [0,1/2)$. *Also, if the load balancing policy is SLTA, then apart from the latter properties, there exists a random variable* $R \geq 1$ *such that*

$$\boldsymbol{r}_n(0) \leq R \quad \textit{and} \quad \boldsymbol{q}_n(0,i,j) < \alpha_n(i) \quad \textit{for all} \quad (i,j) \trianglelefteq \left(i_{\boldsymbol{r}_n(0)}, j_{\boldsymbol{r}_n(0)}\right) \quad \textit{and} \quad n.$$

Proof. This result is a straightforward consequence of (9), (15) and the refined strong law of large numbers for the Poisson process proven in (Goldsztajn et al. 2021, Lemma 4). □

We fix some $\omega \in \Gamma_T$, which we omit from the notation for brevity. Next we prove that the sequences $\{\mathcal{A}_n : n \geq 1\}$ and $\{\mathcal{D}_n : n \geq 1\}$ are relatively compact subsets of $D_{\mathbb{R}^\mathcal{I}}[0,T]$ and that the limit of every convergent subsequence is a function with Lipschitz coordinates. If JLMU is used, it then follows from (19) and (52a) that the sequence $\{\boldsymbol{q}_n : n \geq 1\}$ has the same properties. If SLTA is used, then we need to invoke (21a) instead of (19).

The following characterization of relative compactness with respect to ϱ_T will be useful. Let $D[0,T]$ denote the space of real càdlàg functions on $[0,T]$, with the uniform norm:

$$||\boldsymbol{x}||_T = \sup_{t\in[0,T]} |\boldsymbol{x}(t)| \quad \text{for all} \quad \boldsymbol{x} \in D[0,T].$$

Observe that a sequence of functions $\boldsymbol{x}_n \in D_{\mathbb{R}^\mathcal{I}}[0,T]$ converges to $\boldsymbol{x} \in D_{\mathbb{R}^\mathcal{I}}[0,T]$ with respect to ϱ_T if and only if $\boldsymbol{x}_n(i,j)$ converges to $\boldsymbol{x}(i,j)$ with respect to the uniform norm for all $(i,j) \in \mathcal{I}$. Moreover, we have the following proposition.

PROPOSITION 1. *The sequence* $\{\boldsymbol{x}_n : n \geq 1\}$ *is relatively compact in* $D_{\mathbb{R}^\mathcal{I}}[0,T]$ *if and only if* $\{\boldsymbol{x}_n(i,j) : n \geq 1\}$ *is a relatively compact subset of* $D[0,T]$ *for all* $(i,j) \in \mathcal{I}$.

Proof. We only need to prove the converse, so assume that $\{\boldsymbol{x}_n(i,j) : n \geq 1\}$ is relatively compact for all $(i,j) \in \mathcal{I}$. Given an increasing sequence $\mathcal{K} \subset \mathbb{N}$, we need to establish that there exists a subsequence of $\{\boldsymbol{x}_k : k \in \mathcal{K}\}$ that converges with respect to ϱ_T. For this purpose, we may construct a family of sequences $\{\mathcal{K}_j : j \geq 0\}$ with the following properties.

(a) $\mathcal{K}_{j+1} \subset \mathcal{K}_j \subset \mathcal{K}$ for all $j \geq 0$.

(b) $\{\boldsymbol{x}_k(i,j) : k \in \mathcal{K}_j\}$ has a limit $\boldsymbol{x}(i,j) \in D[0,T]$ for each $(i,j) \in \mathcal{I}$.

Define $\{k_l : l \geq 1\} \subset \mathcal{K}$ such that k_l is the l^{th} element of $\mathcal{K}_l$. Then

$$\lim_{l\to\infty} ||\boldsymbol{x}_{k_l}(i,j) - \boldsymbol{x}(i,j)||_T = 0 \quad \text{for all} \quad (i,j) \in \mathcal{I}.$$

Let $\boldsymbol{x} \in D_{\mathbb{R}^\mathcal{I}}[0,T]$ be the function with coordinates the limiting functions $\boldsymbol{x}(i,j)$ introduced in (b). Then $\boldsymbol{x}_{k_l}$ converges to $\boldsymbol{x}$ with respect to ϱ_T. □

As a result, it suffices to establish that $\{\mathcal{A}_n(i,j):n\geq 1\}$ and $\{\mathcal{D}_n(i,j):n\geq 1\}$ are relatively compact in $D[0,T]$ for all $(i,j)\in\mathcal{I}$. Consider the sets

$$L_M:=\{\boldsymbol{x}\in D[0,T]:\boldsymbol{x}(0)=0 \text{ and } |\boldsymbol{x}(t)-\boldsymbol{x}(s)|\leq M|t-s| \text{ for all } s,t\in[0,T]\}$$

which are compact for each $M>0$ by the Arzelá-Ascoli theorem. For each $(i,j)\in\mathcal{I}$, we prove that there exists M_j such that $\mathcal{A}_n(i,j)$ and $\mathcal{D}_n(i,j)$ approach L_{M_j} as n grows to large. Then we use the compactness of L_{M_j} to show that $\{\mathcal{A}_n(i,j):n\geq 1\}$ and $\{\mathcal{D}_n(i,j):n\geq 1\}$ are relatively compact subsets of $D[0,T]$. To this end, we introduce the spaces

$$L_M^{\varepsilon}:=\{\boldsymbol{x}\in D[0,T]:\boldsymbol{x}(0)=0 \text{ and } |\boldsymbol{x}(t)-\boldsymbol{x}(s)|\leq M|t-s|+\varepsilon \text{ for all } s,t\in[0,T]\}.$$

LEMMA 2. *If $\boldsymbol{x}\in L_M^{\varepsilon}$, then there exists $\boldsymbol{y}\in L_M$ such that $||\boldsymbol{x}-\boldsymbol{y}||_T\leq 4\varepsilon$.*

The above lemma is a restatement of (Bramson 1998, Lemma 4.2), and together with the next lemma, it implies that for each $(i,j)\in\mathcal{I}$ there exists a constant M_j such that $\mathcal{A}_n(i,j)$ and $\mathcal{D}_n(i,j)$ approach L_{M_j} as n grows large.

LEMMA 3. *For each $(i,j)\in\mathcal{I}$ there exist $M_j>0$ and $\{\varepsilon_n(i,j)>0:n\geq 1\}$ such that $\mathcal{A}_n(i,j)$ and $\mathcal{D}_n(i,j)$ lie in $L_{M_j}^{\varepsilon_n(i,j)}$ for all n and $\varepsilon_n(i,j)\to 0$ as n grows large.*

Proof. If $j=0$, then $\mathcal{A}_n(i,j)$ and $\mathcal{D}_n(i,j)$ are identically zero for all i and n, so the claim holds trivially. Therefore, let us fix $(i,j)\in\mathcal{I}_+$, in this case we have

$$|\mathcal{A}_n(t,i,j)-\mathcal{A}_n(s,i,j)|\leq\frac{1}{n}\left|\mathcal{N}_n^{\lambda}(t)-\mathcal{N}_n^{\lambda}(s)\right|\leq\lambda\,|t-s|+2\sup_{r\in[0,T]}\left|\frac{1}{n}\mathcal{N}_n^{\lambda}(r)-\lambda r\right|$$

for all $s,t\in[0,T]$. It follows from Lemma 1 that there exists a vanishing sequence of positive real numbers $\{\delta_n^1>0:n\geq 1\}$ such that

$$|\mathcal{A}_n(t,i,j)-\mathcal{A}_n(s,i,j)|\leq\lambda|t-s|+\delta_n^1\quad\text{for all}\quad s,t\in[0,T].$$

Consider now the non-decreasing function defined as

$$\boldsymbol{f}_n(t):=\int_0^t\mu j\left[\boldsymbol{q}_n(s,i,j)-\boldsymbol{q}_n(s,i,j+1)\right]ds\quad\text{for all}\quad t\in[0,T].$$

Note that $\boldsymbol{f}_n(0)=0$ and $|\boldsymbol{f}_n(t)-\boldsymbol{f}_n(s)|\leq\mu j|t-s|$ for all $s,t\in[0,T]$, so in particular, $\boldsymbol{f}_n(T)\leq\mu jT$. For all $s,t\in[0,T]$, we have

$$\begin{aligned}|\mathcal{D}_n(t,i,j)-\mathcal{D}_n(s,i,j)|&=\frac{1}{n}\left|\mathcal{N}_{(i,j)}\left(n\boldsymbol{f}_n(t)\right)-\mathcal{N}_{(i,j)}\left(n\boldsymbol{f}_n(s)\right)\right|\\&\leq|\boldsymbol{f}_n(t)-\boldsymbol{f}_n(s)|+2\sup_{r\in[0,T]}\left|\frac{1}{n}\mathcal{N}_{(i,j)}\left(n\boldsymbol{f}_n(r)\right)-\boldsymbol{f}_n(r)\right|\\&\leq\mu j|t-s|+2\sup_{r\in[0,\mu jT]}\left|\frac{1}{n}\mathcal{N}_{(i,j)}(nr)-r\right|.\end{aligned}$$

By (52c), there exists a vanishing sequence $\{\delta_n^2(i,j)>0:n\geq 1\}$ such that

$$|\mathcal{D}_n(t,i,j)-\mathcal{D}_n(s,i,j)|\leq\mu j|t-s|+\delta_n^2(i,j)\quad\text{for all}\quad s,t\in[0,T].$$

The result follows setting $M_j:=\max\{\lambda,\mu j\}$ and $\varepsilon_n(i,j):=\max\{\delta_n^1,\delta_n^2(i,j)\}$. □

We now demonstrate that $\{\mathcal{A}_n : n \geq 1\}$, $\{\mathcal{D}_n : n \geq 1\}$ and $\{\boldsymbol{q}_n : n \geq 1\}$ are relatively compact subsets of $D_{\mathbb{R}^{\mathcal{I}}}[0,T]$ with probability one, and that the limit of every convergent subsequence is componentwise Lipschitz.

LEMMA 4. *The sequences $\{\mathcal{A}_n(\omega) : n \geq 1\}$, $\{\mathcal{D}_n(\omega) : n \geq 1\}$ and $\{\boldsymbol{q}_n(\omega) : n \geq 1\}$ are all relatively compact in $D_{\mathbb{R}^{\mathcal{I}}}[0,T]$ for all $\omega \in \Gamma_T$. Also, they have the property that the limit of every convergent subsequence is a function with Lipschitz coordinates.*

Proof. We fix an arbitrary $\omega \in \Gamma_T$ and we omit it from the notation for brevity. As noted above, it suffices to prove that $\{\mathcal{A}_n(i,j) : n \geq 1\}$ and $\{\mathcal{D}_n(i,j) : n \geq 1\}$ are relatively compact subsets of $D[0,T]$ for all $(i,j) \in \mathcal{I}$. We fix some $(i,j) \in \mathcal{I}$ and demonstrate that $\{\mathcal{A}_n(i,j) : n \geq 1\}$ is relatively compact in $D[0,T]$, and that the limit of every convergent subsequence is Lipschitz. The same arguments can be used if $\mathcal{A}_n$ is replaced by $\mathcal{D}_n$.

Let M_j and $\{\varepsilon_n(i,j) : n \geq 1\}$ be as in the statement of Lemma 3. It follows from Lemma 2 that for each n there exists $\boldsymbol{x}_n(i,j) \in L_{M_j}$ such that

$$||\mathcal{A}_n(i,j) - \boldsymbol{x}_n(i,j)||_T \leq 4\varepsilon_n(i,j).$$

Recall that L_{M_j} is compact, thus every increasing sequence of natural numbers has a subsequence $\mathcal{K}$ such that $\{\boldsymbol{x}_k(i,j) : k \in \mathcal{K}\}$ converges to a function $\boldsymbol{x} \in L_{M_j}$. Furthermore,

$$\limsup_{k\to\infty} ||\mathcal{A}_k(i,j) - \boldsymbol{x}(i,j)||_T \leq \lim_{k\to\infty} 4\varepsilon_k(i,j) + \lim_{k\to\infty} ||\boldsymbol{x}_k(i,j) - \boldsymbol{x}(i,j)||_T = 0,$$

where the limits are taken along $\mathcal{K}$. This shows that every subsequence of $\{\mathcal{A}_n(i,j) : n \geq 1\}$ has a further subsequence which converges to a Lipschitz function. □

We are now ready to prove Proposition 2

Proof of Proposition 2. Recall that Γ_T has probability one for all $T \geq 0$. Hence,

$$\Gamma_\infty := \bigcap_{T\in\mathbb{N}} \Gamma_T$$

has probability one as well. Also, (22) and (23) hold within Γ_∞ by Lemma 1.

Next we fix an arbitrary $\omega \in \Gamma_\infty$, which we omit from the notation for brevity, and we prove that $\{\boldsymbol{q}_n : n \geq 1\}$ is a relatively compact subset of $D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ such that the limit of every convergent subsequence has locally Lipschitz coordinate functions. Exactly the same arguments can be used if $\boldsymbol{q}_n$ is replaced by $\mathcal{A}_n$ or $\mathcal{D}_n$.

Fix a sequence $\mathcal{K} \subset \mathbb{N}$. We must show that $\{\boldsymbol{q}_k : k \in \mathcal{K}\}$ has a subsequence which converges uniformly over compact sets to a function with locally Lipschitz coordinates. To this end, we may construct a family of sequences $\{\mathcal{K}_T : T \in \mathbb{N}\}$ such that:

(a) $\mathcal{K}_{T+1} \subset \mathcal{K}_T \subset \mathcal{K}$ for all $T \in \mathbb{N}$;

(b) for each $T \in \mathbb{N}$, there exists $\boldsymbol{q}_T \in D_{\mathbb{R}^{\mathcal{I}}}[0,T]$ such that the coordinates of $\boldsymbol{q}_T$ are Lipschitz and $\varrho_T\left(\boldsymbol{q}_k|_{[0,T]}, \boldsymbol{q}_T\right) \to 0$ as $k \to \infty$ with $k \in \mathcal{K}_T$.

Let k_l denote the l^{th} element of $\mathcal{K}_l$. It follows from (a) and (b) that

$$\lim_{l\to\infty} \varrho_T\left(\boldsymbol{q}_{k_l}|_{[0,T]}, \boldsymbol{q}_T\right) = 0 \quad \text{for all} \quad T \in \mathbb{N}. \tag{53}$$

Note that $\boldsymbol{q}_{k_l}(t) \to \boldsymbol{q}_T(t)$ with l for all $t \leq T$, so $\boldsymbol{q}_S(t) = \boldsymbol{q}_T(t)$ for all $t \leq S,T$. Thus, we may define a function $\boldsymbol{q} \in D_{\mathbb{R}^{\mathcal{I}}}[0,\infty)$ such that $\boldsymbol{q}(t) = \boldsymbol{q}_T(t)$ for all $t \leq T$. Moreover, (b) implies that $\boldsymbol{q}$ has locally Lipschitz coordinates and (53) means that $\{\boldsymbol{q}_{k_l} : l \geq 1\}$ converges uniformly over compact sets to $\boldsymbol{q}$. □

Appendix C: Limiting behavior of SLTA

Proof of Proposition 4. We fix an arbitrary $\omega \in \Gamma$, which is omitted from the notation for brevity. It follows from (21a) that

$$\boldsymbol{s}_n = \boldsymbol{s}_n(0) + \sum_{(i,j)\in\mathcal{I}_+} [\mathcal{A}_n(i,j) - \mathcal{D}_n(i,j)] = \boldsymbol{s}_n(0) + \frac{1}{n}\mathcal{N}_n^{\lambda}(t) - \sum_{(i,j)\in\mathcal{I}_+} \mathcal{D}_n(i,j) \quad \text{for all} \quad n.$$

By Proposition 3, every increasing sequence of natural numbers has a subsequence $\mathcal{K}$ such that the sequences $\{\mathcal{D}_k : k \in \mathcal{K}\}$ and $\{\boldsymbol{q}_k : k \in \mathcal{K}\}$ converge in $D_{\ell_1}[0,\infty)$ to certain functions $\boldsymbol{d}$ and $\boldsymbol{q}$, respectively. It follows from (22c) that

$$\boldsymbol{d}(t,i,j) = \int_0^t \mu j \left[\boldsymbol{q}(s,i,j) - \boldsymbol{q}(s,i,j+1)\right] ds \quad \text{for all} \quad t \geq 0 \quad \text{and} \quad (i,j) \in \mathcal{I}_+.$$

Moreover, $\{\boldsymbol{s}_k : k \in \mathcal{K}\}$ converges uniformly over compact sets to

$$\tilde{\boldsymbol{s}} := \sum_{(i,j)\in\mathcal{I}_+} \boldsymbol{q}(i,j).$$

It follows from (22a) and (22b) that

$$\lim_{k\to\infty} \boldsymbol{s}_k(0) = s_0 \quad \text{and} \quad \lim_{k\to\infty} \sup_{t\in[0,T]} \left| \frac{1}{k}\mathcal{N}_k^{\lambda}(t) - \lambda t \right| = 0 \quad \text{for all} \quad T \geq 0.$$

Furthermore, the fact that $\{\mathcal{D}_k : k \in \mathcal{K}\}$ converges to $\boldsymbol{d}$ in $D_{\ell_1}[0,\infty)$ implies that

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \left| \sum_{(i,j)\in\mathcal{I}_+} \mathcal{D}_k(t,i,j) - \int_0^t \mu \tilde{\boldsymbol{s}}(s) ds \right| =$$

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \left| \sum_{i=1}^{m} \sum_{j=1}^{\infty} \left[\mathcal{D}_k(t,i,j) - \int_0^t \mu j \left[\boldsymbol{q}(s,i,j) - \boldsymbol{q}(s,i,j+1)\right] ds \right] \right| =$$

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \left| \sum_{(i,j)\in\mathcal{I}_+} [\mathcal{D}_k(t,i,j) - \boldsymbol{d}(t,i,j)] \right| = 0 \quad \text{for all} \quad T \geq 0.$$

We conclude that $\tilde{\boldsymbol{s}}$ satisfies

$$\tilde{\boldsymbol{s}}(t) = s_0 + \lambda t - \int_0^t \mu \tilde{\boldsymbol{s}}(s) ds \quad \text{for all} \quad t \geq 0.$$

Equivalently, $\tilde{\boldsymbol{s}}$ solves (35), and therefore $\tilde{\boldsymbol{s}} = \boldsymbol{s}$. This demonstrates that every subsequence of $\{\boldsymbol{s}_n : n \geq 1\}$ has a further subsequence that converges to $\boldsymbol{s}$ uniformly over compact sets. The claim of the proposition follows from this fact. □

Proof of Proposition 5. Let us omit ω from the notation for brevity. Recall that tasks are always assigned to server pools with coordinates $(i,j) \trianglerighteq (i_{\boldsymbol{r}_n}, j_{\boldsymbol{r}_n})$, as explained in Remark 1. By (b), $(i_{\boldsymbol{r}_k}, j_{\boldsymbol{r}_k}) \trianglerighteq (i_r, j_r)$ along the interval $[t_0, t_1]$ for each $k \in \mathcal{K}$. Moreover, (b) implies that the system has green tokens when the equality holds. Therefore, tasks are only assigned to server pools with coordinates $(i,j) \triangleright (i_r, j_r)$ along the interval $[t_0, t_1]$ in all the systems comprised in $\mathcal{K}$. Thus, $\mathcal{A}_k(t,i,j) = \mathcal{A}_k(t_0,i,j)$ for all $t \in [t_0,t_1]$ and $(i,j) \trianglelefteq (i_r, j_r)$. We conclude from (21a) that

$$\boldsymbol{q}_k(t,i,j) = \boldsymbol{q}_k(t_0,i,j) - [\mathcal{D}_k(t,i,j) - \mathcal{D}_k(t_0,i,j)]$$

for all $t \in [t_0, t_1]$, $(i,j) \trianglelefteq (i_r, j_r)$ and $k \in \mathcal{K}$. As in the proof of Proposition 4, we obtain

$$\boldsymbol{q}(t,i,j) = \boldsymbol{q}(t_0,i,j) - \int_{t_0}^{t} \mu j \left[\boldsymbol{q}(s,i,j) - \boldsymbol{q}(s,i,j+1)\right] ds$$

for all $t \in [t_0, t_1]$, $(i,j) \trianglelefteq (i_r, j_r)$ and $k \in \mathcal{K}$. It follows that $\boldsymbol{q}(i,j)$ is differentiable for all $(i,j) \trianglelefteq (i_r, j_r)$ because $\boldsymbol{q}$ has continuous coordinates by Proposition 2. Furthermore, the system of differential equations in the statement of the proposition holds.

Note that $\{\boldsymbol{v}_k(r) : k \geq 1\}$ converges uniformly over compact sets to

$$\boldsymbol{v}(r) := \sum_{(i,j) \trianglelefteq (i_r, j_r)} \boldsymbol{q}(i,j)$$

by (a) and the definition of $\boldsymbol{v}_k$. Let $J_i := \min\{j \geq 1 : (i,j) \trianglelefteq (i_r, j_r)\}$ and note that

$$\begin{aligned}\dot{\boldsymbol{v}}(t,r) &= \sum_{i=1}^{m} \sum_{j=J_i}^{\infty} -\mu j \left[\boldsymbol{q}(t,i,j) - \boldsymbol{q}(t,i,j+1)\right] \\ &= \sum_{i=1}^{m} \left[-\mu(J_i - 1)\boldsymbol{q}(t,i,J_i) - \mu \sum_{j=J_i}^{\infty} \boldsymbol{q}(t,i,j)\right] = -\mu \sum_{i=1}^{m} (J_i - 1)\boldsymbol{q}(t,i,J_i) - \mu \boldsymbol{v}(t,r).\end{aligned}$$

Since the first term on the right-hand side is non-positive, we have

$$\boldsymbol{v}(t,r) \leq \boldsymbol{v}(t_0,r)\mathrm{e}^{-\mu(t-t_0)} < \boldsymbol{s}(t_0)\mathrm{e}^{-\mu(t-t_0)} \quad \text{for all} \quad t \in [t_0, t_1].$$

This completes the proof. □

Proof of Proposition 6. Define

$$\alpha_{\min} := \min_{1 \leq i \leq m} \alpha(i), \quad J := \left\lfloor \frac{\rho + 1}{\alpha_{\min}} \right\rfloor \quad \text{and} \quad r_{\mathrm{bd}} := \max\{r \geq 2 : j_{r-1} \leq J\}.$$

The function τ_{bd} mentioned in the statement of the proposition is defined by

$$\tau_{\mathrm{bd}}(s) = \begin{cases} \frac{1}{\mu}\left[\log\left(\frac{s-\rho}{\sum_{(i,j) \trianglerighteq \sigma_*} - \rho}\right)\right]^+ + \frac{J - j_{r_*}}{\mu}(r_{\mathrm{bd}} - r_*) & \text{if} \quad s > \rho, \\ \frac{J - j_{r_*}}{\mu}(r_{\mathrm{bd}} - r_*) & \text{if} \quad s \leq \rho. \end{cases}$$

For brevity, we omit ω from the notation, and we let

$$\Sigma(r) := \sum_{(i,j) \triangleright (i_r, j_r)} \alpha(i) \quad \text{and} \quad \Sigma_n(r) := \sum_{(i,j) \triangleright (i_r, j_r)} \alpha_n(i) \quad \text{for all} \quad n, r \geq 1.$$

It follows from (10) that we can write $\Sigma(r_* + 1) = \rho + \delta + \varepsilon$ with $\delta, \varepsilon > 0$. We regard ε as a function of δ, determined by the latter identity, and for each

$$0 < \delta < \min\{\Sigma(r_* + 1) - \rho, \alpha_{\min}\} \quad \text{and} \quad \theta \in (0,1)$$

we define $\tau(\delta, \theta) := \min\{t \geq 0 : (s_0 - \rho)\,\mathrm{e}^{-\mu t} \leq \theta\varepsilon\}$. Explicitly,

$$\tau(\delta,\theta) = \begin{cases} \frac{1}{\mu}\left[\log\left(\frac{s_0 - \rho}{\theta[\Sigma(r_* + 1) - \rho - \delta]}\right)\right]^+ & \text{if} \quad s_0 > \rho, \\ 0 & \text{if} \quad s_0 \leq \rho. \end{cases}$$

Thus, the function $\boldsymbol{s}$ defined by (35) satisfies

$$\boldsymbol{s}(t) \leq \rho + \theta\varepsilon \quad \text{for all} \quad t \geq \tau(\delta,\theta). \tag{54}$$

Note that $\tau_{\mathrm{bd}}(s_0)$ is the infimum over δ and θ of

$$\tau(\delta,\theta) + \delta + \frac{J - j_{r_*}}{\mu\theta}(r_{\mathrm{bd}} - r_*).$$

We fix δ and θ such that the above expression is smaller than τ and we choose n_0 such that the following properties hold for all $n \geq n_0$.

(i) $\boldsymbol{r}_n(t) \le R_T$ for all $t \in [0,T]$.

(ii) $|\boldsymbol{s}_n(t) - \boldsymbol{s}(t)| \le (1-\theta)\varepsilon$ for all $t \in [0,T]$.

(iii) $1/n \le \beta_n \le \min\{\alpha_n(i) : 1 \le i \le m\}$ and $\Sigma_n(r_*+1) - L(r_*+1)\beta_n \ge \rho + \varepsilon$, where we used the notation $L(r) := \max\{\ell_i(r) : 1 \le i \le m\}$.

Properties (i) and (ii) are possible by Lemma 2 and Proposition 4, respectively, and (iii) can be enforced by assumptions (9) and (14).

Fix $n \ge n_0$ and note that for each $r > r_*$ and $t \in [\tau(\delta,\theta), T]$ we have

$$\Sigma_n(r) - L(r)\beta_n \ge \Sigma_n(r_*+1) - L(r_*+1)\beta_n \ge \rho + \varepsilon \ge \boldsymbol{s}_n(t).$$

The first inequality follows from the fact that $\Sigma_n(\,\cdot\,) - L(\,\cdot\,)\beta_n$ is non-decreasing provided that $\beta_n \le \min\{\alpha_n(i) : 1 \le i \le m\}$, and the last inequality follows from (ii) and (54). We conclude from the above inequalities that

$$\sum_{i=1}^{m} \boldsymbol{q}_n\left(t, i, \ell_i(r)\right) \le 1 - \beta_n \le 1 - \frac{1}{n} \quad \text{for all} \quad t \in [\tau(\delta,\theta), T]\,, \tag{55}$$

since otherwise we would have $\boldsymbol{s}_n(t) > \Sigma_n(r) - L(r)\beta_n$. It follows from (55) with $r = r_*+1$ that the total number of tokens is at least one if $t \in [\tau(\delta,\theta), T]$ and $\boldsymbol{r}_n(t) \ge r_*$, so we conclude that $\boldsymbol{r}_n$ does not increase beyond r_* along the interval $[\tau(\delta,\theta), T]$. Hence, it only remains to prove that the amount of time after $\tau(\delta,\theta)$ until $\boldsymbol{r}_n \le r_*$ is upper bounded by

$$\delta + \frac{J - j_{r_*}}{\mu\theta}\left(r_{\mathrm{bd}} - r_*\right).$$

It follows from (55) that the number of green tokens is

$$n - \sum_{i=1}^{m} n\boldsymbol{q}_n(t, i, \ell_i\left(\boldsymbol{r}_n(t)\right)) \ge n\beta_n \quad \text{if} \quad \boldsymbol{r}_n(t) > r_*, \quad t \in [\tau(\delta,\theta), T] \quad \text{and} \quad n \ge n_0.$$

Therefore, $\boldsymbol{r}_n$ decreases with the next arrival if the latter conditions hold and

$$\boldsymbol{q}_n(t, i_{\boldsymbol{r}_n(t)-1}, j_{\boldsymbol{r}_n(t)-1}) < \alpha_n(i_{\boldsymbol{r}_n(t)-1}).$$

Choose $n_{\mathrm{bd}}^{\tau,T} \ge n_0$ such that the following conditions hold.

(iv) $\min\{\alpha_n(i) : 1 \le i \le m\} > \alpha(\delta) := \alpha_{\min} - \delta$.

(v) Let $d := \delta\left[2\left(r_{\mathrm{bd}} - r_*\right) + 1\right]^{-1}$. The following two properties hold on every subinterval of $[0,T]$ of length at least d. First, $\mathcal{N}_n^{\lambda}(\,\cdot\,)$ has at least $R_T - r_{\mathrm{bd}}$ jumps. Second, for all $1 \le i \le m$ and $1 \le j \le J$, the process $\mathcal{N}_{(i,j)}(n\,\cdot\,)$ has at least $n\theta d$ jumps followed by at least one jump of $\mathcal{N}_n^{\lambda}(\,\cdot\,)$.

(vi) $\mu n\left[\alpha_n(i) - \alpha(\delta)\right]\theta d \ge 1$ for all $1 \le i \le m$.

Properties (iv) and (v) can be enforced by (9) and Proposition 2, respectively.

We may assume without loss of generality that

$$J \ge \left\lfloor \frac{\rho + \varepsilon}{\alpha(\delta)} \right\rfloor,$$

choosing a smaller δ at the start of the proof if needed. By the above inequality,

$$\boldsymbol{q}_n(t,i,j) \le \frac{\boldsymbol{s}_n(t)}{j} \le \frac{\rho+\varepsilon}{J+1} < \alpha(\delta) \le \alpha_n(i) \quad \text{if} \quad j > J, \quad t \in [\tau(\delta,\theta), T] \quad \text{and} \quad n \ge n_{\mathrm{bd}}^{\tau,T}.$$

Therefore, if $n \geq n_{\mathrm{bd}}^{\tau,T}$, $t \in [\tau(\delta,\theta), T]$ and $\boldsymbol{r}_n(t) > r_{\mathrm{bd}}$, then $\boldsymbol{r}_n$ decreases with each arrival until it reaches r_{bd}. We conclude from (i) and (v) that

$$\boldsymbol{r}_n(t) \leq r_{\mathrm{bd}} \quad \text{for all} \quad t \in [\tau(\delta,\theta)+d, T] \quad \text{and} \quad n \geq n_{\mathrm{bd}}^{\tau,T}.$$

Suppose now that $\boldsymbol{r}_n(t) = r \in (r_*, r_{\mathrm{bd}}]$ for some $t \in [\tau(\delta,\theta)+d, T]$ and $n \geq n_{\mathrm{bd}}^{\tau,T}$. Then $\boldsymbol{r}_n$ decreases as soon as a server pool of class i_{r-1} has a green token and a task arrives. Moreover, since the number of green tokens remains positive, no tasks are sent to server pools of class i_{r-1} before $\boldsymbol{r}_n$ decreases. At least $n\left[\alpha_n(i) - \alpha(\delta)\right]$ server pools of class i_{r-1} have strictly less than $J+1$ tasks, thus it follows from (v) that the time until $\boldsymbol{q}_n(i_{r-1}, j_{r-1})$ drops below $\alpha_n(i)$ is at most

$$\left\lceil \frac{n(J-j_{r-1})\left[\alpha_n(i)-\alpha(\delta)\right]+1}{\mu n\left[\alpha_n(i)-\alpha(\delta)\right] j_{r-1}\theta d} \right\rceil d \leq \frac{J-j_{r-1}}{\mu\theta} + \frac{1}{\mu n\left[\alpha_n(i)-\alpha(\delta)\right]\theta} + d \leq \frac{J-j_{r_*}}{\mu\theta} + 2d.$$

The numerator inside the ceiling function is the maximum number of tasks that need to leave the $n\left[\alpha_n(i)-\alpha(\delta)\right]$ server pools of class i_{r-1} with strictly fewer than $J+1$ tasks to ensure that one of these server pools has strictly less than j_{r-1} tasks, and the quantity $n\left[\alpha_n(i)-\alpha(\delta)\right] j_{r-1}$ in the denominator is a lower bound for the cumulative number of tasks in these server pools while all the server pools have at least j_{r-1} tasks. Also, note that the first inequality uses $j_{r-1} \geq 1$ and the last inequality uses (vi). Since $\boldsymbol{r}_n$ decreases at most $r_{\mathrm{bd}} - r_*$ times before it reaches r_*, we have

$$\boldsymbol{r}_n(t) \leq r_* \quad \text{for all} \quad t \in \left[\tau(\delta,\theta) + \delta + \frac{J-j_{r_*}}{\mu\theta}\left(r_{\mathrm{bd}} - r_*\right), T\right] \quad \text{and} \quad n \geq n_{\mathrm{bd}}^{\tau,T}.$$

This completes the proof. □

Proof of Lemma 6. We omit ω from the notation. It follows from (21a) that

$$\begin{aligned}\sum_{(i,j)\triangleright(i_r,j_r)} \boldsymbol{q}_k(t,i,j) - \sum_{(i,j)\triangleright(i_r,j_r)} \boldsymbol{q}_k\left(\tau_{k,1},i,j\right) = &\sum_{(i,j)\triangleright(i_r,j_r)} \left[\mathcal{A}_k(t,i,j) - \mathcal{A}_k\left(\tau_{k,1},i,j\right)\right] \\ &- \sum_{(i,j)\triangleright(i_r,j_r)} \left[\mathcal{D}_k(t,i,j) - \mathcal{D}_k\left(\tau_{k,1},i,j\right)\right].\end{aligned}$$

Note that all tasks are assigned to server pools with coordinates $(i,j) \triangleright (i_r,j_r)$ during the interval $(\tau_{k,1}, t]$, so the first term on the right-hand side can be expressed as follows:

$$\sum_{(i,j)\triangleright(i_r,j_r)} \left[\mathcal{A}_k(t,i,j) - \mathcal{A}_k\left(\tau_{k,1},i,j\right)\right] = \frac{1}{k}\left[\mathcal{N}_k^{\lambda}(t) - \mathcal{N}_k^{\lambda}\left(\tau_{k,1}\right)\right].$$

Using the process $\delta_k(r)$ defined in (36), we may write

$$\begin{aligned}&\sum_{(i,j)\triangleright(i_r,j_r)} \boldsymbol{q}_k(t,i,j) - \sum_{(i,j)\triangleright(i_r,j_r)} \boldsymbol{q}_k\left(\tau_{k,1},i,j\right) \\ &\qquad = \left(t-\tau_{k,1}\right)\lambda - \sum_{(i,j)\triangleright(i_r,j_r)} \int_{\tau_{k,1}}^{t} \mu j\left[\boldsymbol{q}_k(s,i,j) - \boldsymbol{q}_k(s,i,j+1)\right] ds \\ &\qquad + \delta_k(t,r) - \delta_k(\tau_{k,1},r) \\ &\qquad \geq \left(t-\tau_{k,1}\right)\left[\lambda - \mu \sum_{(i,j)\triangleright(i_r,j_r)} \alpha_k(i)\right] - 2\sup_{s\in[0,T]}\left|\delta_k(s,r)\right|.\end{aligned}$$

For the last inequality, observe that

$$\begin{aligned}\sum_{(i,j)\triangleright(i_r,j_r)} j\left[\boldsymbol{q}_k(i,j)-\boldsymbol{q}_k(i,j+1)\right] &= \sum_{i=1}^{m}\sum_{j=1}^{\ell_i(r)} j\left[\boldsymbol{q}_k(i,j)-\boldsymbol{q}_k(i,j+1)\right]\\ &= \sum_{i=1}^{m}\left[\sum_{j=1}^{\ell_i(r)}\boldsymbol{q}_k(i,j)-\ell_i(r)\boldsymbol{q}_k(i,\ell_i(r)+1)\right]\\ &\leq \sum_{(i,j)\triangleright(i_r,j_r)}\alpha_k(i) \quad \text{for all} \quad k\in\mathcal{K}.\end{aligned}$$

This completes the proof. □

Proof of Lemma 7. We omit ω from the notation and we assume that $r>1$, otherwise the statement holds trivially. Fix an arbitrary $\gamma\in[0,1/2)$ and let

$$\xi_{k,2}:=\inf\left\{t\geq\zeta_{k,1}:\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)\leq\sum_{(i,j)\triangleright(i_r,j_r)}\alpha_k(i)-\min\left\{\beta_k,k^{-\gamma}\right\}\right\},$$

$$\xi_{k,1}:=\sup\left\{t\leq\xi_{k,2}:\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)=\sum_{(i,j)\triangleright(i_r,j_r)}\alpha_k(i)\right\}.$$

By assumption, $\boldsymbol{r}_k\leq r$ along the interval $[\zeta_{k,1},\zeta_{k,2}]$ and $\boldsymbol{r}_k(\zeta_{k,1})=r$, since otherwise the number of tokens at time $\zeta_{k,1}$ would be zero, and this cannot happen by Remark 1. Moreover, the number of green tokens is upper bounded by

$$\sum_{(i,j)\triangleright(i_r,j_r)}k\alpha_k(i)-\sum_{(i,j)\triangleright(i_r,j_r)}k\boldsymbol{q}_k(i,j)<k\beta_k$$

along the interval $[\zeta_{k,1},\xi_{k,2})$, which implies that $\boldsymbol{r}_k$ cannot decrease along $[\zeta_{k,1},\xi_{k,2}]$. We conclude that $\boldsymbol{r}_k(t)=r$ for all $t\in[\zeta_{k,1},\zeta_{k,2}\wedge\xi_{k,2}]$ and $k\in\mathcal{K}$. Moreover,

$$\max\left\{\alpha_k(i)-\boldsymbol{q}_k(i,j):(i,j)\triangleright(i_r,j_r)\right\}\leq\sum_{(i,j)\triangleright(i_r,j_r)}\alpha_k(i)-\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(i,j)\leq k^{-\gamma}$$

along $[\zeta_{k,1},\xi_{k,2}]$. Thus, it suffices to prove that $\xi_{k,2}\geq\zeta_{k,2}$ for all large enough $k\in\mathcal{K}$.

In order to prove that $\xi_{k,2}\geq\zeta_{k,2}$, we show that

$$\begin{aligned}&\sum_{(i,j)\triangleright(i_r,j_r)}\alpha_k(i)-\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)=\\ &\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k\left(\xi_{k,1}^{-},i,j\right)-\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)<\min\left\{\beta_k,k^{-\gamma}\right\}\end{aligned}\tag{56}$$

for all $t\in[\xi_{k,1},\zeta_{k,2}\wedge\xi_{k,2}]$ and all large enough $k\in\mathcal{K}$. Note that, by definition of $\xi_{k,2}$, the latter statement implies that $\xi_{k,2}>\zeta_{k,2}$.

Let $\tau_{k,1}:=\zeta_{k,2}\wedge\xi_{k,1}$ and $\tau_{k,2}:=\zeta_{k,2}\wedge\xi_{k,2}$. In addition, let M_r be the cardinality of the finite set $\{(i,j)\in\mathcal{I}_+:(i,j)\triangleright(i_r,j_r)\}$. For all $t\in[\tau_{k,1},\tau_{k,2}]$, we have

$$\begin{aligned}&\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)-\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k\left(\tau_{k,1}^{-},i,j\right)\\ &\qquad\geq\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)-\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k\left(\tau_{k,1},i,j\right)-\frac{M_r}{k}\end{aligned}$$

$$\geq (t-\tau_{k,1})\left[\lambda-\mu\sum_{(i,j)\triangleright(i_r,j_r)}\alpha_k(i)\right]-2\sup_{s\in[0,T]}|\delta_k(s,r)|-\frac{M_r}{k}$$
$$\geq -2\sup_{s\in[0,T]}|\delta_k(s,r)|-\frac{M_r}{k}.$$

For the first inequality, note that the processes $\boldsymbol{q}_k(i,j)$ have jumps of size $1/k$. The second inequality follows from Lemma 6, since

$$\boldsymbol{r}_k(t)=r \quad \text{and} \quad \sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)<\sum_{(i,j)\triangleright(i_r,j_r)}\alpha_k(i)$$

for all $t\in[\tau_{k,1},\tau_{k,2})$ and $k\in\mathcal{K}$. The third inequality follows from (10) and $r\leq r_*$.

We conclude from (14) and (37) that

$$\lim_{k\to\infty}\frac{1}{\beta_k}\left[2\sup_{s\in[0,T]}|\delta_k(s,r)|+\frac{M_r}{k}\right]=\lim_{k\to\infty}\frac{1}{k^{\gamma_0}\beta_k}\left[2\sup_{s\in[0,T]}k^{\gamma_0}|\delta_k(s,r)|+\frac{M_r}{k^{1-\gamma_0}}\right]=0.$$

If β_k is replaced by $k^{-\gamma}$ on the left-hand side, then the limit is also equal to zero by (37). Consequently, for all sufficiently large $k\in\mathcal{K}$, we have

$$\begin{aligned}\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k(t,i,j)-\sum_{(i,j)\triangleright(i_r,j_r)}\boldsymbol{q}_k\left(\tau_{k,1}^-,i,j\right)&\geq -2\sup_{s\in[0,T]}|\delta_k(s,r)|-\frac{M_r}{k}\\&>-\min\left\{\beta_k,k^{-\gamma}\right\}\end{aligned}$$

for all $t\in[\tau_{k,1},\tau_{k,2}]$. This implies that (56) holds for all sufficiently large $k\in\mathcal{K}$; observe that (56) holds trivially when $\xi_{k,1}>\zeta_{k,2}$. □

References

Bramson M (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* 30(1-2):89–140.

Chen H, Yao DD (2013) *Fundamentals of queueing networks: Performance, asymptotics, and optimization* (Springer Science & Business Media).

Goldsztajn D, Borst SC, Van Leeuwaarden JS (2021) Learning and balancing unknown loads in large-scale systems. *arXiv preprint arXiv:2012.10142* .

Karatzas I, Shreve S (2014) *Brownian motion and stochastic calculus*, volume 113 (Springer).