Online companion for

Scheduling Flexible Servers with Convex Delay Costs In Many-Server Service Systems

Itay Gurvich* Ward Whitt†

A Proofs of Asymptotic Optimality for the Delay-Cost Formulation

In this appendix we provide the proofs of Theorems 3.3 and 3.4 in the paper. We start by proving that asymptotic efficiency (see Definition 3.5) implies the stochastic boundedness and C-tightness of $\hat{Q}^{\lambda}_{\Sigma}(t)$. A family $\{x^{\lambda}, \lambda > 0\}$ of processes in $D^d[0,T]$ is said to be *stochastically bounded* if

$$\lim_{k\to\infty} \limsup_{\lambda\to\infty} P\{\|x^r\|_T > k\} = 0.$$

It is said to be *tight* if every subsequence with $\lambda_k \to \infty$ contains a convergent subsequence and *C-tight* if the limit of each such subsequence is continuous. We refer the reader to §5 Pang et al. (2007) for a detailed discussion of these concepts.

Lemma A.1. (**stochastic boundedness and C-tightness**) *For any family* $\{\pi^{\lambda}, \lambda > 0\} \in \Pi^e$*, the corresponding family* $\{\hat{Q}^{\lambda}_{\Sigma}(t), t \geq 0\}$ *is stochastically bounded and C-tight.*

Proof: By equation (52) of [10], we can write

$$\hat{X}^{\lambda}_{\Sigma}(t) = \hat{X}^{\lambda}_{\Sigma}(0) - \beta t + \sum_{j\in\mathcal{J}} \mu_j \int_0^t \hat{I}^{\lambda}_j(s)ds + \hat{M}^{\lambda}_{\Sigma}(t) + o(1) \quad \text{as} \quad \lambda \to \infty, \tag{A1}$$

where $\hat{M}^{\lambda}_{\Sigma}(t)$ is the square integrable Martingale defined prior to (52) in Gurvich and Whitt (2007b) for each λ. Consequently,

$$|\hat{X}^{\lambda}_{\Sigma}(t)| \leq |\beta|t + \mu_1 \int_0^t \hat{I}^{\lambda}_{\Sigma}(s)ds + |\hat{M}^{\lambda}_{\Sigma}(t)|.$$

*Columbia Business School, 4I Uris Hall, 3022 Broadway, NY, NY 10027. (ig2126@columbia.edu)

†IEOR Department, Columbia University, 304 S. W. Mudd Building, 500 West 120th Street, New York, NY 10027-6699. (ww2040@columbia.edu)

By the asymptotic efficiency assumption, $\|\hat{I}^\lambda_\Sigma - [\hat{X}^\lambda_\Sigma]^-\|_T \Rightarrow 0$ as $\lambda \to \infty$, so that

$$|\hat{X}^\lambda_\Sigma(t)| \le |\beta|t + \mu_1 \int_0^t |\hat{X}^\lambda_\Sigma(s)|ds + |\hat{M}^\lambda_\Sigma(t)| + o(1) \quad \text{as} \quad \lambda \to \infty.$$

Since $\hat{M}^\lambda_\Sigma(t)$ is C-tight - see the proof of Lemma 4.2 in Gurvich and Whitt (2007b)) - it is also stochastically bounded. Hence we can apply Gronwall's inequality to deduce that the family $\hat{X}^\lambda_\Sigma(t)$ is stochastically bounded.

To establish C-tightness, we use (A1) to write

$$\hat{X}^\lambda_\Sigma(t) - \hat{X}^\lambda_\Sigma(s) = -\beta(t-s) + \sum_{j\in\mathcal{J}} \mu_j \int_s^t \hat{I}^\lambda_j(h)dh + \hat{M}^\lambda_\Sigma(t) - \hat{M}^\lambda_\Sigma(s) + o(1) \quad \text{as} \quad \lambda \to \infty$$

and, consequently,

$$|\hat{X}^\lambda_\Sigma(t) - \hat{X}^\lambda_\Sigma(s)| \le |\beta|(t-s) + \mu_1(t-s)\|\hat{X}^\lambda_\Sigma\|_T + \mu_1 \int_0^{t-s} |\hat{X}^\lambda_\Sigma(s+h) - \hat{X}^\lambda_\Sigma(s)|dh + |\hat{M}^\lambda_\Sigma(t) - \hat{M}^\lambda_\Sigma(s)| + o(1).$$

C-tightness now follows from the stochastic boundedness of $\hat{X}^\lambda_\Sigma(t)$ and the C-tightness of $\hat{M}^\lambda_\Sigma(t)$ through an application of Gronwall's inequality, just as in the proof of Lemma 4.2 in Gurvich and Whitt (2007b).

We have thus proved that the family $\hat{X}^\lambda_\Sigma(t)$ is stochastically bounded and C-tight under the asymptotic efficiency condition. To complete the proof, we apply the assumed asymptotic efficiency to deduce that

$$\hat{Q}^\lambda_\Sigma(t) - \hat{Q}^\lambda_\Sigma(s) = [\hat{X}^\lambda_\Sigma(t)]^+ - [\hat{X}^\lambda_\Sigma(s)]^+ + o(1) \quad \text{as} \quad \lambda \to \infty.$$

Consequently, the C-tightness and stochastic boundedness of $\hat{X}^\lambda_\Sigma(t)$ imply these properties for $\hat{Q}^\lambda_\Sigma(t)$. ■

We turn now to the statement of the state-space collapse result for WIR.

Theorem A.1. (**state-space collapse under WIR with pool-dependent rates**)

If $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}^\lambda(0))$ *in* $\mathbb{R}^{I+I\cdot J}$ *and* $\hat{Q}^\lambda_\Sigma(0) = 0$ *for all* λ*, then we have state-space collapse:*

$$\hat{Q}^\lambda_i(t) - \hat{Q}^\lambda_\Sigma(t)p_i \left(\hat{Q}^\lambda_\Sigma(t)\right) \Rightarrow 0 \quad \textit{in} \quad D \quad \textit{as} \quad \lambda \to \infty, \quad i \in \mathcal{I}, \tag{A2}$$

and

$$\hat{I}^\lambda_j(t) - \hat{I}^\lambda_\Sigma(t)v_j \left(\hat{I}^\lambda_\Sigma(t)\right) \Rightarrow 0 \quad \textit{in} \quad D \quad \textit{as} \quad \lambda \to \infty, \quad j \in \mathcal{J}\,. \tag{A3}$$

In addition, we have

$$a_i \hat{W}^{\lambda}_{h,i}(t) - \hat{Q}^{\lambda}_i(t) \Rightarrow 0 \quad \textit{in} \quad D \quad \textit{as} \quad \lambda \to \infty, \quad i \in \mathcal{I}. \tag{A4}$$

Proof: We outline the changes that should be made to the proof of Theorem 4.3 to accommodate the special features of WIR. Theorem 4.3 itself is proved in §4.3 of Gurvich and Whitt (2007b), as a special case of Theorem 3.1 there. We will be making frequent references to that section.

First, the definition of the stopping time σ^{λ} in equation (54) of Gurvich and Whitt (2007b) should be changed to

$$\sigma^{\lambda} := \inf\{t \geq 0 | \hat{B}^{\lambda}(t) \geq 2\hat{B}^{\lambda}(0) \vee 1\} \wedge \tilde{\sigma}^{\lambda},$$

where

$$\tilde{\sigma}^{\lambda} = \inf\{t \geq 0 : \max_{i \in \mathcal{I}} |\hat{Q}^{\lambda}_i(t) - a_i \hat{W}^{\lambda}_{h,i}(t)| \geq \epsilon^{\lambda}\},$$

with $\epsilon^{\lambda} \to 0$ as $\lambda \to \infty$ sufficiently slow (to be precisely defined towards the end of the proof).

Then, the hydrodynamic model equations, (77)-(86) of Gurvich and Whitt (2007b), are augmented by the additional equation

$$\tilde{W}_{h,i}(t) = \tilde{Q}_i(t)/a_i.$$

The proof of state-space collapse now follows identically the proof of Theorem 4.3 with the exception of Lemma 4.6 in Gurvich and Whitt (2007b), which should be slightly changed to take care of the new definition of the stopping time σ^{λ}. Specifically, we add the following argument to the proof of Lemma 4.6: First, we claim that, since state-space collapse holds on $[0, T^{\lambda}]$ and since $\hat{Q}^{\lambda}_{\Sigma}(0) = 0$ by assumption, the C-tightness of the sequence $\hat{Q}^{\lambda}(\cdot \wedge T^{\lambda}) = (\hat{Q}^{\lambda}_1(\cdot \wedge T^{\lambda}), \ldots, \hat{Q}^{\lambda}_I(\cdot \wedge T^{\lambda}))$ follows from that of $\hat{Q}^{\lambda}_{\Sigma}(t)$, which was proved in Lemma A.1. Indeed, by state-space collapse $\hat{Q}^{\lambda}_i(t \wedge T^{\lambda}) \approx \hat{Q}^{\lambda}_{\Sigma}(t \wedge T^{\lambda}) p_i \left(\hat{Q}^{\lambda}_{\Sigma}(t \wedge T^{\lambda})\right)$ with $p_i(\cdot)$ being a locally Hölder continuous function. Consequently, the tightness of $\hat{Q}^{\lambda}_{\Sigma}(t)$ implies that of $\hat{Q}^{\lambda}_i(t)$.

Now let $W^{\lambda}_i(t)$ be the virtual waiting time of class-i at time t in the λ^{th} system and $\hat{W}^{\lambda}_i(t) = \sqrt{\lambda} W^{\lambda}_i(t)$. Note that $\hat{W}^{\lambda}_i(t)$ is not necessarily equal to $\hat{W}^{\lambda}_{h,i}(t)$ as the latter refers to the cumulative waiting time of the customer at the head of the line. Having the C-tightness of $\hat{Q}^{\lambda}_i(\cdot \wedge T^{\lambda})$, we can apply the corollary in

Puhalskii [1] to establish the joint convergence

$$\left(\frac{\hat{Q}_i^\lambda(t\wedge T^\lambda)}{a_i}, \hat{W}_i^\lambda(t\wedge T^\lambda); i\in\mathcal{I}\right) \Rightarrow \left(\hat{Q}_i(t), \hat{Q}_i(t)/a_i; i\in\mathcal{I}\right) \text{ in } D^{2I} \text{ as } \lambda\to\infty,$$

where $\hat{W}_i(t) = \hat{Q}_i(t)/a_i$; see e.g. Lemma A.2 of Puhalskii and Reiman (2000). The convergence of $\hat{W}_i^\lambda(t)$ implies that the family $\{\hat{W}_i^\lambda(t)\}$ is also stochastically bounded.

Since, by definition,

$$\hat{W}_{h,i}^\lambda(t) = \hat{W}_i(t - W_{h,i}^\lambda(t)), \tag{A5}$$

we have that $\hat{W}_{h,i}^\lambda(t)$ is itself stochastically bounded and the unscaled process $W_{h,i}^\lambda(t)$ satisfies

$$W_{h,i}^\lambda(t\wedge T^\lambda) \Rightarrow 0 \text{ in } D \text{ as } \lambda\to\infty.$$

We can then apply the Random-Time-Change Theorem to equation (A5) to have the joint convergence

$$\left(\frac{\hat{Q}_i^\lambda(t\wedge T^\lambda)}{a_i}, \hat{W}_{h,i}^\lambda(t\wedge T^\lambda); i\in\mathcal{I}\right) \Rightarrow \left(\frac{\hat{Q}_i(t)}{a_i}, \frac{\hat{Q}_i(t)}{a_i}\right) \text{ in } D^{2I} \text{ as } \lambda\to\infty.$$

By Theorem 11.4.8 in Whitt (2002) and the continuity of the limit, we then have

$$\max_{i\in\mathcal{I}} \|\hat{Q}_i^\lambda - a_i\hat{W}_{h,i}^\lambda\|_{T^\lambda} \Rightarrow 0 \text{ in } \mathbb{R} \text{ as } \lambda\to\infty,$$

so that

$$P\left\{\max_{i\in\mathcal{I}} \|\hat{Q}_i^\lambda - a_i\hat{W}_{h,i}^\lambda\|_{T^\lambda} > \epsilon^\lambda\right\} \to 0 \text{ as } \lambda\to\infty$$

as long as $\{\epsilon^\lambda\}$ is such that $\epsilon^\lambda\to 0$ sufficiently slowly. Consequently, we have that $\tilde{\sigma}^\lambda\to\infty$. The rest of the proof follows Lemma 4.6 in Gurvich and Whitt (2007b), allowing us to conclude that $\sigma^\lambda\to\infty$. Hence the adaptation of the proof of Theorem 4.3 for WIR is complete. ■

It remains to relate the proof of Theorem 3.3 to that of Theorem 3.1. That is accomplished in the following proposition, which states that both cost criteria share essentially the same lower bound. For the following, we say that a family $\{b^\lambda\}$ is $o_P^\lambda(1)$ if $b^\lambda \Rightarrow 0$ as $\lambda\to\infty$.

Proposition A.1. *If $\{\pi^\lambda\} \in \Pi^e$ is a sequence of admissible policies, then*

$$J_2^\lambda(\pi^\lambda, T) \geq \int_0^T C_i^a \left(q_i^*(\hat{Q}_\Sigma^{\lambda,\pi^\lambda}(t)) \right) dt + o_P^\lambda(1) \quad as \quad \lambda \to \infty,$$

where $C_i^a(\cdot) := C_i(\cdot/a_i)$ for all $i \in \mathcal{I}$.

Proof: The proof builds on the proof of Proposition 6 in Van Meighem (1995). Since there are some differences, we give a detailed proof. Since the family $\hat{Q}_\Sigma^\lambda(t)$ is C-tight by Lemma A.1, we can choose a convergent subsequence $\{\hat{Q}_\Sigma^{\lambda^k}(t),\ k \in \mathbb{N}\}$ with $\lambda_k \to \infty$ whose limit is continuous. We will show that the result of the Proposition holds for every convergent subsequence and consequently for the whole sequence. For simplicity of presentation, we assume that $\{\lambda^k\}$ is the whole family; the reader should remember that the proof applies to the subsequence.

Denote the limit by $\hat{Q}_\Sigma(t)$. Together with the Functional Strong Law of Large Numbers (FSLLN), we have the joint convergence

$$\left(\frac{A_1^\lambda(t)}{\lambda}, \ldots, \frac{A_I^\lambda(t)}{\lambda}, \hat{Q}_\Sigma^\lambda(t) \right) \Rightarrow (a_1 t, \ldots, a_I t, \hat{Q}_\Sigma(t)) \text{ in } D^{I+1} \text{ as } \lambda \to \infty\,. \tag{A6}$$

Since the space D with the J_1 topology is separable (see §11.5 of Whitt (2002)), we can use the Skorohod representation Theorem, Theorem 3.2.2 in Whitt (2002), to construct all the processes on an alternative probability space $(\tilde{\Omega}, \tilde{\mathbb{F}}, \tilde{P})$ such that the convergence holds for almost every $\omega \in \tilde{\Omega}$. We henceforth fix such a realization ω. We consider the sequence $\{t_k,\ k \geq 0\}$ of stopping times defined recursively as follows:

$$t_{k+1} = \min\{T, \inf\{t_k < t \leq T : |\hat{Q}_\Sigma(t) - \hat{Q}_\Sigma(t_k)| \geq \epsilon\}\},$$

where $t_0 = 0$. Fix $\omega \in \Omega$. Note that since $\hat{Q}_\Sigma(t)$ is continuous on the compact interval $[0, T]$ we have that there exists $\delta > 0$ such that

$$\inf_i (t_{k+1} - t_k) > \delta. \tag{A7}$$

By Jensen's inequality,

$$\begin{aligned} J_2^\lambda(\pi^\lambda, T) &= \textstyle\sum_{i \in \mathcal{I}} \frac{1}{A_i^\lambda(t)} \sum_k \int_{t_k}^{t_{k+1}} C_i(\hat{W}_i^\lambda(s)) dA_i^\lambda(s) \\ &\geq \textstyle\sum_{i \in \mathcal{I}} \sum_k \left[\frac{1}{A_i^\lambda(t)} [A_i^\lambda(t_{k+1}) - A_i^\lambda(t_k)] \times C_i \left([A_i^\lambda(t_{k+1}) - A_i^\lambda(t_k)]^{-1} \int_{t_k}^{t_{k+1}} \hat{W}_i^\lambda(s) dA_i^\lambda(s) \right) \right] \end{aligned}$$

Since (A6) holds almost surely on $\tilde{\Omega}$ and by our choice of the realization ω, we have that

$$[A_i^\lambda(t_{k+1}) - A_i^\lambda(t_k)]/\lambda = a_i(t_{k+1} - t_k) + o^\lambda(1) \quad \text{as} \quad \lambda \to \infty\,,$$

where the approximation is uniform on $[0, T]$. Thus,

$$J_2^\lambda(\pi^\lambda, T) \geq \sum_{i\in\mathcal{I}} \sum_k \left[[(t_{k+1} - t_k) + o^\lambda(1)] \times C_i \left([\lambda(t_{k+1} - t_k + o^\lambda(1))]^{-1} \int_{t_k}^{t_{k+1}} \hat{W}_i^\lambda(s) dA_i^\lambda(s) \right) \right]. \tag{A8}$$

We now use Proposition 4 of Van Mieghem (1995), which - with the appropriate modification to our setting - states the following: Fix $0 \leq a < b \leq T$, then

$$\frac{1}{\lambda_i(b-a)} \int_a^b \hat{W}_i^\lambda(s) dA_i^\lambda(s) - \frac{1}{b-a} \int_a^b \hat{Q}_i^\lambda(s) ds \to 0, \text{ as } \lambda \to \infty,$$

with the convergence holding almost surely. The proof of this result is identical to the proof in Van Mieghem (1995), so it is omitted. Using (A7) and recalling that $a_i := \lambda_i/\lambda$, we have

$$[\lambda(t_{k+1} - t_k + o^\lambda(1))]^{-1} \int_{t_k}^{t_{k+1}} \hat{W}_i^\lambda(s) dA_i^\lambda(s) - \frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) \to 0 \text{ as } \lambda \to \infty.$$

Plugging this back into (A8), we then have

$$J_2^\lambda(\pi^\lambda, T) \geq \sum_{i\in\mathcal{I}} \sum_k \left[[(t_{k+1} - t_k) + o^\lambda(1)] \times C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) + o^\lambda(1) \right) \right]. \tag{A9}$$

Since $\hat{Q}_i^\lambda(t) \leq \hat{Q}_\Sigma^\lambda(t)$ and $\hat{Q}_\Sigma^\lambda(t)$ is bounded by its continuity on $[0, T]$, we have that $\hat{Q}_i^\lambda(t)$ is bounded. The continuity of C_i, then implies that (A9) can be written as

$$J_2^\lambda(\pi^\lambda, T) \geq \sum_{i\in\mathcal{I}} \sum_k \left[(t_{k+1} - t_k) \times C_i \left(\frac{1}{a_i} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) \right) \right] + o^\lambda(1) \quad \text{as} \quad \lambda \to \infty.$$

The C-tightness of $\hat{Q}_\Sigma^\lambda(t)$ now implies that

$$C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) \right) \geq C_i^a \left(q_i^*(\hat{Q}_\Sigma^\lambda(t_k)) \right) + o^\lambda(1) + O(\epsilon), \tag{A10}$$

where $C_i^a := C_i(\cdot/a_i)$ and $\{q_i^*(x),\ i \in \mathcal{I}\}$ is the optimal solution for (11) with the cost functions C_i^a

replacing C_i. Finally, $O(\epsilon) \to 0$ as $\epsilon \to 0$. To establish (A10), note that

$$C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^{\lambda}(s) d(s) \right) \geq C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \tilde{q}_i^{\lambda}(s) \right),$$

where $\tilde{q}_i^{\lambda}(s)$ is a solution to

$$\begin{array}{ll} \text{minimize} & \sum_{i \in \mathcal{I}} C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} q_i^{\lambda}(s) ds \right), \\ \text{s.t.} & \sum_{i \in \mathcal{I}} q_i^{\lambda}(s) = \hat{Q}_{\Sigma}^{\lambda}(s), \ s \in [t_k, t_{k+1}], \\ & q_i^{\lambda}(s) \geq 0, \ i \in \mathcal{I}, \ s \in [t_k, t_{k+1}]. \end{array}$$

However, the C-tightness of $\hat{Q}_{\Sigma}^{\lambda}(t)$ and the definition of the stopping times t_k implies that, for all λ large enough, $|\hat{Q}_{\Sigma}^{\lambda}(s) - \hat{Q}_{\Sigma}^{\lambda}(t_k)| \leq 2\epsilon$ for all $s \in [t_k, t_{k+1}]$ so that (A10) follows from the continuity of C_i.

Consequently,

$$J_2^{\lambda}(\pi^{\lambda}, T) \geq \sum_{i \in \mathcal{I}} \sum_k \left[(t_{k+1} - t_k) \times C_i^a \left(q_i^*(\hat{Q}_{\Sigma}^{\lambda}(t_k)) \right) \right] + o^{\lambda}(1) + O(\epsilon).$$

Since ϵ was arbitrary we may invoke the definition of the Rieman integral to obtain the result for almost every $\omega \in \tilde{\Omega}$. Translating this back into the original probability space yields the claimed statement. ■

Proof of Theorem 3.4: Using Proposition A.1 and the state-space collapse result in Theorem A.1, the proof now follows exactly as the proof of Theorem 3.1 with the exception of the $o_P^{\lambda}(1)$ term - whose treatment is trivial - and the replacement of Π_1 by Π^e. ■

References

[1] Puhalskii A. 1994. On the invariance principle for the first passage time, *Math. Oper. Res.* **19**(4) 946 - 954.

[2] Puhalskii A., M. Reiman. 2000. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Prob.* **32**(2) 564-595.