Electronic Companion of “Dynamic Scheduling of Multiclass Many-server Queues with Abandonment: the Generalized $c\mu/h$ Rule”

We prove Theorem 1 in §EC.1.1. The proof of Proposition 1 about the equivalence of the convergence of Q_i and B_i is presented in §EC.1.2. Then we prove Lemma 1 in §EC.1.3. We study the asymptotic analysis of the original queueing system in §EC.2. The proofs of the optimality of our proposed policies is placed in §EC.3. In §EC.3.1, we analyze the flow rates of the fluid model. We provide a proof to the optimality of the target-allocation and the $Gc\mu/h$ rule in §EC.3.2 simultaneously. The optimality of the fixed priority policy is shown in §EC.3.3. In §EC.4, we develop a dynamic programming algorithm to solve the Fractional 0-1 Knapsack Problem.

EC.1. Preliminary Analysis of the Fluid Model

In this section, we start with the analysis of the fluid model (1)–(8). Due to the fact that class-i customers arrive at the system with a constant arrival rate λ_i, we can see from (6) that

$$\eta_{i,t}([0,x]) = \lambda_i \int_0^x F_i^c(s)ds, \tag{EC.1}$$

which implies $\eta_{i,t}(dx) = \lambda_i F^c(x)dx$. This with (7) yields

$$R_i(t) = \lambda_i \int_0^t F_i(w_i(s))ds. \tag{EC.2}$$

For all $i = 1,\dots,I$, let

$$F_{i,d}(x) := \int_0^x F_i^c(y)dy. \tag{EC.3}$$

Combining (4) and (EC.1) yields $w_i(t) = F_{i,d}^{-1}(Q_i(t)/\lambda_i)$. This together with (EC.2) gives

$$R_i(t) = \lambda_i \int_0^t F_i\left(F_{i,d}^{-1}\left(\frac{Q_i(s)}{\lambda_i}\right)\right) ds. \tag{EC.4}$$

Then it follows from (5) that

$$K_i(t) = \lambda_i \int_0^t F_i^c\left(F_{i,d}^{-1}\left(\frac{Q_i(s)}{\lambda_i}\right)\right) ds - Q_i(t) + Q_i(0). \tag{EC.5}$$

We can also see from (1) and (3) that

$$B_i(t) = B_i(0) + K_i(t) - \mu_i \int_0^t B_i(s)ds,$$

of which the solution can be solved as

$$B_i(t) = B_i(0)e^{-\mu_i t} + \int_0^t e^{-\mu_i(t-s)}dK_i(s).$$

Now plugging (EC.5) into the above equation and applying integration by parts yields

$$X_i(t) = X_i(0)e^{-\mu_i t} + \lambda_i \int_0^t F_i^c\left(F_{i,d}^{-1}\left(\frac{Q_i(t-s)}{\lambda_i}\right)\right)e^{-\mu_i s}ds + \mu_i \int_0^t Q_i(t-s)e^{-\mu_i s}ds. \tag{EC.6}$$

The above equation is consistent with (3.21) in Zhang (2013). It reveals the relationship between Q_i and B_i for each class since $X_i = Q_i + B_i$, and will play a central role in the proofs of Theorem 1 and Proposition 1.

Lemma EC.1. *Consider the fluid model (1)–(8). Then all the fluid processes E_i, B_i, X_i, Q_i, D_i, K_i, R_i, i=1,...,I, are absolutely continuous.*

Proof. It is clear the arrival process E_i is absolutely continuous. The absolute continuity of D_i and R_i follows from (3) and (7), respectively. By (1) and (5), $X_i(t) = X_i(0) + E_i(t) - R_i(t) - D_i(t)$. This implies that X_i is absolutely continuous. As a result, $\sum_{i=1}^I Q_i(t) = (\sum_{i=1}^I X_i(t) - n)^+$ is also absolutely continuous. Then the absolute continuity of $\sum_{i=1}^I K_i(t)$ follows from (5). Since the entrance into service process $K_i(t)$ is nondecreasing, it follows that each K_i must be absolutely continuous. Consequently, the absolute continuity of B_i and Q_i follows from (1) and (5). This completes the proof. □

EC.1.1. Underloaded System

If the fluid model is underloaded, i.e., $\sum_{i=1}^I \lambda_i/\mu_i < n$, then any work-conserving policy will be optimal as all the queues vanish in finite time.

Proof of Theorem 1. Let

$$\mathcal{U}(t) = -\sum_{i=1}^I B_i(t) + \sum_{i=1}^I \left[X_i(0)e^{-\mu_i t} + \lambda_i \int_0^t F_i^c\left(F_{i,d}^{-1}\left(\frac{Q_i(t-s)}{\lambda_i}\right)\right)e^{-\mu_i s}ds\right]. \tag{EC.7}$$

Then we can see from (EC.6) that

$$\sum_{i=1}^I Q_i(t) = \mathcal{U}(t) + \sum_{i=1}^I \mu_i \int_0^t Q_i(t-s)e^{-\mu_i s}ds. \tag{EC.8}$$

If $\sum_{i=1}^{I} Q_i(t) = 0$, then by (EC.8), $\mathcal{U}(t) = 0 - \sum_{i=1}^{I} \mu_i \int_0^t Q_i(t-s) e^{-\mu_i s} ds \le 0$. If $\sum_{i=1}^{I} Q_i(t) > 0$, then $\sum_{i=1}^{I} B_i(t) = n$ due to the non-idling constraint (8). Since $\sum_{i=1}^{I} \lambda_i/\mu_i < n$, we can pick $\delta = (n - \sum_{i=1}^{I} \lambda_i/\mu_i)/2$, which is positive, such that

$$\begin{aligned}
&\sum_{i=1}^{I} \left[\lambda_i \int_0^t F_i^c \left(F_{i,d}^{-1} \left(\frac{Q_i(t-s)}{\lambda_i} \right) \right) e^{-\mu_i s} ds \right] \\
&= \sum_{i=1}^{I} \left[\frac{\lambda_i}{\mu_i} \mu_i \int_0^t F_i^c \left(F_{i,d}^{-1} \left(\frac{Q_i(t-s)}{\lambda_i} \right) \right) e^{-\mu_i s} ds \right] \\
&\le n - 2\delta,
\end{aligned}$$

where the last inequality follows since

$$\mu_i \int_0^t F_i^c \left(F_{i,d}^{-1} \left(\frac{Q_i(t-s)}{\lambda_i} \right) \right) e^{-\mu_i s} ds \le \mu_i \int_0^t e^{-\mu_i s} ds = 1 - e^{-\mu_i t} \le 1. \tag{EC.9}$$

For this given $\delta > 0$, there exists a T_1 such that for all $t > T_1$, $\sum_{i=1}^{I} X_i(0) e^{-\mu_i t} \le \delta$. Applying theses estimates to (EC.7), we have $\mathcal{U}(t) \le -n + \delta + n - 2\delta = -\delta$ for all t satisfying $t > T_1$ and $\sum_{i=1}^{I} Q_i(t) > 0$.

Denote by $\mathcal{S} = \{t \ge 0 : \sum_{i=1}^{I} Q_i(t) > 0\}$ the collection of time epochs when the total fluid queue length is larger than 0. Following the discussion of the above two cases, we have that $\mathcal{U}(t) \le 0$ for any $t \in [T_1, +\infty)$ and $\mathcal{U}(t) \le -\delta$ for any $t \in \mathcal{S} \cap [T_1, +\infty)$. We show that $m(\mathcal{S}) < \infty$, where m is the Lebesgue measure of real numbers. Consider the contradictory, i.e., $m(\mathcal{S}) = \infty$. Note that

$$\begin{aligned}
\int_0^\infty e^{-yt} \mathcal{U}(t) dt &= \int_0^{T_1} e^{-yt} \mathcal{U}(t) dt + \int_{T_1}^\infty e^{-yt} \mathcal{U}(t) dt \\
&\le \int_0^{T_1} |\mathcal{U}(t)| dt - \int_{\mathcal{S} \cap [T_1, +\infty)} e^{-yt} \delta dt.
\end{aligned} \tag{EC.10}$$

Since we assume $m(\mathcal{S}) = \infty$, there exists a $T_2 > T_1$ such that $\int_{\mathcal{S} \cap [T_1, T_2]} \delta dt = 2 + 2 \int_0^{T_1} |\mathcal{U}(t)| dt$. Choosing $y_0 = \frac{\ln 2}{T_2} > 0$ yields

$$\int_{\mathcal{S} \cap [T_1, +\infty)} e^{-y_0 t} \delta dt \ge e^{-y_0 T_2} \int_{\mathcal{S} \cap [T_1, T_2]} \delta dt = 1 + \int_0^{T_1} |\mathcal{U}(t)| dt.$$

So we have $\int_0^\infty e^{-y_0 t} \mathcal{U}(t) dt \le -1$ from (EC.10). On the other hand, (EC.8) implies that for all $y > 0$,

$$\int_0^\infty e^{-yt} \sum_{i=1}^{I} Q_i(t) dt = \int_0^\infty e^{-yt} \mathcal{U}(t) dt + \sum_{i=1}^{I} \left[\int_0^\infty e^{-yt} Q_i(t) dt \cdot \int_0^\infty e^{-yt} \mu_i e^{-\mu_i t} dt \right],$$

where the last term follows from the Laplace transform. Due to the fact that $\int_0^\infty e^{-yt}\mu_i e^{-\mu_i t}dt \le 1$ from (EC.9), the above implies $\int_0^\infty e^{-yt}\mathcal{U}(t)dt \ge 0$ for all $y > 0$, which is a contradiction. Hence, we have shown by contradiction that $m(\mathcal{S}) < \infty$.

Since $m(\mathcal{S}) < \infty$, for any $\varepsilon \in (0,1)$ there exists a $\tau \ge 1$ such that $m(\mathcal{S} \cap [\tau - 1, \infty)) < \varepsilon$. So for any $t \ge \tau$, there exists a $\xi \in [t-\varepsilon, t]$ such that $\sum_{i=1}^I Q_i(\xi) = 0$. The balance equation (5) implies

$$Q_i(t) \le Q_i(\xi) + \lambda_i\varepsilon = \lambda_i\varepsilon \quad \text{for all } t \ge \tau. \tag{EC.11}$$

Denote $X_{i,\tau}(t) := X_i(t+\tau)$ and $Q_{i,\tau}(t) := Q_i(t+\tau)$. In other words, we shift the fluid model by time τ. Similar to (EC.6) we have the following "shifted" version:

$$\begin{aligned}\sum_{i=1}^I X_{i,\tau}(t) &= \sum_{i=1}^I \left[X_i(\tau)e^{-\mu_i t} + \lambda_i \int_0^t F_i^c\left(F_{i,d}^{-1}\left(\frac{Q_{i,\tau}(t-s)}{\lambda_i}\right)\right)e^{-\mu_i s}ds + \mu_i \int_0^t Q_{i,\tau}(t-s)e^{-\mu_i s}ds\right]\\ &\le \sum_{i=1}^I X_i(\tau)e^{-\mu_i t} + \sum_{i=1}^I \frac{\lambda_i}{\mu_i} + \sum_{i=1}^I \lambda_i\varepsilon,\end{aligned}$$

where the inequality is due to (EC.9) and (EC.11). We can see that $X_i(\tau)e^{-\mu_i t} \to 0$ as t goes to infinity. Due to the arbitrariness of ε, taking the limsup on both sides of the above equation yields $\limsup_{t\to\infty} \sum_{i=1}^I X_{i,\tau}(t) = \sum_{i=1}^I \lambda_i/\mu_i < n$. Thus, there must exists a $T > 0$ such that $\sum_{i=1}^I Q_i(t) = 0$ for all $t > T$. Consequently, with regards to (9), we have $\lim_{T\to\infty} J_T(\pi) = 0$ for any work-conserving policy $\pi \in \Pi$. Now by (EC.6), $Q_i(t)$ vanishing in finite time implies the convergence of $B_i(t)$. It can also be seen from (EC.6) that $\lim_{t\to\infty} B_i(t) = \frac{\lambda_i}{\mu_i}$. □

EC.1.2. Equivalence of the convergence of Q_i and B_i

Proposition 1 shows that the convergence of Q_i is equivalent to that of B_i. This helps to control the system based on the status of the server pool especially when the queue length of the system is unobservable. This result will be multiply used in the proofs of the optimality of our scheduling polices.

Proof of Proposition 1. We first prove that the convergence of $Q_i(t)$ implies that of $B_i(t)$. The left-hand side of (EC.6) is nothing but $Q_i(t) + B_i(t)$ and the right-hand side of (EC.6) converges to a certain constant as t goes to infinity due to the convergence of $Q_i(t)$. Therefore, $B_i(t)$ also converges.

Now we start to prove that $B_i(t)$ converging implies the convergence of $Q_i(t)$. Assume that $\lim_{t\to\infty} B_i(t) = b_i$, where the limit b_i must satisfy $b_i \in [0, \lambda_i/\mu_i]$ for all $i \in \mathcal{I}$. Indeed, if $b_i > \lambda_i/\mu_i$, then by (1), (3) and (5),

$$X_i'(t) = \lambda_i - R_i'(t) - \mu_i B_i(t) \le -\frac{1}{2}\mu_i(b_i - \frac{\lambda_i}{\mu_i}) \tag{EC.12}$$

for all large enough t, where $R_i'(t) \ge 0$ following from (7) and the inequality holds due to the assumption $b_i > \lambda_i/\mu_i$. The above implies $X_i(t) \to -\infty$ as t goes to infinity, which is a contradiction. Thus, we have $b_i \le \lambda_i/\mu_i$ for all $i = 1, \ldots, I$. Moreover, there must be $\sum_{i=1}^I b_i = n$. Otherwise, assume to the contrary that $\sum_{i=1}^I b_i < n \le \sum_{i=1}^I \lambda_i/\mu_i$, where the last inequality is due to Assumption 1. This implies there must exist an $i_0 \in \{1, \ldots, I\}$ satisfying $b_{i_0} < \lambda_{i_0}/\mu_{i_0}$. Moreover, there will be $\sum_{i=1}^I B_i(t) < n$ for large enough t, which means all the arrivals enter into service upon arriving. For class i_0, we have for any $\epsilon > 0$ there will be $B_{i_0}(t) \le b_{i_0} + \epsilon$ for all large t. Then by (1) and (3), we have

$$B_{i_0}'(t) = \lambda_{i_0} - \mu_{i_0} B_{i_0}(t) \ge \lambda_{i_0} - \mu_{i_0}(b_{i_0} + \epsilon) \ge \frac{1}{2}(\lambda_{i_0} - \mu_{i_0} b_{i_0})$$

for small enough ϵ. The above implies $B_{i_0}(t) \to +\infty$, which is a contradiction. This proves $\sum_{i=1}^I b_i = n$. Now let

$$X_{i,\infty} := b_i + \lambda_i \int_0^{F_i^{-1}(1-\mu_i b_i/\lambda_i)} F_i^c(s) ds.$$

Plugging (EC.4) into the equation in (EC.12) yields

$$X_i'(t) = \lambda_i F_i^c \left(F_{i,d}^{-1} \left(\frac{X_i(t) - B_i(t)}{\lambda_i} \right) \right) - \mu_i B_i(t).$$

For any $\epsilon > 0$, there exists a $T_0 > 0$ such that for all $t > T_0$, $b_i - \epsilon \le B_i(t) \le b_i + \epsilon$, and as well there exists $\delta_1, \delta_2 > 0$ depending only on ϵ such that

$$X_i'(t) \le -\epsilon \quad \text{whenever } X_i(t) \ge X_{i,\infty} + \delta_1, \tag{EC.13}$$

$$X_i'(t) \ge \epsilon \quad \text{whenever } X_i(t) \le X_{i,\infty} - \delta_2, \tag{EC.14}$$

for all $t \ge T_0$, where δ_1 and δ_2 will be determined in the following. It can be easily checked that

$$\lambda_i F_i^c \left(F_{i,d}^{-1} \left(\frac{X_{i,\infty} - b_i}{\lambda_i} \right) \right) = \mu_i b_i, \tag{EC.15}$$

where $F_{i,d}^{-1}(\cdot)$ is defined in (EC.3). One can find $F_i^c(F_{i,d}^{-1}(\cdot))$ is strictly decreasing. Therefore, when $X_i(t) \geq X_{i,\infty} + \delta_1$, we have

$$X_i'(t) = \lambda_i F_i^c\left(F_{i,d}^{-1}\left(\frac{X_i(t) - B_i(t)}{\lambda_i}\right)\right) - \mu_i B_i(t) \leq \lambda_i F_i^c\left(F_{i,d}^{-1}\left(\frac{X_{i,\infty} - b_i + \delta_1 - \epsilon}{\lambda_i}\right)\right) - \mu_i(b_i - \epsilon).$$

Solving the equation

$$\lambda_i F_i^c\left(F_{i,d}^{-1}\left(\frac{X_{i,\infty} - b_i + \delta_1 - \epsilon}{\lambda_i}\right)\right) - \mu_i(b_i - \epsilon) = -\epsilon$$

yields $\delta_1 = \delta_1(\epsilon) > 0$ following from (EC.15) and the fact that $F_i^c(F_{i,d}^{-1}(\cdot))$ is strictly decreasing. Moreover, $\delta_1(\epsilon) \to 0$ as ϵ goes to zero also following from (EC.15). This determines δ_1 in (EC.13). The δ_2 in (EC.14) can be determined in a same way. Let $\mathcal{L}(t) = (X_i(t) - X_{i,\infty})^2$. Then

$$\mathcal{L}'(t) = 2(X_i(t) - X_{i,\infty})\left[\lambda_i F_i^c\left(F_{i,d}^{-1}\left(\frac{X_i(t) - B_i(t)}{\lambda_i}\right)\right) - \mu_i B_i(t)\right] \leq -2\epsilon \min\{\delta_1, \delta_2\},$$

whenever $X_i(t) \leq X_{i,\infty} - \delta_1$ or $X_i(t) \geq X_{i,\infty} + \delta_2$. So there must be a $T > T_0$ such that $X_i(t) \in (X_{i,\infty} - \delta_1, X_{i,\infty} + \delta_2)$ for all $t > T$. Since δ_1 and δ_2 can be arbitrarily small, we have $\lim_{t\to\infty} X_i(t) = X_{i,\infty}$. Thus $Q_i(t)$ also converges. More specifically, $\lim\limits_{t\to\infty} Q_i(t) = \lambda_i \int_0^{F_i^{-1}(1-\mu_i b_i/\lambda_i)} F_i^c(s)ds$. This implies (11). And we proved that the convergence of B_i and Q_i are equivalent.

In view of (EC.4) and (EC.15), we have $\lim\limits_{T\to\infty} \frac{1}{T} R_i(T) = \mu_i b_i$ for a convergent policy. Thus, the convergence of the total cost $J_T(\pi)$ immediately follows from (9) and satisfies (12) for the cost of each class. □

EC.1.3. Types of the Optimization Problem

In this paper three scheduling polices are proposed to cater the different types of the nonlinear programming (13). Lemma 1 provides sufficient conditions to each type of the optimization problem.

Proof of Lemma 1. It is evident that (14) is a nondecreasing function in b_i for convex cost function C_i and nonincreasing hazard rate function h_i. The reason is simply that $c_i(\lambda_i \int_0^x F_i^c(s)ds)\mu_i/h_i(x)$ is nondecreasing in x, so is the derivative $(d/db_i)J_i(b_i)$. Then the objective function $\sum_{i=1}^I J_i(b_i)$ is a convex function, and the optimization problem (13) is a convex programming.

On the other hand, if the cost function C_i is concave and the hazard rate function h_i is nondecreasing then the objective function $J_i(b_i)$ in (13) is a concave function of b_i. Indeed, it follows that $c_i(\lambda_i \int_0^x F_i^c(s)ds)\mu_i/h_i(x)$ becomes nonincreasing in x. Thus, the derivative $(d/db_i)J_i(b_i)$ is non-increasing in b_i. Therefore the objective function $\sum_{i=1}^I J_i(b_i)$ is a concave function, and the optimization problem (13) becomes a concave optimization. □

EC.2. The Original Queueing System

As explained in §2, our fluid model of a multiclass $G/M/n+GI$ queue follows directly from Atar et al. (2014). Since they focused on the fixed priority policy, the dynamic priority policy (16) was only proved when the priority value function is specified to be (21). Thus, we still need to prove that (16) holds under our proposed policies. This is shown in Theorem EC.1. We also prove in Theorem EC.2 that under the stochastic version of our proposed policies the fluid-scaled queueing system can also asymptotically achieve the optimal value J^* of the nonlinear programming (13).

To this end, let $(E^N, B^N, X^N, Q^N, D^N, K^N, R^N, \eta^N)$ be the prelimit stochastic processes of the fluid limits $(E, B, X, Q, D, K, R, \eta)$ defined in §2. For the Nth system, there are n^N homogeneous servers that serve customers of I classes. The arrival processes $\{E_i^N : i = 1, \ldots, I\}$ are mutually independent renewal processes with mean interarrival times $(\lambda_i^N)^{-1}$, respectively. Upon arrival customers who cannot be served immediately will join an infinite-capacity queue dedicated to their class. Each class-i customer has an independent and identically distributed (i.i.d.) patience time following distribution F_i for waiting in queue, and abandons the queue once the waiting time exceeds the patience time. Once admitted to service, a class-i customer will be served with exponentially distributed service time with mean $1/\mu_i$, i.e., the service time follow the distribution function $G_i(x) = 1 - e^{-\mu_i x}$ depending on the customer class i. Note that the service and patience time distributions are independent of N. It's worth pointing out that only η^N the measure-valued process of the buffer is needed since the measure-valued process of the server pool just becomes an auxiliary process due to the exponential service time distributions. The stochastic processes characterize exactly the same dynamics of a multiclass many-server queueing system as that of §2 in Atar et al. (2014) except the scheduling policy. Thus, we won't repeat the dynamics of the original queueing model here. The stochastic version of our policies will be proposed in (EC.19) together with three priority value functions in (EC.20), (EC.21) and (EC.22). Before that, we denote by Π^N the class of all work-conserving policies that satisfy, for all $t \geq 0$,

$$\big(n^N - \sum_{i=1}^{I} B_i^N(t)\big) \sum_{i=1}^{I} Q_i^N(t) = 0.$$

Stochastic Cost Function. Assume that each queue i incurs a per unit time queue length cost

$$C_i^N(Q_i^N(t)) = C_i(Q_i^N(t)/N), \tag{EC.16}$$

where $C_i(\cdot)$ is a nondecreasing function with the additional properties in Assumption 1 and the cost function is rescaled as the parameter N changes. Actually, the same scaling has also been used in §7 of Mandelbaum and Stolyar (2004). It also incurs a penalty cost γ_i for each class-i customer who abandons the queue before being admitted to service. For any work-conserving policy $\pi^N \in \Pi^N$, consider the rescaled average cost function

$$J_T^N(\pi^N) = \frac{1}{T} \sum_{i=1}^{I} \left[\int_0^T C_i(Q_i^N(s)/N) ds + \gamma_i R_i^N(T)/N \right], \tag{EC.17}$$

where the first term on the right-hand side of (EC.17) is due to (EC.16) and the last term about the abandon penalties is also rescaled by N. The idea of the above cost function also follows from Mandelbaum and Stolyar (2004), where the authors study the almost sure convergence of the cost function using Skorohod representation theorem. An alternative way is to consider the convergence in mean, e.g., in Atar et al. (2010, 2014) the authors consider the expectation of the cost function and the expectation in their papers can be directly appended to the queue length process due to their assumption of linear cost.

Stochastic Scheduling Policies. In the Nth system, let $P_i^N(t)$ be the priority value function of each level. Then the stochastic version of the fluid *dynamic priority policy* (15) is said to be: at time t, given that a customer is to be served by an idle server, it chooses the head-of-the-line customer from the class with index

$$i \in \arg\max_{i=1,\dots,I} P_i^N(t), \tag{EC.18}$$

where $P_i^N(t)$ is the *priority value* for class i at time t of the Nth system. If queue i with the highest priority value is empty, the idle server will check classes with the second largest priority value, so on so forth. Ties are broken arbitrarily once there are multiple queues with same priority value, for example, in favor of the smallest index i. It can be easily seen that the stochastic dynamic priority policy (EC.18) is equivalent to

$$\int_0^t \sum_{\{j=1,\dots,I:P_j^N(s)>P_i^N(s)\}} Q_j^N(s)dK_i^N(s)=0, \quad i=1,\dots,I. \tag{EC.19}$$

Note that $\sum_{\{j=1,\dots,I:P_j^N(s)>P_i^N(s)\}} Q_j^N(s)=0$ if $\{j=1,\dots,I:P_j^N(s)>P_i^N(s)\}=\emptyset$.

In the following, we consider three stochastic scheduling policies that correspond to the three fluid scheduling polices proposed in §3.

- *Target-allocation Policy*, which we denote by $\pi_{b^*}^N$ given the priority value function

$$P_i^N(t)=b_i^*-B_i^N(t)/N, \tag{EC.20}$$

where we apply the same scaling as in (EC.16) and b_i^* is an optimal solution of the nonlinear programming (13).

- *The Generalized $c\mu/h$ Rule*, which we denote by π_G^N given the priority value function

$$P_i^N(t)=\frac{c_i\big(\frac{\lambda_i^N}{N}\int_0^{F_i^{-1}(1-B_i^N(t)\mu_i/\lambda_i^N)}F_i^c(s)ds\big)\mu_i}{h_i(F_i^{-1}(1-B_i^N(t)\mu_i/\lambda_i^N))}+\gamma_i\mu_i. \tag{EC.21}$$

Recall that h_i is the hazard rate function of the patience time distribution F_i and in (10) we have $(d/dx)C_i(x)=c_i(x)$, to which we also apply the same scaling as in (EC.16).

- *Fixed Priority Policy*, which we denote by $\pi_{P^*}^N$ given the priority value function (after re-ordering the class indices if needed)

$$P_i^N(t)=I-i. \tag{EC.22}$$

Asymptotic Analysis. In the many-server heavy traffic regime, both the arrival rates λ_i^N, $i=1,\dots,I$, and the number of agents n^N increase to infinity. More precisely, as $N\to\infty$,

$$\frac{\lambda_i^N}{N}\to\lambda_i,\ i=1,\dots,I, \quad \text{and} \quad \frac{n^N}{N}\to n. \tag{EC.23}$$

Define the fluid-scaled processes $\bar{X}_i^N=N^{-1}X_i^N$ and define $\bar{E}_i^N$, $\bar{B}_i^N$, $\bar{Q}_i^N$, $\bar{D}_i^N$, $\bar{K}_i^N$, $\bar{R}_i^N$ analogously. Similarly, $\bar{\eta}_i^N=N^{-1}\eta_i^N$ for the measure-valued process. We assume that the initial states satisfy $\bar{X}_i^N(0)\Rightarrow X_i(0)$ and $\bar{\eta}_{i,0}^N\Rightarrow\eta_{i,0}$ as $N\to\infty$ for all $i=1,\dots,I$.

Theorem EC.1 (Fluid Limits). *The sequence of fluid-scaled stochastic processes* $\{(\bar{E}^N, \bar{B}^N, \bar{X}^N, \bar{Q}^N, \bar{D}^N, \bar{K}^N, \bar{R}^N, \bar{\eta}^N) : N \in \mathbb{N}\}$ *under any one of the three policies* $\pi^N_{b^*}$, π^N_G *and* $\pi^N_{P^*}$ *is tight in the Skorohod-*J_1 *topology and any subsequential limit of the fluid-scaled stochastic processes satisfies the fluid model equations* (1)–(8) *together with the fluid dynamic priority policy* (16) *specified by* π_{b^*}, π_G *and* π_{P^*} *in §3, respectively.*

Proof. Following the same argument as Theorem 4.3 of Atar et al. (2014), we can conclude that the fluid-scaled stochastic processes $\{(\bar{E}^N, \bar{B}^N, \bar{X}^N, \bar{Q}^N, \bar{D}^N, \bar{K}^N, \bar{R}^N, \bar{\eta}^N) : N \in \mathbb{N}\}$ are tight for any work-conserving policy (including our proposed policies $\pi^N_{b^*}$, π^N_G and $\pi^N_{P^*}$). From their argument, we can also conclude that any subsequential limit also satisfies the fluid model equations (1)–(8). This together with (EC.23) implies that any subsequential limit of $P^N_i(t)$ in (EC.20), (EC.21) and (EC.22) is just $P_i(t)$ in (17), (20) and (21), respectively. For notational simplicity, we still use index N for the convergent subsequence.

It remains to prove that (16) also holds under the three fluid scheduling policies π_{b^*}, π_G and π_{P^*}. By Lemma EC.1, let $K'_i(t) = (d/dt)K_i(t)$. For any fixed $i \in \{1,\ldots,I\}$, it suffices to prove that $K'_i(t) = 0$ if $\sum_{\{j=1,\ldots,I: P_j(t) > P_i(t)\}} Q_j(t) > 0$, which gives (16). So assume that there exists $t > 0$ and $j \in \{1,\ldots,I\}$ such that $P_j(t) > P_i(t)$ and $Q_j(t) > 0$. Due to the continuity of P_j and P_i (which are defined in (17), (20) and (21) for our proposed fluid policies π_{b^*}, π_G and π_{P^*}, respectively) and the continuity of Q_j by Lemma EC.1, we can conclude that for N large enough $P^N_j(s) > P^N_i(s)$ and $\bar{Q}^N_j(s) > 0$ for $|s-t| < \delta$ and some $\delta > 0$. Here, we map all the random objects to the same probability space such that all weak convergence becomes almost sure convergence by Skorohod representation theorem (see, for example, Lemma C.1 in Zhang (2013)). According to the stochastic dynamic priority policy (EC.18) (or equivalently (EC.19)), $\bar{K}^N_i(t+\delta) - \bar{K}^N_i(t-\delta) = 0$, and therefore $K_i(t+\delta) - K_i(t-\delta) = 0$. This gives us the desired result. □

Recall that J^* is the minimum value of the nonlinear programming (13) and by Proposition 1 it is actually the lower bound of any fluid convergent policies. We have proven in Theorems 2, 3 and 4 that the fluid model can achieve the minimum value J^* under the three fluid scheduling policies π_{b^*}, π_G, and π_{P^*}. For the original queueing system, our goal is similar to find a scheduling policy such that J^* can also be asymptotically achieved in the many-server heavy traffic regime. We refer to such a scheduling policy as an *asymptotically stationary optimal* control policy. The following theorem shows that the optimal value J^* can actually be asymptotically achieved under any one of the stochastic policies $\pi^N_{b^*}$, π^N_G and $\pi^N_{P^*}$. Actually, it is exactly the stochastic version of Theorems 2, 3 and 4. Since Theorem EC.1 ensures the subsequential limit of the fluid-scaled stochastic processes, we need to consider both the limit inferior and limit superior of the cost function.

Theorem EC.2 (Asymptotically Stationary Optimality of Our Policies). *Given the conditions in Theorems 2, 3 and 4 respectively, there is*

$$\liminf_{T\to\infty}\liminf_{N\to\infty} J^N_T(\pi^N) = \limsup_{T\to\infty}\limsup_{N\to\infty} J^N_T(\pi^N) = J^* \tag{EC.24}$$

almost surely, where $\pi^N = \pi^N_{b^*}$, π^N_G and $\pi^N_{P^*}$ accordingly.

Proof. We first consider the target-allocation policy $\pi^N_{b^*}$. By Theorem EC.1, for the sequence of the target-allocation policies $\{\pi^N_{b^*}\}$ we can always choose a convergent subsequence as the supremum. By Skorohod representation theorem (see, for example, Lemma C.1 in Zhang (2013)) we can map all the random objects to the same probability space so that all weak convergence becomes almost sure convergence. Thus, there is a fluid target-allocation policy π_{b^*} such that $\limsup_{N\to\infty} J^N_T(\pi^N_{b^*}) = J_T(\pi_{b^*})$ almost surely. It then follows from Theorem 2 that the second equation in (EC.24) holds. The limit inferior in (EC.24) follows due to the same reason. The proof for the other two policies π^N_G and $\pi^N_{P^*}$ is exactly the same. Thus, we omit it. □

EC.3. Proofs of the Optimality of the Fluid Scheduling Policies

EC.3.1. Flow Rates of the Fluid Model

The following lemma extends Theorem 3.2 in Atar et al. (2014) and characterizes a notable property of the dynamic priority policy that the entrance into service process can be represented by the external arrival and departure processes.

Let ${}^*I_j(t)$ be the collection of indices with the first jth highest priority value at time t recursively defined as follows:

$$^*I_1(t) = \arg\max_{i\in\{1,\dots,I\}} P_i(x), \tag{EC.25}$$

and for $1 \le j \le I$,

$$^*I_{j+1}(t) = {}^*I_j(t) \cup \arg\max_{i\in\{1,\dots,I\}\setminus {}^*I_j(t)} P_i(t).$$

Lemma EC.2. *Consider the fluid model* (1)–(8) *given any continuous priority value function $P_i(t)$. Then the entrance into service processes $K_i(t)$ are absolutely continuous, and the derivatives $K_i'(t) := (d/dt)K_i(t)$ satisfy a.e. for $j = 1,\ldots,I$,*

$$\sum_{i\in {}^*I_j(t)} K_i'(t) = \begin{cases} \sum_{i=1}^I \mu_i B_i(t) & \text{if } \sum_{i\in {}^*I_j(t)} Q_i(t) > 0, \\ [\sum_{i=1}^I \mu_i B_i(t)] \wedge \sum_{i\in {}^*I_j(t)} \lambda_i & \text{if } \sum_{i\in {}^*I_j(t)} Q_i(t) = 0, \sum_{i=1}^I B_i(t) = n, \\ \sum_{i\in {}^*I_j(t)} \lambda_i & \text{if } \sum_{i=1}^I B_i(t) < n, \end{cases} \tag{EC.26}$$

where $a \wedge b$ is the minimum of a and b.

Proof. We prove this lemma following a similar argument to Theorem 3.2 in Atar et al. (2014). The absolutely continuity of K_i has been proven in Lemma EC.1.

If $\sum_{i=1}^I B_i(t) < n$ for some t, then by the continuity of B_i's (which follows from (1) using the continuity of K_i and D_i) this holds on a neighborhood of t. For any s in such a neighborhood, it is easily seen that $Q_i(s) = 0$ by (8) and $R_i'(s) = 0$ by (EC.4). Hence, by (5), we have $K_i(s) - K_i(t) = E_i(s) - E_i(t)$. This shows $K_i'(t) = \lambda_i$ for all $i = 1,\ldots,I$.

On the other hand, if $\sum_{i\in {}^*I_j(t)} Q_i(t) > 0$, then we have $\sum_{i\in {}^*I_j(t)} Q_i(s) > 0$ for any $s \ge t$ in a right neighborhood of t by the continuity of Q_i's (which follows from (5) using the continuity of E_i, R_i, and K_i). By (8), for any s in such a neighborhood, $\sum_{i=1}^I B_i(s) = n$. We also have ${}^*I_j(s) \subset {}^*I_j(t)$ for small enough neighborhood, which is due to continuity of the priority value function. According to the definition of the dynamic priority policy (15), customers with lower priority value can be served only if those with higher priority are all in service. This together with the fact $\sum_{i\in {}^*I_j(t)} Q_i(s) > 0$ implies that there must be $K_i'(s) = 0$ for all $i \notin {}^*I_j(t)$ for small enough neighborhood. It then follows from (1) that

$$\sum_{i\in {}^*I_j(t)} K_i(s) - \sum_{i\in {}^*I_j(t)} K_i(t) = \sum_{i=1}^I D_i(s) - \sum_{i=1}^I D_i(t).$$

By (3), the above implies that $\sum_{i\in {}^*I_j(t)} K_i'(t) = \sum_{i=1}^I \mu_i B_i(t)$ if $\sum_{i\in {}^*I_j(t)} Q_i(t) > 0$.

Now we start to prove the second entry in (EC.26). Since $\sum_{i=1}^I B_i(t)$ and $\sum_{i\in {}^*I_j(t)} Q_i(t)$ are absolutely continuous, it follows that $\sum_{i=1}^I B_i'(t) = 0$ a.e. on $S_1 := \{t : \sum_{i=1}^I B_i(t) = n\}$ and $\sum_{i\in {}^*I_j(t)} Q_i'(t) = 0$ a.e. on $S_2 := \{t : \sum_{i\in {}^*I_j(t)} Q_i(t) = 0\}$ by Theorem A.6.3 in Dupuis and Ellis (1997). Moreover, from (1) and (5) we have

$$\sum_{i=1}^I B_i'(t) = \sum_{i=1}^I K_i'(t) - \sum_{i=1}^I \mu_i B_i(t),$$

$$\sum_{i\in {}^*I_j(t)} Q_i'(t) = \sum_{i\in {}^*I_j(t)} \lambda_i - \sum_{i\in {}^*I_j(t)} K_i'(t) - \sum_{i\in {}^*I_j(t)} R_i'(t).$$

Note that $R_i'(t)=0$ whenever $Q_i(t)=0$ by (EC.4). Thus a.e. on $S_1\cap S_2$, we have $\sum_{i=1}^I K_i'(t) = \sum_{i=1}^I \mu_i B_i(t)$ and $\sum_{i\in {}^*I_j(t)} K_i'(t) = \sum_{i\in {}^*I_j(t)} \lambda_i$. Hence a.e. on $S_1\cap S_2$, $\sum_{i\in {}^*I_j(t)} K_i'(t) = \sum_{i\in {}^*I_j(t)} \lambda_i(t) = [\sum_{i=1}^I \mu_i B_i(t)]\wedge \sum_{i\in {}^*I_j(t)} \lambda_i$. This completes the proof. □

EC.3.2. Optimality of the Target-allocation Policy and the $Gc\mu/h$ Rule

In view of the fact that the priority value functions go to an equal constant under both policies. We will see that the proofs of the optimality of the target-allocation policy and the $Gc\mu/h$ rule are exactly the same. Thus we prove Theorems 2 and 3 simultaneously, which is presented in the end of this subsection. Before that, some auxiliary Lemmas EC.3 – EC.6 are analyzed. First we introduce the following auxiliary functions.

For the target-allocation policy π_{b^*} proposed in §3.1, let

$$A_i(x) = \alpha_0 + b_i^* - x, \tag{EC.27}$$

where α_0 can be chosen as any constant. In order to have a same proof as the optimality of the $Gc\mu/h$ rule, we choose α_0 to be the one in (18). With a little bit abuse of notation, for the $Gc\mu/h$ rule, we also introduce $A_i(\cdot)$ as follows:

$$A_i(x) = \frac{c_i\big(\lambda_i \int_0^{F_i^{-1}(1-x\mu_i/\lambda_i)} F_i^c(u)du\big)\mu_i}{h_i(F_i^{-1}(1-x\mu_i/\lambda_i))} + \gamma_i\mu_i. \tag{EC.28}$$

Note that by (18) and (EC.27), we have

$$A_i(b_i^*) = \alpha_0 \tag{EC.29}$$

for both $A_i(\cdot)$ in (EC.27) and (EC.28). Obviously, $A_i(\cdot)$ in (EC.27) is strictly decreasing. And $A_i(\cdot)$ in (EC.28) is also a strictly decreasing function under Assumption 2. Thus, within this subsection $A_i(x)$ could be either (EC.27) or (EC.28). Now introduce

$${}^*A(B(t)) := \max_{i=1,\dots,I} A_i(B_i(t)). \tag{EC.30}$$

In view of (17) and (EC.27), for the target-allocation policy, we can consider $A_i(B_i(t))$ as the priority value function instead of the one in (17). Then ${}^*I_1(t)$ in (EC.25) can be replaced by

$${}^*I_1(t) := \{i\in\{1,\dots,I\} : A_i(B_i(t)) = {}^*A(B(t))\}, \tag{EC.31}$$

which is the collection of indices with the highest priority value at time t. And define

$$^*B_i(t) \doteq \{\zeta \geq 0 : A_i(\zeta) = {}^*A(B(t))\}. \tag{EC.32}$$

Lemma EC.3. *Consider the fluid model* (1)–(8) *given the priority value function* (17) *or* (20)*. The following properties hold at any time* $t \geq 0$*.*
(1) *The process* $B_i(t)$ *is absolutely continuous and the derivative* $B_i'(t) := (d/dt)B_i(t)$ *satisfies a.e.*

$$\sum_{i \in {}^*I_1(t)} B_i'(t) \geq 0. \tag{EC.33}$$

(2) *Moreover, if* $\sum_{i=1}^{I} {}^*B_i(t) \leq n - \delta$*, for some* $\delta > 0$*, then there exists a constant* $\epsilon_0 > 0$ *depending only on* δ *such that*

$$B_i(t) \leq b_i^* - \epsilon_0 \quad \textit{for all } i \in {}^*I_1(t), \tag{EC.34}$$

and there also exists a constant $\epsilon_1 > 0$ *depending only on* δ *such that*

$$\sum_{i \in {}^*I_1(t)} B_i'(t) \geq \epsilon_1. \tag{EC.35}$$

Proof. First, the absolute continuity of $B_i(t)$ follows from (1) and Lemma EC.2. Now, we claim that there must be $B_i(t) \leq b_i^*$ for all $i \in {}^*I_1(t)$. Suppose there exists an $i_0 \in {}^*I_1(t)$ satisfying $B_{i_0}(t) > b_{i_0}^*$. Together this with (EC.29) yields $A_{i_0}(B_{i_0}(t)) \leq A_{i_0}(b_{i_0}^*) = \alpha_0$. By (EC.31), this implies ${}^*A(B(t)) \leq \alpha_0$, which yields $A_i(B_i(t)) \leq \alpha_0$ for all $i \in \{1, \ldots, I\}$. Thus $B_i(t) \geq b_i^*$ for all $i \in \{1, \ldots, I\}$ following from (EC.29). Due to the strict inequality of $B_{i_0}(t) > b_{i_0}^*$, we obtain $\sum_{i=1}^{I} B_i(t) > n$. This contradicts (2) and then it follows $B_i(t) \leq b_i^*$ for all $i \in {}^*I_1(t)$. From (1),

$$\sum_{i \in {}^*I_1(t)} B_i'(t) = \sum_{i \in {}^*I_1(t)} K_i'(t) - \sum_{i \in {}^*I_1(t)} D_i'(t). \tag{EC.36}$$

By Lemma EC.2, the above expression is nonnegative once $\sum_{i \in {}^*I_1(t)} K_i'(t) = \sum_{i=1}^{I} D_i'(t) = \sum_{i=1}^{I} \mu_i B_i(t)$. So we just need to consider the other possible case $\sum_{i \in {}^*I_1(t)} K_i'(t) = \sum_{i \in {}^*I_1(t)} \lambda_i$ when proving (EC.33), which still holds since $D_i'(t) = B_i(t)\mu_i \leq b_i^* \mu_i \leq \lambda_i$ for all $i \in {}^*I_1(t)$. Thus (EC.33) holds.

We show that the condition $\sum_{i=1}^{I} {}^*B_i(t) \le n-\delta$ implies there exists an $\epsilon' > 0$, such that

$$B_i(t) \le b_i^* - \epsilon' \quad \text{for all } i \in {}^*I_1(t), \tag{EC.37}$$

where ϵ' depends only on the subset ${}^*I_1(t)$ and δ. Indeed, there must be $B_i(t) < b_i^*$ for all $i \in {}^*I_1(t)$ with strict inequalities. Otherwise, we will have $B_i(t) = b_i^*$ for at least one $i \in {}^*I_1(t)$, which causes ${}^*A(B(t)) = \alpha_0$ following from (EC.29) and (EC.31). Then ${}^*B_i(t) = b_i^*$ for all $i \in \mathcal{I}$ deducing from (EC.32). This is a contradiction to the assumption $\sum_{i=1}^{I} {}^*B_i(t) < n$ since $\sum_{i=1}^{I} b_i^* = n$. Therefore ${}^*A(B(t)) = \alpha_0 + \varepsilon$, for some $\varepsilon > 0$. From (EC.32), we have

$$\sum_{i=1}^{I} {}^*B_i(t) = \sum_{i=1}^{I} A_i^{-1}(\alpha_0 + \varepsilon) \le s - \delta.$$

Let ε^* satisfy $\sum_{i=1}^{I} A_i^{-1}(\alpha_0 + \varepsilon^*) = n - \delta$. There must be $0 < \varepsilon^* \le \varepsilon$ since A_i^{-1}, $i \in \{1, \ldots, I\}$, are decreasing. By (EC.31), for all $i \in {}^*I_1(t)$, $B_i(t) = A_i^{-1}(\alpha_0 + \varepsilon) \le A_i^{-1}(\alpha_0 + \varepsilon^*) = b_i^* - (b_i^* - A_i^{-1}(\alpha_0 + \varepsilon^*))$. Now let $\epsilon' = \min_{i \in {}^*I_1(t)}(b_i^* - A_i^{-1}(\alpha_0 + \varepsilon^*))$ which is positive and depends only on the subset ${}^*I_1(t) \subset \{1, \ldots, I\}$ and δ. This proves (EC.37). Because there is only a finite number of subsets of $\{1, \ldots, I\}$, we have proved (EC.34) and ϵ_0 only depends on δ.

From (3) and (EC.36), if $\sum_{i \in {}^*I_1(t)} K_i'(t) = \sum_{i \in {}^*I_1(t)} \lambda_i$, then

$$\begin{aligned} \sum_{i \in {}^*I_1(t)} B_i'(t) &= \sum_{i \in {}^*I_1(t)} \lambda_i - \sum_{i \in {}^*I_1(t)} B_i(t)\mu_i \\ &\ge \sum_{i \in {}^*I_1(t)} \lambda_i - \sum_{i \in {}^*I_1(t)} (b_i^* - \epsilon_0)\mu_i \\ &\ge \sum_{i \in {}^*I_1(t)} \mu_i \epsilon_0 \\ &\ge \min_{i \in \{1, \ldots, I\}} \mu_i \epsilon_0, \end{aligned}$$

where the first inequality uses (EC.34), the second inequality is due to the fact $\lambda_i \ge b_i^* \mu_i$. Another case is $\sum_{i \in {}^*I_1(t)} K_i'(t) = \sum_{i=1}^{I} D_i'(t)$, which happens only when $\sum_{i=1}^{I} B_i(t) = n$ deduced from Lemma EC.2. In this case the set $\{1, \ldots, I\} \setminus {}^*I_1(t)$ is nonempty, otherwise, observing (EC.34), $\sum_{i=1}^{I} B_i(t) = \sum_{i \in {}^*I_1(t)} B_i(t) < \sum_{i \in {}^*I_1(t)} b_i^* \le n$ becoming a contradiction. Then there must be an $i_1 \in \{1, \ldots, I\} \setminus {}^*I_1(t)$ satisfying $B_{i_1}(t) \ge b_{i_1}^*$. Thus, by (EC.36),

$$\sum_{i \in {}^*I_1(t)} B_i'(t) = \sum_{i \in \{1, \ldots, I\} \setminus {}^*I_1(t)} D_i'(t) \ge B_{i_1}(t)\mu_{i_1} \ge b_{i_1}^* \mu_{i_1} \ge \min_{i \in \{1, \ldots, I\}} b_i^* \mu_i.$$

Combining the above two inequalities yields (EC.35). □

It follows from (EC.27) and the absolutely continuous of $B_i(t)$ proved in Lemma EC.3 that $A_i(B_i(t))$ is absolutely continuous for the target-allocation policy. For the $Gc\mu/h$ rule, with the fact c_i and h_i are differentiable assumed in Theorem 3, the function $A_i(x)$ in (EC.28) is absolutely continuous. Thus $A_i(B_i(t))$ is also absolutely continuous for the $Gc\mu/h$ rule. This implies that ${}^*A(B(t))$ is absolutely continuous, so is ${}^*B_i(t)$ by (EC.32). Let us call such points t *strictly regular*. This concept was also used in Mandelbaum and Stolyar (2004) (see Page 847 for reference).

Lemma EC.4. *Consider the fluid model* (1)–(8) *given the priority value function* (17) *or* (20)*. Suppose t is a strictly regular point, then*

$$\frac{d}{dt}[A_i(B_i(t))] = \frac{d}{dt}[{}^*A(B(t))] \quad \text{for all } i \in {}^*I_1(t). \tag{EC.38}$$

Proof. Suppose contrarily

$$\frac{d}{dt}[A_{i_0}(B_{i_0}(t))] = \max_{i\in {}^*I_1(t)} \frac{d}{dt}[A_i(B_i(t))] > \min_{i\in {}^*I_1(t)} \frac{d}{dt}[A_i(B_i(t))] = \frac{d}{dt}[A_{i_1}(B_{i_1}(t))]$$

for some $i_0, i_1 \in {}^*I_1(t)$. There exist sequences $\{\epsilon_1^n, \epsilon_2^n\}$ both converging to 0 such that $A_{i_0}(B_{i_0}(t+\epsilon_1^n)) > A_{i_1}(B_{i_1}(t+\epsilon_1^n))$ and $A_{i_0}(B_{i_0}(t-\epsilon_2^n)) < A_{i_1}(B_{i_1}(t-\epsilon_2^n))$. Thus $\lim\limits_{s\to t+} \frac{{}^*A(B(s)) - {}^*A(B(t))}{s-t} = \lim\limits_{\epsilon_1^n \to 0} \frac{A_{i_0}(B_{i_0}(t+\epsilon_1^n)) - A_{i_0}(B_{i_0}(t))}{\epsilon_1^n} = \frac{d}{dt}[A_{i_0}(B_{i_0}(t))]$. Similarly, $\lim\limits_{s\to t-} \frac{{}^*A(B(s)) - {}^*A(B(t))}{s-t} = \frac{d}{dt}[A_{i_1}(B_{i_1}(t))] \neq \frac{d}{dt}[A_{i_0}(B_{i_0}(t))]$, which contradicts the strict regularity at t. This completes the proof. □

Lemma EC.5. *Consider the fluid model* (1)–(8) *given the priority value function* (17) *or* (20)*. The following inequalities hold for almost all $t \geq 0$,*

$${}^*A(B(t)) \geq \alpha_0, \tag{EC.39}$$

$$\frac{d}{dt}[{}^*A(B(t))] \leq 0. \tag{EC.40}$$

And if $\sum_{i=1}^I {}^*B_i(t) \leq n - \delta$, for some $\delta > 0$, then there exists an $\epsilon_1 > 0$ depending only on δ such that for almost all $t > 0$,

$$\frac{d}{dt}\Big[\sum_{i=1}^I {}^*B_i(t)\Big] \geq \epsilon_1, \tag{EC.41}$$

where ϵ_1 is given in (EC.35).

Proof. In view of (2) and the fact that $\sum_{i=1}^{I} b_i^* = n$, there must be an $i \in \{1, \ldots, I\}$ such that $B_i(t) \le b_i^*$. Then by (EC.29) and (EC.30) the inequality (EC.39) follows.

We have shown in the above of Lemma EC.4 that ${}^*A(B(t))$ and ${}^*B_i(t)$ for all $i = 1, \ldots, I$ are absolutely continuous, which means they have derivatives almost everywhere. Consider an arbitrary strictly regular point $t > 0$. We cannot have $(d/dt){}^*A(B(t)) > 0$ since by Lemma EC.4 this would imply $B_i'(t) < 0$ for all $i \in {}^*I_1(t)$. This contradicts (EC.33). So we have (EC.40).

Next we prove (EC.41). Using (EC.31), (EC.32), and (EC.38) yields ${}^*B_i'(t) = B_i'(t)$ for all $i \in {}^*I_1(t)$. Therefore,

$$\sum_{i=1}^{I} {}^*B_i'(t) \ge \sum_{i \in {}^*I_1(t)} {}^*B_i'(t) = \sum_{i \in {}^*I_1(t)} B_i'(t) \ge \epsilon_1,$$

where the first inequality comes from the fact that ${}^*B_i'(t) \ge 0$ for all $i = 1, \ldots, I$ (which is implied by (EC.40)) and the second inequality follows from (EC.35). □

The following lemma is similar to Proposition 7 in van Mieghem (1995), which is essentially a sufficient condition of the optimality of our policies.

Lemma EC.6. *Consider the fluid model* (1)–(8) *given the priority value function* (17) *or* (20). *If*

$$\max_{1 \le k,l \le I} |A_k(B_k(t)) - A_l(B_l(t))| \to 0 \quad \textit{as } t \to \infty, \tag{EC.42}$$

then the amount of fluid content in service $B_i(t)$ satisfies $\lim_{t \to \infty} B_i(t) = b_i^*$ for all $i = 1, \ldots, I$.

Proof. We first claim that for any $\epsilon_0 > 0$ and $i \in \{1, \ldots I\}$,

$$B_i(t) \le \lambda_i/\mu_i + \epsilon_0 \quad \text{for large enough } t. \tag{EC.43}$$

Otherwise, there must be an $i_0 \in \{1, \ldots, I\}$ and a subsequence $t_n \to \infty$ such that

$$B_{i_0}(t_n) > \lambda_{i_0}/\mu_{i_0} + \epsilon_0 \ge b_{i_0}^* + \epsilon_0. \tag{EC.44}$$

We have $A_{i_0}(B_{i_0}(t_n)) \le A_{i_0}(b_{i_0}^*) = \alpha_0$ by (EC.29) and the fact that $A_{i_0}(\cdot)$ is decreasing. Then by (EC.42), $A_i(B_i(t_n)) \le \alpha_0 + \epsilon'$ for all $i \ne i_0$ and large enough t_n, where $\epsilon' > 0$ could be arbitrarily small. Thus we can chose ϵ' small enough such that for all $i \ne i_0$, $B_i(t_n) \ge b_i^* - \epsilon_0/(2(I-1))$ for large enough t_n. This together with the assumption (EC.44) yields $\sum_{i=1}^{I} B_i(t_n) \ge \sum_{i=1}^{I} b_i^* + \epsilon_0/2 > n$, contradicting (2). Thus (EC.43) holds.

Now we use (EC.43) to prove

$$\lim_{t\to\infty}\sum_{i=1}^{I} B_i(t) = n. \tag{EC.45}$$

To this end, we show that for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\sum_{i=1}^{I} B_i'(t) \geq \delta \quad \text{whenever } \sum_{i=1}^{I} B_i(t) \leq n - \varepsilon. \tag{EC.46}$$

Since $\sum_{i=1}^{I} \lambda_i/\mu_i \geq n$, there must exist $i_1 \in \{1,\ldots,I\}$ such that $B_{i_1}(t) \leq \frac{\lambda_{i_1}}{\mu_{i_1}} - \frac{\varepsilon}{2I}$. Then we can choose the ϵ_0 in (EC.43) small enough such that

$$\sum_{i=1}^{I} D_i'(t) = \sum_{i\neq i_1} \mu_i B_i(t) + \mu_{i_1} B_{i_1}(t) \leq \sum_{i=1}^{I} \lambda_i - c\varepsilon,$$

where c is a small enough constant. Note that $\sum_{i=1}^{I} K_i'(t) = \sum_{i=1}^{I} \lambda_i$ whenever $\sum_{i=1}^{I} B_i(t) < n$ by (EC.26). Thereby, $\sum_{i=1}^{I} B_i'(t) \geq c\varepsilon$ is strictly positive deduced from the above and (1). Let $\delta = c\varepsilon$, then (EC.46) holds. This yields (EC.45).

Next we consider the following two cases:

Case 1: $A_i(x)$ is given in (EC.27). Fix a class, say $l \in \{1,\ldots,I\}$. Then by (EC.27) and (EC.42),

$$\lim_{t\to\infty} |b_k^* - B_k(t) - (b_l^* - B_l(t))| = 0.$$

Summing over the classes $k = 1,\ldots,I$,

$$\lim_{t\to\infty} \Big|\sum_{k=1}^{I} (b_k^* - B_k(t)) - I\cdot(b_l^* - B_l(t))\Big| = 0.$$

From (EC.45), the above implies $B_l(t) \to b_l^*$. Thus, $B_i(t) \to b_i^*$ for all $i = 1,\ldots,I$.

Case 2: $A_i(x)$ is given in (EC.28). We also fix a class, say $l \in \{1,\ldots,I\}$. The limit (EC.42) shows that for all $\epsilon_1 > 0$ there exists a T such that for all $t > T$,

$$|A_k(B_k(t)) - A_l(B_l(t))| < \epsilon_1 \quad \text{for all } k \in \{1,\ldots,I\}.$$

Since $A_k(x)$ is strictly decreasing and continuous in x according to (EC.28), its inverse A_k^{-1} is also strictly decreasing and continuous. Thus by (EC.43) and the above, for all $\epsilon > 0$ there exists a $\delta' > 0$ such that if $\epsilon_0, \epsilon_1 < \delta'$, then

$$\left|B_k(t) - A_k^{-1}(A_l(B_l(t)))\right| < \epsilon.$$

Summing over the classes $k = 1, \ldots, I$,

$$\left| \sum_{k=1}^{I} B_k(t) - \sum_{k=1}^{I} A_k^{-1}(A_l(B_l(t))) \right| < \epsilon I.$$

Because the function $\sum_{k=1}^{I} A_k^{-1}(A_l(\cdot))$ is strictly decreasing , $B_l(t)$ converges by (EC.45). The policy satisfying (EC.42) controls the service capacity such that $b^* = (b_1^*, \ldots, b_I^*)$ is the solution to the sufficient first order conditions of the minimization problem (13). Thus, $B_i(t) \to b_i^*$ for all $i = 1, \ldots, I$. Combining the above two cases yields the result of this lemma. □

Proof of Theorems 2 and 3. From the definition of ${}^*B_i(t)$ in (EC.32), we have $A_i({}^*B_i(t)) \geq A_i(B_i(t))$. Since A_i is decreasing, this inequality implies ${}^*B_i(t) \leq B_i(t)$ for all $i = 1, \ldots, I$. Then it can be seen from (2) that $\sum_{i=1}^{I} {}^*B_i(t) \leq n$. This yields $\lim\limits_{t \to \infty} \sum_{i=1} {}^*B_i(t) = n$ by (EC.41). Then, we also have $\lim_{t\to\infty} \sum_{i=1}^{I} B_i(t) = n$ following from (2). Hence, $\lim_{t\to\infty}(B_i(t) - {}^*B_i(t)) = 0$ for all $i = 1, \ldots, I$. Thus we can conclude from (EC.32) that

$$\lim_{t\to\infty} \max_{1 \leq k,l \leq I} |A_k(B_k(t)) - A_l(B_l(t))| = 0.$$

It then follows from Lemma EC.6 that $\lim\limits_{t\to\infty} B_i(t) = b_i^*$. This together with Proposition 1 yields $\lim\limits_{T\to\infty} J_T(\pi_{b^*}) = \lim\limits_{T\to\infty} J_T(\pi_G) = J^*$. Till now we complete the proof. □

EC.3.3. Optimality of the Fixed Priority Policy

Proposition 2 shows that the fluid model given any fixed priority order converges to an equilibrium with a special form as (23). For concave holding cost functions and nondecreasing hazard rate functions, Theorem 4 states that the optimal scheduling policy must be in the family of the fixed priority policies. The proof is placed in the end of this subsection.

Recall from the definition of i_0 in (23), i_0 is the biggest number such that $\sum_{i=1}^{i_0-1} \lambda_i/\mu_i$ is strictly less than n, which implies that the traffic intensity of the first $i_0 - 1$ classes with high priorities are actually underloaded. Intuitively, their queue lengths should vanish after a finite time under a fixed priority scheduling. The following lemma verifies such a phenomena and claims that the first $i_0 - 1$ queues will become empty eventually.

Lemma EC.7. *Under Assumption 1, for any class $i \in \{1, \cdots, i_0 - 1\}$, where i_0 is given in* (23)*, the queue length vanishes after a finite time and the amount of fluid content in* *service converges to* (23)*. In other words, there exists a* $T > 0$ *such that* $Q_i(t) = 0$ *for all* $t \geq T$ *and* $i \in \{1, \cdots, i_0 - 1\}$*. And*

$$\lim_{t\to\infty} B_i(t) = \lambda_i/\mu_i \quad \textit{for all } i \in \{1, \cdots, i_0 - 1\}. \tag{EC.47}$$

Proof. We prove the result by induction.

Step 1: As a first step, we show this lemma holds for $i = 1$. To prove this, we first show that $\liminf_{t\to\infty} B_1(t) \geq b_1 = \frac{\lambda_1}{\mu_1}$. Suppose that $B_1(t) \leq \frac{\lambda_1}{\mu_1} - \delta$ for some $\delta > 0$. Combining (1) with (EC.26) yields

$$B_1'(t) = K_1'(t) - D_1'(t) = \begin{cases} \sum_{i=1}^I \mu_i B_i(t) - \mu_1 B_1(t) & \text{if } Q_1(t) > 0, \\ [\sum_{i=1}^I \mu_i B_i(t)] \wedge \lambda_1 - \mu_1 B_1(t) & \text{if } Q_1(t) = 0 \text{ and } \sum_{i=1}^I B_i(t) = n, \\ \lambda_1 - \mu_1 B_1(t) & \text{if } \sum_{i=1}^I B_i(t) < n. \end{cases}$$

Then, one can easily see from the above equation that $B_1'(t) \geq c > 0$ for small constant c only depending on δ. Due to the arbitrariness of δ, the result $\liminf_{t\to\infty} B_1(t) \geq b_1 = \frac{\lambda_1}{\mu_1}$ thus follows. Now for any $\epsilon > 0$, we have $B_1(t) \geq b_1 - \epsilon$ for all large t. This together with (3), (2) and the first entry of (EC.26) implies when $Q_1(t) > 0$ we have

$$\begin{aligned} K_1'(t) = \sum_{i=1}^I D_i'(t) &\geq \mu_1 B_1(t) + \mu_{\min}(n - B_1(t)) \\ &\geq \mu_1(b_1 - \epsilon) + \mu_{\min}(n - b_1 + \epsilon) \\ &\geq \mu_1 b_1 + \frac{1}{2}\mu_{\min}(n - b_1) \end{aligned}$$

for small enough $\epsilon > 0$, where $\mu_{\min} = \min_{i=1,\ldots,I} \mu_i$. Thus, $Q_1'(t) \leq -\frac{1}{2}\mu_{\min}(n - b_1)$ whenever $Q_1(t) > 0$ from (5). Therefore there exists $t_1 > 0$ such that $Q_1(t) = 0$ for all $t \geq t_1$. Thus by Proposition 1 we have $\lim_{t\to\infty} B_1(t) = \lambda_1/\mu_1$.

Step 2: Suppose that Lemma EC.7 is true for all $i = 1, \cdots, k - 1 \in \{1, \cdots, i_0 - 1\}$, i.e.,

$$\lim_{t\to\infty} B_i(t) = b_i \quad \text{for all } i = 1, \cdots, k - 1. \tag{EC.48}$$

And there exists a $T_{k-1} > 0$ such that $\sum_{i=1}^{k-1} Q_i(t) = 0$ for all $t \geq T_{k-1}$. From this, we need to show that Lemma EC.7 continues to hold for $k \in \{1, \cdots, i_0 - 1\}$. Now by (5) we have

$\sum_{i=1}^{k-1} K_i'(t) = \sum_{i=1}^{k-1} \lambda_i$ for all $t \geq T_{k-1}$. So from (3) and (EC.26), one can see that for all $t \geq T_{k-1}$,

$$K_k'(t) = \begin{cases} \sum_{i=1}^{I} \mu_i B_i(t) - \sum_{i=1}^{k-1} \lambda_i, & \text{if } Q_k(t) > 0, \\ \lambda_k \wedge \left(\sum_{i=1}^{I} \mu_i B_i(t) - \sum_{i=1}^{k-1} \lambda_i \right) & \text{if } Q_k(t) = 0 \text{ and } \sum_{i=1}^{I} B_i(t) = n, \\ \lambda_k & \text{if } \sum_{i=1}^{I} B_i(t) < n. \end{cases} \tag{EC.49}$$

By (1),

$$B_k'(t) = \begin{cases} \sum_{i=1}^{I} \mu_i B_i(t) - \mu_k B_k(t) - \sum_{i=1}^{k-1} \lambda_i, & \text{if } K_k'(t) = \sum_{i=1}^{I} \mu_i B_i(t) - \sum_{i=1}^{k-1} \lambda_i, \\ \lambda_k - \mu_k B_k(t), & \text{if } K_k'(t) = \lambda_k. \end{cases}$$

Similar to Step 1, we also show that $\liminf_{t\to\infty} B_k(t) \geq b_k = \frac{\lambda_k}{\mu_k}$. Suppose that $B_k(t) \leq \frac{\lambda_k}{\mu_k} - \delta$ for some $\delta > 0$. From (EC.48) and the above, one can conclude that $B_k'(t) \geq c > 0$ for a small constant c only depending on δ. As a consequence, we have $\liminf_{t\to\infty} B_k(t) \geq b_k = \frac{\lambda_k}{\mu_k}$. Note that $\mu_i b_i = \lambda_i$ for all $i \in \{1, \cdots, i_0 - 1\}$. Thus (EC.48) implies that for any $\epsilon > 0$

$$\sum_{i=1}^{k-1} \lambda_i - \epsilon \leq \sum_{i=1}^{k-1} \mu_i B_i(t) \leq \sum_{i=1}^{k-1} \lambda_i + \epsilon$$

for all large t. According to the above proved limit inferior of $B_k(t)$, for any $\epsilon' > 0$, we have $B_k(t) \geq b_k - \epsilon'$ for all large t. When $Q_k(t) > 0$, using (2) and (EC.49), we have

$$\begin{aligned} K_k'(t) &= \sum_{i=1}^{I} \mu_i B_i(t) - \sum_{i=1}^{k-1} \lambda_i \\ &\geq \sum_{i=1}^{k-1} (\mu_i B_i(t) - \lambda_i) + \mu_k B_k(t) + \mu_{\min}(n - \sum_{i=1}^{k} B_i(t)) \\ &\geq -\epsilon + (\mu_k - \mu_{\min})(b_k - \epsilon') + \mu_{\min} \left(n - \sum_{i=1}^{k-1} (b_i + \epsilon) \right) \\ &= \mu_k b_k + \mu_{\min}(n - \sum_{i=1}^{k} b_i) - \epsilon - \epsilon' \mu_k - (k-1)\epsilon' \mu_{\min} + \epsilon' \mu_{\min} \\ &\geq \mu_k b_k + \frac{1}{2} \mu_{\min}(n - \sum_{i=1}^{k} b_i) \end{aligned} \tag{EC.50}$$

for small enough $\epsilon, \epsilon' > 0$. The above and (5) implies $Q_k'(t) \leq -\frac{1}{2}\mu_{\min}(n - \sum_{i=1}^{k} b_i)$ whenever $Q_k(t) > 0$. Therefore, there exists a t_k such that $Q_k(t) = 0$ for all $t \geq t_k$. Therefore, the result $\lim_{t\to\infty} B_k(t) = \lambda_k / \mu_k$ follows from Proposition 1. □

With Lemma EC.7, we now proceed with the proof of Proposition 2.

Proof of Proposition 2. Lemma EC.7 shows that the first i_0-1 classes with high priorities satisfy $\lim_{t\to\infty} B_i(t)=b_i$ and there exists a T such that $Q_i(t)=0$, $t\ge T$, for all $i\in\{1,\dots,i_0-1\}$. And $\sum_{i=1}^{i_0-1} K_i'(t)=\sum_{i=1}^{i_0-1}\lambda_i$ for all $t\ge T$ from (5) and (7). Then it follows from (3) and (EC.26) that for all $t\ge T$,

$$K_{i_0}'(t)=\begin{cases}\sum_{i=1}^{I}\mu_i B_i(t)-\sum_{i=1}^{i_0-1}\lambda_i & \text{if } Q_{i_0}(t)>0,\\ \lambda_{i_0}\wedge\left(\sum_{i=1}^{I}\mu_i B_i(t)-\sum_{i=1}^{i_0-1}\lambda_i\right) & \text{if } Q_{i_0}(t)=0 \text{ and } \sum_{i=1}^{I} B_i(t)=n,\\ \lambda_{i_0} & \text{if } \sum_{i=1}^{I} B_i(t)<n.\end{cases}\tag{EC.51}$$

In order to complete the proof of this theorem, a critical step is to prove $\lim_{t\to\infty} B_{i_0}(t)=b_{i_0}=n-\sum_{i=1}^{i_0-1}\frac{\lambda_i}{\mu_i}$, which is less than or equal to λ_{i_0}/μ_{i_0} according to the definition of i_0 in (23). Deducing from (2), (23) and (EC.47), there must be $\limsup_{t\to\infty} B_{i_0}(t)\le b_{i_0}$. Then it suffices to show that $\liminf_{t\to\infty} B_{i_0}(t)\ge b_{i_0}$. To this end, we consider the following two cases.

Case 1: $i_0=I$. Suppose that $B_{i_0}(t)\le b_{i_0}-\delta$ for some $\delta>0$. For large enough t, this could happen only when $\sum_{i=1}^{I} B_i(t)<n$. Otherwise, we have $\sum_{i=1}^{I} B_i(t)=n$. And by (2) and (EC.47) this causes $B_{i_0}(t)=n-\sum_{i=1}^{i_0-1} B_i(t)>b_{i_0}-\delta$ for all large enough t. So we just need to consider $\sum_{i=1}^{I} B_i(t)<n$. Then by (1) and (EC.51), $B_{i_0}'(t)=\lambda_{i_0}-\mu_{i_0}B_{i_0}(t)\ge\mu_{i_0}\delta$. This implies $\liminf_{t\to\infty} B_{i_0}(t)\ge b_{i_0}$. Combining the limit superior in the above, it immediately follows $\lim_{t\to\infty} B_{i_0}(t)=b_{i_0}$.

Case 2: $i_0<I$. Deduce from (1) and (EC.51) that

$$B_{i_0}'(t)=\begin{cases}\sum_{i=1}^{I}\mu_i B_i(t)-\mu_{i_0}B_{i_0}(t)-\sum_{i=1}^{i_0-1}\lambda_i, & \text{if } K_{i_0}'(t)=\sum_{i=1}^{I}\mu_i B_i(t)-\sum_{i=1}^{i_0-1}\lambda_i,\\ \lambda_{i_0}-\mu_{i_0}B_{i_0}(t), & \text{if } K_{i_0}'(t)=\lambda_{i_0}.\end{cases}$$

Here we also suppose that $B_{i_0}(t)\le b_{i_0}-\delta$ for some $\delta>0$. Together this with the above equation, one can find that if $K_{i_0}'(t)=\sum_{i=1}^{I}\mu_i B_i(t)-\sum_{i=1}^{i_0-1}\lambda_i$, then

$$\begin{aligned}B_{i_0}'(t)&=\sum_{i=1}^{i_0-1}[\mu_i B_i(t)-\lambda_i]+\sum_{i=i_0+1}^{I}\mu_i B_i(t)\\ &\ge\sum_{i=1}^{i_0-1}[\mu_i B_i(t)-\lambda_i]+\mu_{\min}(n-\sum_{i=1}^{i_0-1}B_i(t)-b_{i_0}+\delta)\\ &\ge\frac{1}{2}\mu_{\min}\delta,\end{aligned}$$

where the last inequality follows from (EC.47). If $K'_{i_0}(t) = \lambda_{i_0}$, then $B'_{i_0}(t) \geq \lambda_{i_0} - \mu_{i_0} b_{i_0} + \mu_{i_0}\delta \geq \mu_{i_0}\delta$. It then follows that $\liminf_{t\to\infty} B_{i_0}(t) \geq b_{i_0}$. As argued in the above this implies $\lim_{t\to\infty} B_{i_0}(t) = b_{i_0}$. Apparently, together this with (2) and (EC.47) yields $\lim_{t\to\infty} B_i(t) = 0$ for all $i = i_0 + 1, \cdots, I$. The convergence of queue length processes can be seen from Proposition 1. This completes the proof. □

Proof of Theorem 4. We claim that there exists a global minimum for which $0 < b_i < \lambda_i/\mu_i$ for at most one index i. From Lemma 1, the nonlinear programming (13) is a concave optimization problem if the cost functions C_i's are concave and the hazard rate functions h_i's are nondecreasing. Note that the constraint set is a convex set (acutally a convex polytope), then it follows that the optimization problem admits a global minimum at an extreme point, i.e., at one the vertices of this polytope. And at a vertex we have that $0 < b_i < \lambda_i/\mu_i$ for at most one index i. Corresponding to any optimal vertex, we can define an optimal fixed priority order. Then this theorem immediately follows from Propositions 1 and 2 (after re-ordering the class indices if needed). □

EC.4. Dynamic Programming Algorithm

This section is devoted to developing a dynamic programming (DP) algorithm to solve the Fractional 0-1 Knapsack Problem (28). It is easy to see that there exists a straightforward algorithm, especially when K is relatively small. According to each possible order of items, items are packed into the knapsack until the weight limit W is reached. Note that the last item packed might be divided. After evaluating all of the sequences, the optimal solution and the maximum value can be determined. However, such a brute-force algorithm is NP-hard. Fortunately, the DP algorithm of the classical 0-1 Knapsack Problem inspired us to develop a dynamic programming to solve it efficiently.

A DP Algorithm for the Fractional 0-1 Knapsack Problem. We determine how to optimally pack items into a knapsack, allowing at most one item to be divided, using a four-step procedure.

Step 1: Decompose the problem into subproblems.

In view of (28), any feasible solution contains at most one fractionally packed item. This suggests constructing a three-dimensional array $M[0..K, 0..W, 0..K]$, where the third dimension is used to track the fractionally packed item. For $1 \leq k \leq K$, $0 \leq w \leq W$ and $0 \leq l \leq K$, we consider the following two cases:

Case 1: $l = 0$. The entry $M[k, w, 0]$ stores the maximum rewarded value of items packed in their entirety from any subset of items $\{1, 2, \ldots, k\}$ with total weight at most w. The component 0 in $M[k, w, 0]$ indicates that there is no fractionally packed item.

Case 2: $l \neq 0$. The entry $M[k, w, l]$ stores the maximum rewarded value of the fractionally packed item l and the items packed in their entirety from any subset of items $\{1, 2, \cdots, k\} \setminus \{l\}$ with total weight at most w.

We also need the following initial setting for $k = 0$,

$$M[0, w, l] = \begin{cases} 0 & \text{if } l = 0, \\ V_l(w) & \text{if } l > 0 \text{ and } w_l > w, \\ -\infty & \text{if } l > 0 \text{ and } w_l \leq w. \end{cases} \tag{EC.52}$$

The first entry means no item is packed in the knapsack. The second one implies that item l is fractionally packed with weight w since its full weight w_l exceeds the weight limit w. The third entry is illegal, since item l cannot be divided. Thus, we simply set the value to be $-\infty$. For the case with weight limit $w < 0$, which is also illegal, we set

$$M[k, w, l] = -\infty \quad \text{for all } w < 0 \text{ and } k, l \geq 0. \tag{EC.53}$$

Step 2: Recursively define the value of an optimal solution.

We use the above notations to define the rewarded value of an optimal solution recursively. Similar to the definition of $M[k, w, l]$, we recursively define it for two cases as well. For $l = 0$, which means no item is fractionally packed, the optimal solution corresponding to $M[k, w, 0]$ is to either leave item k behind, in which case $M[k, w, 0] = M[k-1, w, 0]$, or pack item k, in which case $M[k, w, 0] = V_k(w_k) + M[k-1, w-w_k, 0]$ given $w_k \leq w$. Due to the penalty for a negative weight in (EC.53), we conclude that

$$M[k, w, 0] = \max\{M[k-1, w, 0], V_k(w_k) + M[k-1, w-w_k, 0]\} \tag{EC.54}$$

for all $1 \leq k \leq K$, $0 \leq w \leq W$. Actually, (EC.54) is exactly the recursive equation of the classical 0-1 Knapsack Problem (see §2.6 in Martello and Toth (1990)). For $l = 1, \ldots, K$, where item l is exactly the fractionally packed item, we can similarly derive

$$M[k, w, l] = \begin{cases} M[k-1, w, l] & \text{if } k = l, \\ \max\{M[k-1, w, l], V_k(w_k) + M[k-1, w-w_k, l]\} & \text{if } k \neq l. \end{cases} \tag{EC.55}$$

for all $1 \leq k \leq K$, $0 \leq w \leq W$, where the first entry means that item k has been fractionally packed, and thus it cannot also be packed in its entirety. The second entry relies on a similar explanation to that of (EC.54). Since this time item k is not the fractionally packed item, it can be either left behind or packed in the optimal solution corresponding to the maximum value $M[k,w,l]$.

We show in the proposition below that these recursions can indeed be described by a single recursive equation.

Proposition EC.1 (Recursive Equation). *The Fractional 0-1 Knapsack Problem* (28) *can be solved using dynamic programming. Namely, for any $l \in \{0,1,\ldots,K\}$, we have the following recursive equation*

$$M[k,w,l] = \max\left\{M[k-1,w,l], V_k(w_k) + M[k-1,w-w_k,l] + \mathrm{Inf}\mathbf{1}_{\{k=l\}}\right\}, \qquad \text{(EC.56)}$$

holds for all $k \in \{1,\ldots,K\}$ and $w \in \{0,1,\ldots,W\}$, where $\mathrm{Inf} = -\infty$.

Proof. From the condition of this proposition, only $k \geq 1$ should be considered and $n = 0$ for the boundary condition has been given in (EC.52). Thus, it's easy to see that the recursions (EC.54) and (EC.55) can be expressed as a unified equation (EC.56). In order to prove (EC.56), we first consider a possible case $k = l$, which implies that item k is the fractionally added item. Then $M[k,w,l] = M[k-1,w,l]$ since in this case item k cannot be wholly taken. It remain to prove the case $k \neq l$. To compute $M[k,w,l]$ we note that there are only two choices for item k. If we leave the whole item k, then limited by the maximum weight w the maximum reward with the wholly added items taken from $\{1,2,\cdots,k-1\}$ and the fractionally added being item l is $M[k-1,w,l]$. If instead we take the whole item k (only possible if $w \geq w_k$), then we gain $V_k(w_k)$ immediately, but consume w_k weight of our storage. Now the rest weight limit becomes $w - w_k$, then the maximum reward with the remaining items $\{1,2,\cdots,k-1\}$ is $M[k-1,w-w_k,l]$. In all, we obtain $V_k(w_k) + M[k-1,w-w_k,l]$. Note that if $w < w_k$, then $M[k-1,w-w_k,l] = -\infty$ from (EC.53). So the recursion (EC.56) holds in both cases. □

Step 3: Compute the value of an optimal solution.

For any fixed $l \in \{0,1\ldots,K\}$, the above recursive equation (EC.56) suggests a two-dimensional recursive equation. In all, there are $K+1$ independent recursive equations. To reach our goal, we just need to recursively calculate $K+1$ two-dimensional recursions for $k \in \{1,\ldots,K\}$ and $w \in \{0,1,\ldots,W\}$ based on the boundary conditions (EC.52) and (EC.53). Thus the running time of the dynamic programming algorithm is $O(K^2W)$. Finally the optimal value of the Fraction 0-1 Knapsack Problem (28) is obtained as follows:

$$\max \sum_{k=1}^{K} V_k(y_k) = \max_{l \in \{0,1,\ldots,K\}} M[K, W, l]. \tag{EC.57}$$

Step 4: Construct an optimal solution.

From (EC.57), we find that $\mathit{Frac} := \arg\max_{l \in \{0,1,\ldots,K\}} M[K, M, l]$ is the index of the fractionally packed item of the optimal solution. The only remaining problem is to obtain the indices of the items that are packed in their entirety. To that end, we need one auxiliary three-dimensional array $\mathcal{T}[0..K, 0..W, 0..K]$ to be a Boolean array to find their indices. Each entry $\mathcal{T}[k, w, l]$ records whether item k is packed in its entirety in realizing the highest value $M[k, w, l]$. That is, $\mathcal{T}[k, w, l] = 1$ if item k is packed in its entirety and $\mathcal{T}[k, w, l] = 0$ otherwise. In the optimal solution, item K is packed in its entirety if $\mathcal{T}[K, W, \mathit{Frac}] = 1$. We can now repeat this argument for $\mathcal{T}[K-1, W - w_K, \mathit{Frac}]$. And item K is not packed in its entirety if $\mathcal{T}[K, W, \mathit{Frac}] = 0$. In this case, we can repeat the argument for $\mathcal{T}[K-1, W, \mathit{Frac}]$. Iterating the argument K times from item K downward to item 1 will give the indices of all items that are packed in their entirety.

Thus far we have identified the optimal value and the solution to (28). The step-by-step procedures are described in Algorithm 1.

Remark EC.1. To the best of our knowledge, the problem (28) was only studied in Burke et al. (2008). They also proposed an exact algorithm to solve that problem. The complexity of their approach is $O(UK^2W)$, where $U = \max_{k=1,\ldots,K} w_k$. The additional U is needed because they have to further calculate each possible value of the fractionally packed item. In contrast, the complexity our algorithm is only $O(K^2W)$ as shown in Step 3. Obviously, our proposed dynamic programming algorithm is more efficient. Note that the classical 0-1 Knapsack Problem needs $O(KW)$ time. More importantly, Propositions 3 and 4 reveal the internal connection between queueing and knapsack problems.

Algorithm 1 The Fractional 0-1 Knapsack (Dynamic Programming)

```
procedure Initialization according to (EC.52)
procedure Recursively define values
for $k \leftarrow 1$ to $K$ do
    for $w \leftarrow 0$ to $W$ do
        for $l \leftarrow 0$ to $K$ do
            if $w_k \leq w$ and $k \neq l$ and $M[k-1,w,l] < V_k(w_k) + M[k-1,w-w_k,l]$ then
                begin
                $M[k,w,l] \leftarrow V_k(w_k) + M[k-1,w-w_k,l]$
                $\mathcal{T}[k,w,l] \leftarrow 1$
                end
            else
                begin
                $M[k,w,l] \leftarrow M[k-1,w,l]$
                $\mathcal{T}[k,w,l] \leftarrow 0$
                end
procedure Search for the optimal value and the fractionally packed item
$Max \leftarrow M[K,W,0]$; $Frac \leftarrow 0$
for $k \leftarrow 1$ to $K$ do
    if $M[K,W,l] > Max$ then
        $Max \leftarrow M[K,W,l]$; $Frac \leftarrow l$
procedure Find indices of items packed in their entirety
$S \leftarrow W$
for $k \leftarrow K$ to $1$ do
    if $\mathcal{T}[k,S,Frac] = 1$ then
        $S \leftarrow S - w_k$; output $k$
```

References

Atar, R., C. Giat, and N. Shimkin (2010). The $c\mu/\theta$ rule for many server queues with abandonment. *Oper. Res. 58*(5), 1427–1439.

Atar, R., H. Kaspi, and N. Shimkin (2014). Fluid limits for many-server systems with reneging under a priority policy. *Math. Oper. Res. 39*(3), 672–696.

Burke, G. J., J. Geunes, H. Edwin Romeijn, and A. Vakharia (2008). Allocating procurement to capacitated suppliers with concave quantity discounts. *Operations Research Letters 36*(1), 103–109.

Dupuis, P. and R. S. Ellis (1997). *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley Series in Probability and Statistics. Wiley.

Mandelbaum, A. and A. L. Stolyar (2004). Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res. 52*(6), 836–855.

Martello, S. and P. Toth (1990). *Knapsack problems: algorithms and computer implementations.* Wiley-Interscience series in discrete mathematics and optimization. J. Wiley & Sons.

van Mieghem, J. A. (1995). Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab. 5*(3), 809–833.

Zhang, J. (2013). Fluid models of many-server queues with abandonment. *Queueing Syst. 73*(2), 147–193.