Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with non-degenerate slowdown

Rami Atar*† Nir Solomon†

August 26, 2010; revised January 25, 2011

Abstract

A parallel server system is considered, with $\mathbf{I}$ customer classes and many servers, operating at a heavy traffic diffusion regime where the queueing delay and service time are of the same order of magnitude. Denoting by $\widehat{X}^n$ and $\widehat{Q}^n$ the diffusion scale deviation of the headcount process from the quantity corresponding to the underlying fluid model and, respectively, the diffusion scale queue-length, we consider minimizing r.v.'s of the form $c_X^n = \int_0^u C(\widehat{X}^n(t))dt$ and $c_Q^n = \int_0^u C(\widehat{Q}^n(t))dt$ over policies that allow for service interruption. Here, $C : \mathbb{R}^{\mathbf{I}} \to \mathbb{R}_+$ is continuous and $u > 0$. Denoting by θ the so called workload vector, it is assumed that $C^*(w) := \min\{C(q) : q \in \mathbb{R}_+^{\mathbf{I}}, \theta \cdot q = w\}$ is attained along a continuous curve as w varies in $\mathbb{R}_+$. We show that any weak limit point of c_X^n stochastically dominates the r.v. $\int_0^u C^*(W(t))dt$ for a suitable reflected Brownian motion W, and construct a sequence of policies that asymptotically achieve this lower bound. For c_Q^n, an analogous result is proved when, in addition, C^* is convex. The construction of the policies takes full advantage of the fact that in this regime the number of servers is of the same order as the typical queue-length.

1 Introduction

Gurvich, Mandelbaum, Shaikhet, and Whitt ([5], [8], [11] and see further references in [1]) analyzed a many-server queueing system (in the form of an M/M/N queue) in a critically loaded diffusion regime that is unique in that the typical queueing delay and service time are of the same order of magnitude. In [1] it was proposed to refer to this as the *non-degenerate slowdown* (NDS) diffusion regime, and the setting was extended to cover heterogeneous servers. Moreover, the sojourn time asymptotics were identified and shown to be distinct from the many-server regime of Halfin and Whitt [6], as well as the conventional heavy traffic (where a critically loaded system with a *fixed* number of servers is considered). In both the conventional and Halfin-Whitt heavy traffic regimes there has been work on control problems for queueing models to achieve asymptotic optimality. An

*Research supported in part by the ISF (Grant 1349/08), the US–Israel BSF (Grant 2008466), and the Technion's fund for promotion of research

†Department of Electrical Engineering, Technion, Haifa 32000, Israel

important work in this direction is that of van Mieghem [10] on a multi-class single-server queueing model in conventional heavy traffic, where a generalized $c\mu$ rule is proposed and proved to be asymptotically optimal for minimizing a convex delay-related cost. The goal of this paper is to study a similar control problem in a setting with a single pool of identical servers, parameterized in the NDS regime. The assumption on the service time distribution is more restrictive than in [10] in that the service time are exponential rather than general. Also, the cost we focus on is associated with headcount and queue-length processes, rather than delay. On the other hand, apart from working in the NDS regime, a main novelty of our approach is the ability to handle a cost function of quite a general structure.

We denote by $\widehat{X}^n$ and $\widehat{Q}^n$ the diffusion scale deviation of the headcount process from the quantity corresponding to the underlying fluid model and, respectively, the diffusion scale queue-length (see precise definitions in Section 2). We consider minimizing r.v.'s of the form

$$c_X^n = \int_0^u C(\widehat{X}^n(t))dt \quad \text{and} \quad c_Q^n = \int_0^u C(\widehat{Q}^n(t))dt, \tag{1}$$

where $C : \mathbb{R}^{\mathbf{I}} \to \mathbb{R}_+$ is a continuous function, non-decreasing in the usual partial order, and $u > 0$ is a constant. We show that any weak limit point of c_X^n stochastically dominates the r.v.

$$\int_0^u C^*(W(t))dt \tag{2}$$

for a suitable one-dimensional reflected Brownian motion W. Denoting by θ the so called *workload vector* (see Section 2), we then assume that $C^*(w) := \min\{C(q) : q \in \mathbb{R}_+^{\mathbf{I}}, \theta \cdot q = w\}$ is attained along a continuous curve, as w varies in $\mathbb{R}_+$. The minimizing curve is denoted by f. We define an interruptible service policy based on the function f, that attempts to keep

$$\widehat{X}^n(t) \approx f(\theta \cdot \widehat{X}^n(t))$$

for large values of n, so that $\widehat{X}^n$ evolves close to the minimizing curve. We show that, as a consequence, the policy asymptotically achieves the lower bound (2). For c_Q^n, analogous lower bound and asymptotic attainability result are proved when, in addition, C^* is convex.

While in conventional heavy traffic the difference between $\widehat{X}^n$ and $\widehat{Q}^n$ (or rather between processes defined analogously to $\widehat{X}^n$ and $\widehat{Q}^n$) is negligible when n is large, they may differ significantly in the NDS regime due to the fact that the number of servers is large (while $\widehat{Q}^n$ is non-negative, $\widehat{X}^n$ may assume negative values that are $O(1)$). The results that we obtain are indeed different, with those for $\widehat{X}^n$ stronger than those for $\widehat{Q}^n$. The cost considered by [10] in conventional heavy traffic is of the form $\sum_{i=1}^I \sum_{k \in K_i} C_i(\tau_k)$, where I is the number of customer classes, i is an index to the class, K_i is the set of all class-i customers arriving within a given finite time horizon, C_i are convex nondecreasing functions, and τ_k is the delay experienced by customer k. It is well understood (and was used, for example, by Mandelbaum and Stolyar [9]) that delay costs are closely related to headcount and queue-length costs. Recast in terms of headcount or queue-length, the cost of [10] corresponds to (1), where $C(x)$ takes the form $\sum_{i=1}^I C_i(x_i)$, and C_i are convex. Viewed this way, the assumption on C made in this paper is much weaker, as far as the results on c_X^n are concerned, and in particular, C need not be of sum form nor convex. Moreover, only partial convexity is needed for the results regarding the c_Q^n cost (i.e., that of C^*).

To work in the NDS regime we consider a sequence of systems indexed by n, and scale the number of servers as $O(n^{1/2})$, the arrival rates as $O(n)$ and the individual service rates as $O(n^{1/2})$. The form of the proposed policy takes full advantage of the fact that the typical queue length $O(n^{1/2})$ is of the same order as the number of servers.

In a work in progress [2] the problem is studied for a multiple pool model with non-interruptible service policies.

The model and results are presented in Section 2. The proofs appear in Section 3.

2 Model and main results

We use the following notation. For $x \in \mathbb{R}$, $x^{\pm} = \max(\pm x, 0)$. For $x \in \mathbb{R}^k$, $\|x\| = \sum_{i=1}^k |x_i|$. For $x : [0,u] \to \mathbb{R}^k$ and $t \in [0,u]$ denote $\|x\|_t^* = \sup_{s\in[0,t]} \|x(s)\|$, and in the case $k = 1$ write $|x|_t^*$ in place of $\|x\|_t^*$.

2.1 The queueing model and asymptotic regime

Let a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ be given. Unless otherwise stated, the stochastic processes introduced below are defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We write $\mathbb{E}$ for expectation w.r.t. $\mathbb{P}$. We consider a parallel server system with $\mathbf{I}$ classes of customers and a pool of identical servers. The index set for the classes is $\mathcal{I} = \{1, 2, \ldots, \mathbf{I}\}$. We will consider a sequence of systems, indexed by n, that have the same structure but differ in the values of the parameters.

The arrivals are modeled as independent renewal processes, denoted by A_i^n, $i \in \mathcal{I}$. To define these processes, let A_i, $i \in \mathcal{I}$ be independent renewal processes, where, for each i, the time of the first arrival and the inter-arrival times are positive i.i.d. random variables with mean 1 and variance $(C_{i,IA})^2 \geq 0$. The processes A_i^n are defined from A_i via time acceleration, namely

$$A_i^n(t) = A_i(\lambda_i^n t), \qquad t \geq 0,\, i \in \mathcal{I}.$$

The acceleration parameters satisfy $\lim_n \lambda_i^n/n = \lambda_i > 0$, and moreover

$$\widehat{\lambda}_i^n := n^{-1/2}(\lambda_i^n - n\lambda_i) \to \widehat{\lambda}_i \in (-\infty, \infty), \quad i \in \mathcal{I}, \tag{3}$$

as $n \to \infty$. The index n will take values in $\mathbb{N}' = \{k^2 : k \in \mathbb{N}\}$, so that $n^{1/2}$ will always be integer. The number of servers in the nth system is given by

$$N^n = n^{1/2}. \tag{4}$$

Service time distributions are exponential, with class-dependent parameter. We denote by μ_i^n the rate at which a class-i customer is served, and assume

$$\mu_i^n = n^{1/2}\mu_i, \quad i \in \mathcal{I}, \tag{5}$$

for some constants $\mu_i > 0$. The traffic intensity for class i, namely $\lambda_i^n/(N^n\mu_i^n)$, has the limit $\rho_i = \lambda_i/\mu_i$, $i \in \mathcal{I}$. It is assumed that the system is critically loaded, namely

$$\sum_{i\in\mathcal{I}} \rho_i = 1.$$

Let B_i^n represent the number of servers working on class-i customers. Let X_i^n, Q_i^n and I^n denote the number of class-i customers in the system, the number of class-i customers in the buffer, and the number of servers that are idle, respectively. Note that

$$X_i^n = Q_i^n + B_i^n, \qquad i \in \mathcal{I}, \tag{6}$$

$$N^n = I^n + \sum_{i\in\mathcal{I}} B_i^n. \tag{7}$$

We are given standard Poisson processes S_i, $i \in \mathcal{I}$. The number of service completions of class-i jobs by time t is given by

$$D_i^n(t) = S_i(T_i^n(t)), \tag{8}$$

where

$$T_i^n(t) = \mu_i^n \int_0^t B_i^n(s)ds. \tag{9}$$

We have

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t). \tag{10}$$

The processes A_i^n, S_i, X_i^n, Q_i^n, B_i^n, I^n will always be assumed to have càdlàg sample paths. For each n, the processes A_i^n, $i \in \mathcal{I}$, S_i, $i \in \mathcal{I}$, and the initial condition $((B_i^n(0), i \in \mathcal{I})), Q_i^n(0))$ are assumed to be mutually independent.

The diffusion-scaled queueing processes are defined by

$$\widehat{Q}_i^n(t) = n^{-1/2}Q_i^n(t), \qquad i \in \mathcal{I}. \tag{11}$$

The diffusion scale deviation of the headcount process from the quantity reflected by the fluid model, is given by

$$\widehat{X}_i^n(t) = n^{-1/2}(X_i^n(t) - \rho_i N^n), \qquad i \in \mathcal{I}. \tag{12}$$

We also define the diffusion scale idleness process $\widehat{I}^n = n^{-1/2}I^n$. For simplicity, we will impose the following assumption on the initial value of the above process, namely

$$\widehat{X}^n(0) \Rightarrow 0, \qquad \text{as } n \to \infty. \tag{13}$$

Note that this is an assumption on the initial condition alluded to above, because $X_i^n(0)$ is the sum of $Q_i^n(0)$ and $B_i^n(0)$.

2.2 The cost function and an asymptotic lower bound

We are given a continuous function $C : \mathbb{R}^{\mathbf{I}} \to \mathbb{R}_+$, non-decreasing with respect to the usual partial order ($x \leq y$ iff $x_i \leq y_i$ for all i). It will serve as a 'running cost'. That is, we will be interested in the random variables

$$\int_0^u C(\widehat{Q}^n(t))dt \quad \text{and} \quad \int_0^u C(\widehat{X}^n(t))dt,$$

where $u > 0$, and attempt to find control policies that asymptotically minimize them. The so-called *workload process* plays an important role in solving this asymptotic control problem. By *workload* one means the time it takes a single server to complete the service of all customers present in the system. Note that the conditional mean of the workload at time t, given $X^n(t)$, is equal to

$$W^n(t) = \sum_{i \in \mathcal{I}} \frac{X_i^n(t)}{\mu_i} = \theta \cdot X^n(t),$$

where we denote $\theta = (\theta_i)_{i \in \mathcal{I}}$, $\theta_i = 1/\mu_i$. W^n is therefore called the workload process. The process

$$\widehat{W}^n(t) = \theta \cdot \widehat{X}^n(t) \tag{14}$$

represents the diffusion scale deviations of W^n from the nominal value. To present an asymptotic lower bound on the cost we need the following notation. Denote

$$C^*(w) = \inf\{C(q) : \, q \in \mathbb{R}_+^{\mathbf{I}}, \, \theta \cdot q = w\}, \qquad w \geq 0. \tag{15}$$

Define the *Skorohod map* $\Gamma : D([0,\infty) : \mathbb{R}) \to D([0,\infty), \mathbb{R}_+)$ by

$$\Gamma[\zeta](t) = \zeta(t) + \sup_{s \leq t}(-\zeta(s))^+, \qquad t \geq 0, \tag{16}$$

and denote by W the reflected Brownian motion $\Gamma[Z_\theta]$. Here, Z_θ is a Brownian motion starting from zero, with drift $\theta \cdot \widehat{\lambda}$ and diffusion coefficient $(\sum_i \theta_i^2(\lambda_i C_{i,IA}^2 + \mu_i \rho_i))^{1/2}$.

Theorem 2.1. *Let an arbitrary sequence of policies be given and let $\widehat{Q}^n$, $\widehat{X}^n$ and $\widehat{W}^n$ denote the corresponding processes from* (11) *and* (14)*. Fix $u > 0$. Then there exists a sequence of processes, $\{L^n\}$, defined on $[0,u]$, taking values in $\mathbb{R}$, and converging in law to W, such that the following holds.*

1. *For each n, L^n is a.s. dominated by $\widehat{W}^n$. As a result, any weak limit point of the sequence $\int_0^u C(\widehat{X}^n(t))dt$ stochastically dominates the r.v. $\int_0^u C^*(W(t))dt$.*
2. *For each interval $[a,b] \subset [0,u]$, $[\int_a^b \theta \cdot \widehat{Q}^n(s)ds - \int_a^b L^n(s)ds]^- \to 0$ in probability. As a result, provided that C^* is convex, any weak limit point of the sequence $\int_0^u C(\widehat{Q}^n(t))dt$ stochastically dominates the r.v. $\int_0^u C^*(W(t))dt$.*

2.3 Asymptotic optimality

Preemptive policies. We will be interested in policies that are preemptive and of feedback form, by which we mean that, at every time t, the value of $Q^n(t)$ is determined solely by n and the current 'state' of the system, $X^n(t)$. (While it is convenient to refer to X^n as state, note that this is an abuse of usual terminology because this process is not a true state descriptor of a controlled Markov process, except in the case when the arrivals are Poisson processes). Note that once $Q^n(t)$ is selected, $B^n(t)$ is determined via $B^n = X^n - Q^n$. Given n and a vector $X^n \in \mathbb{Z}^{\mathbf{I}}$, a vector $Q^n \in \mathbb{Z}_+^{\mathbf{I}}$ is said to be *feasible for the state X^n (for the nth system)*, if it satisfies the following conditions:

$$\begin{aligned} &(a) \qquad 0 \le Q_i^n \le X_i^n \text{ for all } i, \\ &(b) \qquad 1 \cdot X^n - 1 \cdot Q^n \le N^n. \end{aligned} \tag{17}$$

These relations express the facts that the ith queue length cannot exceed the number of class-i customers, and that the total number of customers in service cannot exceed the number of servers. Under a preemptive policy, the queue length vector Q^n can be selected among all feasible vectors for X^n. A policy is said to be *work conserving* if, for every t, $I^n(t) > 0$ implies $1 \cdot Q^n(t) = 0$. Equivalently, the following relation holds at all times:

$$(1 \cdot X^n - N^n)^+ = 1 \cdot Q^n. \tag{18}$$

Note that (17b) is automatically satisfied when relation (18) holds. Expressed in terms of the diffusion-scale processes, (17a) is equivalent to

$$0 \le \widehat{Q}_i^n \le \widehat{X}_i^n + \rho_i, \qquad i \in \mathcal{I}, \tag{19}$$

while (18) can be written as

$$(1 \cdot \widehat{X}^n)^+ = 1 \cdot \widehat{Q}^n \quad [\text{equivalently, } (1 \cdot \widehat{X}^n)^- = \widehat{I}^n]. \tag{20}$$

Construction of a policy. The main goal of this paper is to construct a sequence of policies that asymptotically attains the lower bound identified in Theorem 2.1. We keep the continuity and monotonicity assumptions on the cost function C, and add the following.

Assumption 2.1. (Existence of a continuous minimizing curve) *The function C and the corresponding function C^* of* (15) *satisfy the following. There exists a continuous function $f : \mathbb{R}_+ \to \mathbb{R}_+^{\mathbf{I}}$ such that*

$$\theta \cdot f(w) = w \quad \text{and} \quad C^*(w) = C(f(w)), \qquad w \in \mathbb{R}_+. \tag{21}$$

Note that $f(0) = 0$. For notational purposes it will be convenient to extend f to $\mathbb{R}$ by letting $f = 0$ on $(-\infty, 0)$. We refer to the curve $w \mapsto f(w)$, $w \in \mathbb{R}$ as the *minimizing curve*.

Remark 2.1. *Two important families of functions C satisfying Assumption 2.1 are as follows. 1. Continuous, homogeneous of degree $\alpha > 0$ (linear being a special case): $C(ax) = a^\alpha C(x)$, $a > 0$. If $x^* \in \arg\min\{C(x) : \theta \cdot x = 1\}$ then it is easy to check $f(w) = wx^*$ is a minimizing curve.*

2. Strictly convex. In this case, $\min\{C(x) : x \in \mathbb{R}_+^{\mathbf{I}}, \theta \cdot x = w\}$*, as the minimum of a strictly convex function over a compact convex set, is attained at a unique point,* $f(w)$*. Let us show that* f *is continuous. Arguing by contradiction, assume there exist* $w \geq 0$ *and a sequence* $w_n \to w$ *such that* $x_n = f(w_n) \to \tilde{x} \neq x = f(w)$*. Consider two cases. Case 1:* $w > 0$*. Set* $\hat{x}_n = xw_n/w$*. Then* $\hat{x}_n$ *satisfies* $\hat{x}_n \cdot \theta = w_n$ *hence* $C(x_n) \leq C(\hat{x}_n)$*. By continuity of* C*,* $C(\tilde{x}) \leq C(x)$*, a contradiction. Case 2:* $w = 0$*. In this case it is easy to see that both* x *and* $\tilde{x}$ *must be zero, a contradiction. This shows* f *is a continuous minimizing curve.*

The main step toward the asymptotic attainability is to establish this result (in Theorem 2.2) under a more restrictive condition, that is later dropped (in Theorem 2.3).

Condition A. *There exists* $r > 0$ *such that (a)* $f_{\mathbf{I}}(w) > r$ *for all* $w > r$*; and (b)* $f_i(w) = 0$ *for all* $i < \mathbf{I}$ *and* $w \in [0, r]$*.*

The asymptotic attainability under Assumption 2.1 and Condition A will be established by constructing a sequence of preemptive, work conserving policies of feedback form, under which

$$Q^n(t) = g^n(X^n(t)), \qquad t \geq 0,\, n \in \mathbb{N},$$

for some functions g^n mapping $\mathbb{Z}_+^{\mathbf{I}}$ into itself. The structure of the policies is motivated by the attempt to keep the relations

$$\widehat{X}^n(t) \approx f(\widehat{W}^n(t)), \qquad \widehat{Q}^n(t) \approx f(\widehat{W}^n(t))$$

hold for large values of n, so that the state and queue-length processes evolve close to the minimizing curve.

For a precise definition of the policy we have to specify how $Q^n(t)$ is determined from $X^n(t)$. Let $X^n = X^n(t)$ be given, and recall that $\widehat{X}^n = n^{-1/2}(X^n - \rho N^n)$.

1. First, define a *candidate queue-length vector*, $Q^{n,*}$ (the actual queue-length will typically be equal to the candidate queue-length, but not always; this is stated precisely in item 3). To this end, consider two cases.

   (a) $1 \cdot X^n \leq N^n$. In this case set $Q^{n,*} = 0$.

   (b) $1 \cdot X^n > N^n$. Let $[x]$ denote the integer part of a real number x, and for $x \geq 0$, define $\text{round}_n(x) := n^{-1/2}[n^{1/2}x]$. Set

   $$\widehat{Q}_i^{n,*} = \text{round}_n\, f_i(\widehat{W}^n) = \text{round}_n\, f_i(\theta \cdot \widehat{X}^n), \qquad \text{for all } i < \mathbf{I}. \tag{22}$$

   Note that the rounding operation assures that the quantities $Q_i^{n,*}$ take integer values. Next, for the last component, let

   $$\widehat{Q}_{\mathbf{I}}^{n,*} = 1 \cdot \widehat{X}^n - \sum_{i<\mathbf{I}} \widehat{Q}_i^{n,*}. \tag{23}$$

   Finally, let $Q^{n,*} = n^{1/2}\widehat{Q}^{n,*}$.

2. Next we need to fix some auxiliary work conserving policy, that will play a secondary role. For concreteness, it can be the policy that assigns absolute preemptive priority according to the order of the indices. That is, let $Q^{n,0} = X^n - B^n$, where

$$B_1^n = X_1^n \wedge N^n, \quad B_2^n = X_2^n \wedge (N^n - B_1^n), \ldots, \quad B_{\mathbf{I}}^n = X_{\mathbf{I}}^n \wedge \Big(N^n - \sum_{i<\mathbf{I}} B_i^n\Big).$$

3. Now determine Q^n as follows

$$Q^n = \begin{cases} Q^{n,*}, & \text{if } Q^{n,*} \text{ is feasible for } X^n, \\ Q^{n,0}, & \text{otherwise.} \end{cases}$$

Note that, by definition, Q^n is always feasible for X^n. Given n and f, we refer to the above policy as the (n, f) policy.

Theorem 2.2. *Let Assumption 2.1 and Condition A hold. For each n, let $\widehat{Q}^n$ and $\widehat{X}^n$ denote the processes associated with the policy (n, f) defined above. Then $(\widehat{Q}^n, \widehat{X}^n)$ converge in law, u.o.c., to $(f(W), f(W))$. As a result,*

$$\int_0^u C(\widehat{Q}^n(t))dt \Rightarrow \int_0^u C^*(W(t))dt \quad \textit{and} \quad \int_0^u C(\widehat{X}^n(t))dt \Rightarrow \int_0^u C^*(W(t))dt.$$

We now drop Condition A by invoking approximations. Recall that, under Assumption 2.1, the function f is continuous, vanishes at the origin, and satisfies $f_{\mathbf{I}} \geq 0$. Thus, it is easy to see that a sequence $\{f^k\}$ exists, where, for each k, $f^k : \mathbb{R}_+ \to \mathbb{R}_+^{\mathbf{I}}$ satisfies

1. Equation (21),
2. Condition A with $r = 1/k$,
3. $\sup_{\mathbb{R}_+} \|f^k - f\| \leq c_0/k$, for some c_0 that does not depend on k.

The asymptotic attainability under Assumption 2.1 alone is established by applying the policies (n, f^k), with $k = k(n)$.

Theorem 2.3. *Let Assumption 2.1 hold. Then the conclusions of Theorem 2.2 hold for the sequence of policies $(n, f^{k(n)})$, for a suitable choice of $\{k(n), n \in \mathbb{N}'\}$ (i.e., growing sufficiently slowly).*

3 Proofs

3.1 Proof of Theorem 2.2

In this subsection we prove Theorem 2.2. Throughout this subsection, Assumption 2.1 and Condition A are in force, and the stochastic processes Q^n, X^n, etc., correspond to the sequence of

policies (n, f). We begin by introducing rescaled versions of the processes involved, and specifying relations that they satisfy. Let

$$\bar{T}_i^n(t) = n^{-1}T_i^n(t) \equiv n^{-1/2}\mu_i \int_0^t B_i(s)ds, \tag{24}$$

$$\widehat{A}_i^n(t) = n^{-1/2}(A_i^n(t) - \lambda_i^n t), \tag{25}$$

$$\widehat{S}_i^n(t) = n^{-1/2}(S_i(nt) - nt), \tag{26}$$

$$\widetilde{B}_i^n = B_i^n - \rho_i N^n, \qquad \widehat{B}_i^n = n^{-1/2}\widetilde{B}_i^n, \tag{27}$$

$$V_i^n = n^{-1/2}(D_i^n - T_i^n) \equiv \widehat{S}_i^n \circ \bar{T}_i^n, \tag{28}$$

and

$$Z_i^n(t) = \widehat{\lambda}_i^n t + \widehat{A}_i^n(t) - V_i^n(t). \tag{29}$$

The second and third terms in the above display represent diffusion scale deviations of the arrival and, respectively, departure processes. As we will argue below, they asymptotically behave as Brownian motions (under some assumptions). This will be the basis of the proof of convergence of the diffusion-scale workload process to a reflected Brownian motion. Since $\sum_i \rho_i = 1$, we have by (7)

$$I^n + \sum_i \widetilde{B}_i^n = 0. \tag{30}$$

By (6) we have

$$\widehat{X}_i^n = \widehat{Q}_i^n + \widehat{B}_i^n. \tag{31}$$

Next, by (8), (9), (10), (12),

$$\begin{aligned}\widehat{X}_i^n(t) &= \widehat{X}_i^n(0) + n^{-1/2}A_i^n - n^{-1/2}D_i^n(t) \\ &= \widehat{X}_i^n(0) + Z_i^n(t) - n^{-1/2}\mu_i^n \int_0^t \widetilde{B}_i^n(s)ds.\end{aligned}$$

Thus

$$\widehat{X}_i^n(t) = \widehat{X}_i^n(0) + Z_i^n(t) - \mu_i \int_0^t \widetilde{B}_i^n(s)ds. \tag{32}$$

Using (31), we obtain

$$\widehat{Q}_i^n(t) = \widehat{X}_i^n(0) + Z_i^n(t) - \mu_i \int_0^t \widetilde{B}_i^n(s)ds - \widehat{B}_i^n(t). \tag{33}$$

Identities (32) and (33) will be particularly important in the sequel.

Let $\varepsilon_0 = \frac{\min_i \rho_i}{2\mathbf{I}} \wedge \frac{r}{2}$ and denote

$$\mathcal{E}_\varepsilon = \{x \in \mathbb{R}^{\mathbf{I}} : \|x - f(\theta \cdot x)\| \le \varepsilon\}, \qquad \mathcal{E} = \mathcal{E}_{\varepsilon_0}, \qquad \mathcal{E}' = \mathcal{E}_{\varepsilon_0/2}.$$

We have the following (the proofs of the lemmas stated in this subsection appear in the next subsection).

Lemma 3.1. *For each n, the policy (n, f) is work-conserving. Moreover, if $n^{-1/2} < \frac{\varepsilon_0}{2\mathbf{I}}$ and $\widehat{X}^n \in \mathcal{E}$ then $Q^{n,*}$ is feasible for X^n (hence so is Q^n).*

Let

$$\tau_n = \inf\{t : \widehat{X}^n(t) \notin \mathcal{E}'\} \wedge u.$$

For $v > 0$, denote by $\mathbb{D}_v$ the space of RCLL functions from $[0, v]$ to $\mathbb{R}$. Endow $\mathbb{D}_v$ with the usual (J_1) Skorohod topology. A sequence of processes with sample paths in $\mathbb{D}_v$ is said to be *C-tight* if it is tight, and every subsequential weak limit has continuous sample paths a.s. For $x \in \mathbb{D}_v$ and $\delta > 0$, denote

$$\bar{w}_v(x, \delta) = \sup_{s,t\in[0,v];|s-t|\leq\delta} |x(s) - x(t)|.$$

A useful characterization of C-tightness is as follows. A sequence R_n is C-tight if and only if one has

$$\begin{aligned} &(i) \quad |R_n|_v^* \text{ is a tight sequence of r.v.s, and} \\ &(ii) \quad \text{for every positive } \varepsilon \text{ and } \varepsilon' \text{ there exist } n_0 \text{ and } \delta \text{ such that} \\ &\qquad n > n_0 \quad \text{implies} \quad \mathbb{P}(\bar{w}_v(R_n, \delta) > \varepsilon) < \varepsilon' \end{aligned} \tag{34}$$

(see e.g., [7, Proposition VI.3.26]).

Toward proving the theorem, let us first argue that, for each i, the sequence of processes Z_i^n, $n \in \mathbb{N}'$ is C-tight. As is well known, for each i, the scaled processes A_i^n and S_i^n converge in distribution, uniformly on compacts, to a zero mean Brownian motion (BM) with diffusion coefficient $\lambda_i^{1/2} C_{i,IA}$, and 1, respectively [3, Section 17]. Because of the independence assumption of the primitive processes, the collection $(\widehat{A}_i^n, i \in \mathcal{I}, \widehat{S}_i^n, i \in \mathcal{I})$ jointly converges in law, as $n \to \infty$, to a collection $((Z_{i,A})_{i\in\mathcal{I}}, Z_{i,S})$ of $2\mathbf{I}$ independent BMs, with the laws specified above.

By (24), clearly $0 \leq \frac{d}{dt}\bar{T}_i^n(t) \leq \mu_i$, thus $\bar{T}_i^n$ are uniformly Lipschitz. It is easy to see, using the C-tightness characterization above, using (28), that V_i^n are C-tight. In turn, using (29), so are Z_i^n.

Based on this, one can show the following.

Lemma 3.2. *For each i, $(\widehat{X}_i^n)^- \Rightarrow 0$.*

Lemma 3.3. *The exists a continuous function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$, vanishing at zero, such that the following holds. Whenever $\widehat{X}^n \in \mathcal{E}$,*

$$|\widehat{Q}_i^n - f_i(\theta \cdot \widehat{X}^n)| \leq \gamma\Big(\max_{j\in\mathcal{I}}(\widehat{X}_j^n)^-\Big) + n^{-1/2}, \qquad i < \mathbf{I}.$$

Next, by (31) and (32), we have

$$\widehat{B}_i^n(t) = F_i^n(t) - \mu_i n^{1/2} \int_0^t \widehat{B}_i^n(s)ds, \tag{35}$$

where

$$F_i^n(t) = \widehat{X}_i^n(0) + Z_i^n(t) - \widehat{Q}_i^n(t). \tag{36}$$

A crucial element of the proof is Lemma 3.4 below, which is based on equation (35). For each i and n, consider $\widehat{B}_i^n$ as the solution to this equation with data F_i^n. If F_i^n were, say, C^1 functions with bounded first derivative, one could express $\frac{d}{dt}\widehat{B}_i^n$ as the convolution of $\frac{d}{dt}F_i^n$ with $e^{-\mu_i n^{1/2} t}$, and immediately obtain a uniform estimate on $\widehat{B}_i^n$, that vanishes as $n \to \infty$. The purpose of the lemma is to show that merely C-tightness of F_i^n is sufficient for a similar estimate to hold.

Lemma 3.4. *Fix i. Given $u > 0$ and $\delta \in (0, u)$, for all n,*

$$|\widehat{B}_i^n|_u^* \le 2|F_i^n|_u^* e^{-\mu_i \sqrt{n}\delta} + \bar{w}_u(F_i^n, \delta).$$

Thus if the sequence F_i^n is C-tight then $\widehat{B}_i^n$ converges to zero u.o.c. in probability.

To present the proof we need one more result.

Lemma 3.5. *The sequence $\widehat{W}^n$ is C-tight. Moreover, if (Z, W) is any subsequential limit of the pair $(\theta \cdot Z^n, \widehat{W}^n)$ then $W = \Gamma(Z)$.*

The main obstacle in applying Lemma 3.4 is that we do not a priori know the F_i^n's are C-tight. We have to argue step by step.

To this end, note first, using the fact that all jumps of $\widehat{X}^n$ are of order $n^{-1/2}$ a.s. and the definition of τ_n, that, for all sufficiently large n, $\widehat{X}^n(t) \in \mathcal{E}$ for all $t \in [0, \tau_n]$, a.s. on the event $\widehat{X}^n(0) \in \mathcal{E}'$.

By Lemmas 3.2 and 3.3,

$$\max_{i<\mathbf{I}} |\widehat{Q}_i^n - f_i(\theta \cdot \widehat{X}^n)|_{\tau_n}^* \to 0, \text{ in probability.} \tag{37}$$

Combined with Lemma 3.5, the fact $\widehat{W}^n = \theta \cdot \widehat{X}^n$, and the continuity of the function f, this shows that for $i < \mathbf{I}$, $\widehat{Q}_i^n(\cdot \wedge \tau_n)$ are C-tight. As a result, using Lemma 3.4, for all $i < \mathbf{I}$, $\widehat{B}_i^n(\cdot \wedge \tau_n) \to 0$ uniformly on $[0, u]$, in probability.

Next, we argue that for all $i \le \mathbf{I}$, $\widehat{X}_i^n(\cdot \wedge \tau_n)$ are C-tight. For $i < \mathbf{I}$ this is immediate from (31) using the tightness, shown above, of $\widehat{Q}_i^n$ and $\widehat{B}_i^n$, stopped at τ_n. As for $i = \mathbf{I}$, write $\widehat{X}_{\mathbf{I}}^n$ as a linear combination of $\theta \cdot \widehat{X}^n = \widehat{W}^n$ and $\widehat{X}_j^n$, $j < \mathbf{I}$ (where we use the fact that $\theta_j > 0$ for all j). Now use Lemma 3.5 and the C-tightness of $\widehat{X}_j^n$, $j < \mathbf{I}$, stopped at τ_n.

As verified in both cases (a) and (b) of item 1 in the definition of the policy, for $t \le \tau_n$,

$$\widehat{Q}_{\mathbf{I}}^n = (1 \cdot \widehat{X}^n)^+ - \sum_{i<\mathbf{I}} \widehat{Q}_i^n.$$

As a result, $\widehat{Q}_{\mathbf{I}}^n(\cdot \wedge \tau_n)$ are C-tight. Using again Lemma 3.4, now for $i = \mathbf{I}$ shows that $\widehat{B}_{\mathbf{I}}^n(\cdot \wedge \tau_n) \to 0$. Summarizing what we have shown thus far, we have, for all $i \le \mathbf{I}$,

$$\widehat{Q}_i^n(\cdot \wedge \tau_n) \text{ and } \widehat{X}_i^n(\cdot \wedge \tau_n) \text{ are } C\text{-tight, and } \|\widehat{B}^n\|_{\tau_n}^* = \|\widehat{X}^n - \widehat{Q}^n\|_{\tau_n}^* \to 0 \text{ in probability.} \tag{38}$$

Now this statement can be lifted to hold for the unstopped processes, provided

$$\mathbb{P}(\tau_n < u) \to 0. \tag{39}$$

To show (39), note that (38) and (37) together show that for $i < \mathbf{I}$, $|\widehat{X}_i^n - f_i(\theta \cdot \widehat{X}^n)|^*_{\tau_n} \to 0$. As for $i = \mathbf{I}$, using $\theta \cdot f(w) = w$ for $w \geq 0$, we have on $[0, \tau_n]$,

$$f_{\mathbf{I}}(\widehat{W}^n) - \widehat{X}_{\mathbf{I}}^n = \theta_{\mathbf{I}}^{-1} \sum_{i<\mathbf{I}} \theta_i (\widehat{X}_i^n - f_i(\widehat{W}^n)),$$

provided $\widehat{W}^n \geq 0$. When $\widehat{W}^n < 0$, we have

$$f_{\mathbf{I}}(\widehat{W}^n) - \widehat{X}_{\mathbf{I}}^n = -\widehat{X}_{\mathbf{I}}^n.$$

Denote $\xi^n = \max_i \sup_{t\in[0,u]} (\widehat{X}_i^n(t))^-$, $\theta_{\min} = \min_i \theta_i$, $\theta_{\max} = \max_i \theta_i$. Since $\widehat{W}^n < 0$, we have $\sum_i \theta_i (\widehat{X}_i^n)^+ < \sum_i \theta_i (\widehat{X}_i^n)^-$, and it follows that $\theta_{\min} \sum_i (\widehat{X}_i^n)^+ \leq I\theta_{\max}\xi^n$. As a result, for every i, $|\widehat{X}_i^n| \leq c_2 \xi^n$, where $c_2 = I\theta_{\max}/\theta_{\min}$. Thus

$$|f_{\mathbf{I}}(\widehat{W}^n) - \widehat{X}_{\mathbf{I}}^n| \leq c_2 \xi^n.$$

Combining the two cases,

$$|f_{\mathbf{I}}(\widehat{W}^n) - \widehat{X}_{\mathbf{I}}^n|^*_{\tau_n} \leq c_1 \sum_{i<\mathbf{I}} |\widehat{X}_i^n - f_i(\widehat{W}^n)|^*_{\tau_n} + \xi^n.$$

Thus, based on the result for $i < \mathbf{I}$, and on Lemma 3.1, the r.h.s. of the above display converges to zero in probability. As a result, $\|\widehat{X}^n - f(\theta \cdot \widehat{X}^n)\|^*_{\tau_n} \to 0$ in probability. By the definitions of τ_n and $\mathcal{E}'$, this shows (39). We thus have, for all $i \leq \mathbf{I}$,

$$\widehat{Q}_i^n \text{ and } \widehat{X}_i^n \text{ are } C\text{-tight, and } \|\widehat{B}^n\|^*_u = \|\widehat{X}^n - \widehat{Q}^n\|^*_u \to 0 \text{ in probability.} \tag{40}$$

Similarly, (37) holds with τ_n replaced by u.

By (24) and (27),

$$\bar{T}_i^n(t) = \mu_i \int_0^t \widehat{B}_i^n(s)ds + \mu_i \rho_i t. \tag{41}$$

Denoting $T_i(t) = \mu_i \rho_i t$, we have from (40) that $\bar{T}_i^n \Rightarrow T_i$. By a lemma regarding random change of time [3, p. 151], it follows from (28) that $((\hat{A}_i^n)_{i\in\mathcal{I}}, (V_i)_{i\in\mathcal{I}}) \Rightarrow ((Z_{i,A})_{i\in\mathcal{I}}, (Z_{i,S} \circ T_i)_{i\in\mathcal{I}})$. In view of (29), the processes $\theta \cdot Z^n$ converge to the process we denote Z_θ, that is a BM starting from zero, with drift $\theta \cdot \widehat{\lambda}$ and diffusion coefficient $(\sum_i \theta_i^2 (\lambda_i C_{i,IA}^2 + \mu_i \rho_i))^{1/2}$.

A use of Lemma 3.5 along with (40) now shows that $(\widehat{W}^n, \theta \cdot Z^n)$, as well as $(\theta \cdot \widehat{Q}^n, \theta \cdot Z^n)$, converge to $(\Gamma(Z_\theta), Z_\theta)$.

Let X be any subsequential limit of $\widehat{X}^n$. Then, along the subsequence, it is also a limit of $\widehat{Q}^n$, due to (40). Moreover, $W := \theta \cdot X$ is the limit of $\widehat{W}^n$. Hence by (37) and continuity of f,

$$X_i = f_i(\theta \cdot X) = f_i(W), \qquad i < \mathbf{I}.$$

A similar relation is true for $i = \mathbf{I}$, because, using the above display and arguing again by the property $\theta \cdot f(w) = w$, $w \geq 0$,

$$f_{\mathbf{I}}(W) = \theta_{\mathbf{I}}^{-1}\Big(\theta \cdot f(W) - \sum_{i<\mathbf{I}} \theta_i f_i(W)\Big) = \theta_{\mathbf{I}}^{-1}\Big(W - \sum_{i<\mathbf{I}} \theta_i X_i\Big) = X_{\mathbf{I}}.$$

Thus,

$$X = f(\theta \cdot X) = f(W) = f(\Gamma(Z)),$$

where we used Lemma 3.5. We have shown that $(\widehat{X}^n, \widehat{Q}^n)$ converge in law to $(f(\Gamma(Z)), f(\Gamma(Z)))$, uniformly on $[0, u]$. This completes the proof of the first assertion in Theorem 2.2. The second assertion follows by continuity of C and the definition of f, by which $C^* = C \circ f$. □

3.2 Proof of Lemmas

Proof of Lemma 3.1. To see that the work conservation condition (18) (equivalently (20)) holds for $Q^{n,*}$, note by property 1(a) in its definition that (18) holds when $1 \cdot X^n \leq N^n$, and by (23), that (20) holds when $1 \cdot X^n > N^n$. For $Q^{n,0}$ the property is verified directly. Consequently, Q^n satisfies the same condition.

Next we prove the assertion regarding feasibility. Let n satisfy $2\mathbf{I}n^{-1/2} < \varepsilon_0$ and let $\widehat{X}^n \in \mathcal{E}$. Since the work conservation condition holds, by the discussion on preemptive policies it suffices to verify that $\widehat{Q}^{n,*}$ and $\widehat{X}^n$ satisfy (19). Note that the rounding operation performed in (22) has the property $|x - \mathrm{round}_n(x)| \leq n^{-1/2}$ for $x \geq 0$. Hence by (22), for $i < \mathbf{I}$,

$$|\widehat{Q}_i^{n,*} - \widehat{X}_i^n| \leq |f_i(\theta \cdot \widehat{X}^n) - \widehat{X}_i^n| + n^{-1/2} \leq \varepsilon_0 + n^{-1/2} \leq \rho_i.$$

Since we also have that $f_i(\cdot) \geq 0$, we see that (19) holds for $i < \mathbf{I}$. For $i = \mathbf{I}$, by (23),

$$|\widehat{Q}_{\mathbf{I}}^{n,*} - \widehat{X}_{\mathbf{I}}^n| \leq \sum_{i<\mathbf{I}} |\widehat{Q}_i^{n,*} - \widehat{X}_i^n| \leq \mathbf{I}(\varepsilon_0 + n^{-1/2}) \leq \rho_{\mathbf{I}},$$

and the second inequality of (19) holds for $\mathbf{I}$ as well.

It remains to prove the non-negativity of $\widehat{Q}_{\mathbf{I}}^{n,*}$. This is where Condition A is used. In case 1(a) of the definition of the policy, the non-negativity is immediate (because $Q^{n,*}$ is set to zero). In case 1(b) we have $1 \cdot \widehat{X}^n > 0$. Consider two subcases. First, if $\widehat{W}^n \leq r$, then by Condition A(b), $Q_i^{n,*} = 0$ for $i < \mathbf{I}$, hence by (23) $\widehat{Q}_{\mathbf{I}}^{n,*} = 1 \cdot \widehat{X}^n > 0$. Next, if $\widehat{W}^n > r$, we have by Condition A(a) that $f_{\mathbf{I}}(\widehat{W}^n) > r$, hence by (23),

$$\begin{aligned}\widehat{Q}^{n,*} &= f_{\mathbf{I}}(\widehat{W}^n) + [\widehat{X}_{\mathbf{I}}^n - f_{\mathbf{I}}(\widehat{W}^n)] + \sum_{i<\mathbf{I}} [\widehat{X}_i^n - \mathrm{round}_n \circ f_i(\widehat{W}^n)] \\ &\geq r - \|\widehat{X}^n - f(\widehat{W}^n)\| - \mathbf{I}n^{-1/2} \geq r - \frac{r}{2} - \frac{r}{4} > 0.\end{aligned}$$

This proves that $Q^{n,*}$ is feasible for X^n. □

Proof of Lemma 3.2. Fix i. Fix an interval $[0,u]$. Since by assumption $\widehat{X}^n(0) \to 0$ in probability, it suffices to prove that, for every $a > 0$, $\mathbb{P}(\widehat{X}_i^n(0) > -a, \inf_{[0,u]} \widehat{X}_i^n \leq -3a) \to 0$. Using the fact that, a.s., all the jumps of $\widehat{X}_i^n$ are of size $n^{-1/2}$, it suffices to prove that, for every $a > 0$,

$$\mathbb{P}(\text{there exist } 0 \leq \sigma < \theta \leq u \text{ s.t. } \widehat{X}_i^n(\sigma) \geq -2a;\ \widehat{X}_i^n \leq -a \text{ on } [\sigma,\theta]; \text{ and } \widehat{X}_i^n(\theta) \leq -3a) \to 0. \quad (42)$$

By (31), $\widehat{X}_i^n \geq \widehat{B}_i^n$. Hence on the event indicated in (42), for $s \in [\sigma,\theta]$ one has $\widetilde{B}_i^n(s) = \sqrt{n}\widehat{B}_i^n(s) \leq \sqrt{n}\widehat{X}_i^n(s) \leq -\sqrt{n}a$. As a result, on this event, by (32),

$$-a \geq \widehat{X}_i^n(\theta) - \widehat{X}_i^n(\sigma) \geq Z_i^n(\theta) - Z_i^n(\sigma) + \sqrt{n}\mu_i a(\theta - \sigma).$$

Fix $b > 0$. Note that if $\theta - \sigma \geq b$ then a necessary condition for the above display to hold is $\bar{w}_u(Z_i^n, b) \geq a$. Hence the probability on the l.h.s. of (42) is bounded by

$$\mathbb{P}(\bar{w}_u(Z_i^n, b) \geq a) + \mathbb{P}(2|Z_i^n|_u^* \geq \sqrt{n}\mu_i ab).$$

Since Z_i^n are C-tight, the expression in the above display vanishes upon taking $n \to \infty$ and then $b \to 0$. Indeed, this follows for the first and second terms in the above display, by parts (ii) and (i), respectively, of the characterization (34). The result follows. □

Proof of Lemma 3.3. Let $g : \mathbb{R}^{\mathbf{I}} \to \mathbb{R}^{\mathbf{I}-1}$ be defined as

$$g(x) = \begin{cases} \Big(f_1(\theta \cdot x), \ldots, f_{\mathbf{I}-1}(\theta \cdot x)\Big), & 1 \cdot x > 0, \\ 0, & 1 \cdot x \leq 0. \end{cases}$$

Let $\widehat{X}^n \in \mathcal{E}$. By Lemma 3.1, $Q^n = Q^{n,*}$. By (22), for $i < \mathbf{I}$, $|\widehat{Q}_i^n - g_i(\widehat{X}^n)| \leq n^{-1/2}$. Denote $b = \max_j(\widehat{X}_j^n)^-$. Fix $i < \mathbf{I}$. Write x for $\widehat{X}^n$. Toward bounding $g_i(x) - f_i(\theta \cdot x)$, note that

$$|g_i(x) - f_i(\theta \cdot x)| = |f_i(\theta \cdot x)| 1_{\{1 \cdot x \leq 0\}}.$$

If $1 \cdot x = \sum x_j^+ - \sum x_j^- \leq 0$ and, for all j, $x_j^- \leq b$ then $\sum x_j^+ \leq \mathbf{I}b$ hence $\|x\| \leq 2\mathbf{I}b$. Thus

$$|g_i(x) - f_i(\theta \cdot x)| \leq \sup\{\|f(\theta \cdot x)\| : \|x\| \leq 2\mathbf{I}b\}.$$

Recall that f is assumed to be continuous. By its definition, it vanishes at zero. It follows that there exists a continuous function γ vanishing at the origin, such that $\gamma(b)$ in an upper bound on the r.h.s. of the above display. We have thus shown that

$$|\widehat{Q}_i^n - f_i(\theta \cdot \widehat{X}^n)| \leq n^{-1/2} + |g_i(\widehat{X}^n) - f_i(\theta \cdot \widehat{X}^n)| \leq n^{-1/2} + \gamma(\max_{j \in \mathcal{I}}(\widehat{X}_j^n)^-),$$

for $i < \mathbf{I}$. □

Proof of Lemma 3.4. The proof is based on the equation (35) satisfied by $\widehat{B}_i^n$. The solution X to the integral equation

$$X(t) = F(t) - \mu \int_0^t X(s)ds, \qquad t \geq 0, \quad (43)$$

is given by

$$X(t) = F(t) - \mu \int_0^t F(s)e^{-\mu(t-s)}ds, \quad (44)$$

as can be readily checked. For $\delta \in (0,t)$ one has $X = D + E$, where

$$D(t) = -\mu \int_0^{t-\delta} F(s) e^{-\mu(t-s)} ds, \qquad E(t) = F(t) - \mu \int_{t-\delta}^{t} F(s) e^{-\mu(t-s)} ds.$$

We have

$$|D|_t^* \le |F|_t^* e^{-\mu\delta}.$$

Moreover, denoting $b = \bar{w}_t(F, \delta)$,

$$E(t) \le F(t) - \mu \int_{t-\delta}^{t} (F(t) - b) e^{-\mu(t-s)} ds = F(t) e^{-\mu\delta} + b(1 - e^{-\mu\delta}),$$

and in conjunction with an analogous lower bound, we obtain $|E|_t^* \le |F|_t^* e^{-\mu\delta} + b$. Combining the bounds on D and E yields the result. □

Proof of Lemma 3.5. Denote $Z_0^n(t) = \theta \cdot \widehat{X}^n(0) + \theta \cdot Z^n(t)$. Let $\zeta^n(t) = \int_0^t I^n(s) ds$ and note that it is nondecreasing. By (30), $I^n = -1 \cdot \widetilde{B}^n$. Since $\widehat{W}^n = \theta \cdot \widehat{X}^n$ and $\theta_i = \mu_i^{-1}$, we have from (32) that

$$\widehat{W}^n = Z_0^n + \zeta^n.$$

Fix $a > 0$. Let $Y^n = (\widehat{W}^n - a)^+$. Then

$$Y^n(t) = (\widehat{W}^n - a)^- - a + Z_0^n + \zeta^n(t). \tag{45}$$

Let $b = \frac{1}{2} \frac{a}{1 \cdot \theta}$. by Lemma 3.2, the probability of the event $\Omega_n = \{\min_j \inf_{[0,u]} \widehat{X}_j^n \ge -b\}$ converges to 1. Let us argue that, on Ω_n, $Y^n(t) > 0$ implies $I^n(t) = 0$. This is based on work conservation. Indeed, $Y^n(t) > 0$ implies $\theta \cdot \widehat{X}^n(t) > a$. Hence $\max_j \widehat{X}^n(t) > c := \frac{a}{1 \cdot \theta}$, and

$$1 \cdot \widehat{X}^n(t) \ge \max_j \widehat{X}_j^n(t) - \mathbf{I}b \ge c - b > 0.$$

Since the policy is work conserving we have by (20) that $I^n = n^{1/2} \widehat{I}^n = n^{1/2} (1 \cdot \widehat{X}^n)^-$. This proves that the implication, alluded to above, holds on Ω_n. As a result, $\int 1_{\{Y^n > 0\}} d\zeta^n = 0$ holds on Ω_n. Coupled with (45) and the fact that $Y^n \ge 0$, this shows that, on Ω_n, the relation

$$Y^n = \Gamma((\widehat{W}^n - a)^- - a + Z_0^n)$$

holds. Since the map Γ satisfies the Lipschitz condition w.r.t. the sup norm, we have on Ω_n, for some constant L,

$$\begin{aligned} |\widehat{W}^n - \Gamma(\theta \cdot Z^n)| &\le |\widehat{W}^n - Y^n| + |Y^n - \Gamma(Z_0^n)| + |\Gamma(Z_0^n) - \Gamma(\theta \cdot Z^n)| \\ &\le (\widehat{W}^n - a)^- + a + L\{|(\widehat{W}^n - a)^-|_u^* + a\} + L|\theta \cdot \widehat{X}^n(0)|. \end{aligned}$$

Recall that $\theta \cdot Z^n$ are C-tight, and consider any subsequential limit Z. Using the bound in the above display, the continuity of Γ, the convergence $(\widehat{W}^n)^- \Rightarrow 0$, the assumption $\widehat{X}^n(0) \Rightarrow 0$, and the fact that $a > 0$ is arbitrary, we obtain $\widehat{W}^n \Rightarrow \Gamma(Z)$, along the subsequence. This shows C-tightness of $\widehat{W}^n$. The second claim of the lemma follows. □

3.3 Proof of Theorem 2.3

For each k, f^k satisfies equation (21) as well as Condition A (with $r = 1/k$). The proof of Theorem 2.2 presented above establishes, under the sequence of policies (n, f^k) with k *fixed*, the convergence of $(\widehat{X}^n, \widehat{Q}^n)$ to $(f^k(W), f^k(W))$. Since f^k converge to f uniformly as $k \to \infty$, choosing $k(n)$ to increase to infinity sufficiently slowly, gives rise to $(f(W), f(W))$ as the limit under $(n, f^{k(n)})$. The convergence of the cost to the minimal cost follows as in Theorem 2.2. □

3.4 Proof of Theorem 2.1

Consider the stochastic processes X^n, Q^n, etc., under an arbitrary, fixed sequence of policies. A review of the proof of Theorem 2.2 shows that the following elements do not rely on Assumption 2.1 or Condition A, nor do they depend on the particular properties of the policy. Thus, they continue to hold for the arbitrary policies we have fixed.

- C-tightness of Z^n,

- The convergence $(X_i^n)^- \Rightarrow 0$.

An event Ω^n is said to occur *with high probability* (w.h.p.) if $\mathbb{P}(\Omega^n) \to 1$ as $n \to \infty$. The symbol $o(1)$ will serve as generic notation for an n-dependent r.v. or stochastic process defined on $[0, u]$, converging to zero in probability (uniformly on $[0, u]$, in the case of a process).

Let

$$\sigma_n = \inf \left\{ t : \max_{i \in \mathcal{I}} \left| \int_0^t \widehat{B}_i^n(s) ds \right| \geq \varepsilon_n \right\} \wedge u,$$

where $\varepsilon_n > 0$ converge to zero, while $\varepsilon_n n^{1/2} \to \infty$. If $\sigma_n < u$ then by (32) there exists i for which

$$\widehat{X}_i^n(\sigma_n) = \widehat{X}_i^n(0) + Z_i^n(\sigma_n) \pm \mu_i \varepsilon_n n^{1/2}.$$

Since $\max_j (\widehat{X}_j^n)^- \to 0$ in probability, and Z_i^n are tight, the probability that $\widehat{X}_i^n(s) = \widehat{X}_i^n(0) + Z_i^n(s) - \mu_i \varepsilon_n n^{1/2}$ for some $s \in [0, u]$, some i, converges to zero as $n \to \infty$. Hence, w.h.p.,

$$\widehat{W}^n(t) 1_{\{t \in (\sigma_n, u]\}} \geq r_n := c \varepsilon_n n^{1/2}, \tag{46}$$

some constant $c > 0$, where $(u, u]$ is regarded as the empty set. Similarly to the proof of Lemma 3.5, using (30) and (32),

$$\widehat{W}^n = \theta \cdot X^n(0) + \theta \cdot Z^n + \int_0^{\cdot} I^n(s) ds = o(1) + \theta \cdot Z^n + \int_0^{\cdot} I^n(s) ds.$$

Since we also have $(\widehat{W}^n)^- = o(1)$,

$$(\widehat{W}^n)^+ = o(1) + \theta \cdot Z^n + \int_0^{\cdot} I^n(s) ds.$$

In view of the non-negativity of $(\widehat{W}^n)^+$ and the monotonicity of $\int I^n$ we can invoke the well-known minimality property of the Skorohod map [4], and obtain

$$(\widehat{W}^n)^+ \geq \Gamma[o(1) + \theta \cdot Z^n].$$

As a result,

$$\widehat{W}^n \geq o(1) + \Gamma[o(1) + \theta \cdot Z^n] = o(1) + \Gamma[\theta \cdot Z^n].$$

Combined with (46), this shows that w.h.p.,

$$\widehat{W}^n(t) \geq o(1) + \Gamma[\theta \cdot Z^n] 1_{\{t \in [0,\sigma_n]\}} + r_n 1_{\{t \in (\sigma_n, u]\}}.$$

Recall from the proof of Theorem 2.2 the notation $T_i(t) = \mu_i \rho_i t$ and the representation (41) for the process $\bar{T}_i^n$ and the fact that $\frac{d}{dt}\bar{T}_i^n$ are uniformly bounded. By the definition of σ_n, $a_n := |\bar{T}_i^n - T_i|^*_{\sigma_n} \to 0$ in probability. Hence the process $V_i^n = \widehat{S}_i \circ \bar{T}_i^n$ (28) satisfies $|\widehat{S}_i^n \circ \bar{T}_i^n - \widehat{S}_i^n \circ T_i|^*_{\sigma_n} \leq b_n := \bar{w}_{c_1 u}(\widehat{S}_i^n, a_n)$, for some constant c_1. Since $\widehat{S}_i^n$ are C-tight, $b_n \to 0$ in probability. Hence by (29), denoting

$$G_i^n(t) = \widehat{\lambda}^n t + \widehat{A}_i^n(t) - \widehat{S}_i^n \circ T_i(t)$$

and

$$L_n = \Gamma[\theta \cdot G^n],$$

we have w.h.p.,

$$\widehat{W}^n(t) \geq o(1) + L_n 1_{\{t \in [0,\sigma_n]\}} + r_n 1_{\{t \in (\sigma_n, u]\}}.$$

Note that L_n converges in law to the process $W = \Gamma[Z_\theta]$, thus $\mathbb{P}(\sup_{(\sigma_n,u]} L_n > r_n) \to 0$. In particular, w.h.p.,

$$\widehat{W}^n(t) \geq o(1) + L_n, \tag{47}$$

and the r.h.s. converges in law to W. This proves the first assertion of item 1 of the theorem. The second assertion, regarding $\int_0^u C(\widehat{X}^n(t))dt$, follows immediately by continuity of C and the definition of C^*.

We turn to the second part of the theorem. Let $[a, b] \subset [0, u]$ be given. By (31), $\theta \cdot \widehat{Q}^n$ and $\widehat{W}^n$ differ by $\theta \cdot \widehat{B}^n$. Hence by the definition of σ_n,

$$\int_a^{a \vee (\sigma_n \wedge b)} \theta \cdot \widehat{Q}^n(s)ds \geq \int_a^{a \vee (\sigma_n \wedge b)} \widehat{W}^n(s)ds - c\varepsilon_n.$$

Moreover, since $\widehat{B}^n$ are uniformly bounded, a lower bound of the form (46) holds for $\theta \cdot \widehat{Q}^n$ as well. Thus,

$$\int_a^b \theta \cdot \widehat{Q}^n(s)ds \geq \int_a^{a \vee (\sigma_n \wedge b)} \widehat{W}^n(s)ds - c\varepsilon_n + \int_{a \vee (\sigma_n \wedge b)}^b r_n ds \geq o(1) + \int_a^b L_n(s)ds,$$

where we used (47). This proves the first assertion of item 2.

To prove the second assertion, fix k, and write $\delta = u/k$, $\Delta_j = [(j-1)\delta, j\delta)$, $j = 1, 2, \ldots, k$. Recall that in this case we assume convexity of C^*. Thus, using Jensen's inequality,

$$\int_0^u C(\widehat{Q}^n(s))ds \geq \int_0^u C^*(\theta \cdot Q^n(s))ds \geq \sum_{j=1}^k C^*\Big(\frac{1}{\delta}\int_{\Delta_j} \theta \cdot \widehat{Q}^n(s)ds\Big)\delta$$
$$\geq \sum_{j=1}^k C^*\Big(\frac{1}{\delta}\int_{\Delta_j} L^n(s)ds\Big)\delta - o(1).$$

For $A > 0$, denote $\mathrm{mod}(A, \cdot)$ the modulus of continuity of $C^*|_{[-A,A]}$. Then

$$\int_0^u C(\widehat{Q}^n(s))ds \geq \int_0^u C^*(L^n(s))ds - o(1) - \mathrm{mod}(Y_n, y_n)u,$$

where $Y_n = |L^n|_u^*$ and $y_n = \bar{w}_u(L^n, \delta)$. Since L^n converge to the process W, any subsequential limit in distribution of the l.h.s. stochastically dominates $\int_0^u C^*(W(s))ds$, provided we have, for every $\varepsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}(\mathrm{mod}(Y_n, y_n) > \varepsilon) = 0.$$

However, this is clear by C-tightness of L^n. This completes the proof of the theorem. □

References

[1] R. Atar. A diffusion regime with non-degenerate slowdown. Preprint.

[2] R. Atar and I. Gurvich. Work in progress.

[3] P. Billingsley. *Convergence of Probability Measures.* Second edition. Wiley, New York, 1999.

[4] H. Chen and A. Mandelbaum. Leontief systems, RBVs and RBMs. in *Applied stochastic analysis (London, 1989),* 1–43, Stochastics Monogr., 5, Gordon and Breach, New York, 1991.

[5] I. Gurvich. *Design and control of the M/M/N queue with multi-class customers and many servers.* M.Sc. Thesis, Technion, July 2004

[6] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29 (1981), no. 3, 567–588.

[7] J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes.* Springer-Verlag, 1987.

[8] A. Mandelbaum. QED Q's. Notes from a lecture delivered at the Workshop on Heavy Traffic Analysis and Process Limits of Stochastic Networks, EURANDOM, September 2003

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

[10] J. A. van Mieghem. (1995). Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab.* 5 809833.

[11] W. Whitt. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* Vol. 51 No. 4 531–542 (2003)