The Annals of Applied Probability
2015, Vol. 25, No. 1, 406–428
DOI: 10.1214/13-AAP1001

NECESSARY CONDITION FOR NULL CONTROLLABILITY IN MANY-SERVER HEAVY TRAFFIC

BY GENNADY SHAIKHET[1]

Carleton University

Throughput sub-optimality (TSO), introduced in Atar and Shaikhet [*Ann. Appl. Probab.* **19** (2009) 521–555] for static fluid models of parallel queueing networks, corresponds to the existence of a resource allocation, under which the total service rate becomes greater than the total arrival rate. As shown in Atar, Mandelbaum and Shaikhet [*Ann. Appl. Probab.* **16** (2006) 1764–1804] and Atar and Shaikhet (2009), in the many server Halfin–Whitt regime, TSO implies null controllability (NC), the existence of a routing policy under which, for every finite T, the measure of the set of times prior to T, at which at least one customer is in the buffer, converges to zero in probability at the scaling limit. The present paper investigates the question whether the converse relation is also true and TSO is both sufficient and necessary for the NC behavior.

In what follows we do get the affirmation for systems with either two customer classes (and possibly more service pools) or vice-versa and state a condition on the underlying static fluid model that allows the extension of the result to general structures.

1. Introduction.

In this paper we consider many-server parallel queueing networks in heavy traffic regime. Despite the criticality, as shown in [5, 6], there may exist a scheduling rule, with high probability maintaining the system without waiting customers, for "most of the time." Called *null controllability*, such unusual phenomena occurs under the *throughput sub-optimality* of the underlying (critically loaded in a standard sense), fluid model. In the current work we try to understand if the effect can still be achieved when the underlying fluid is throughput optimal, and conclude that it is not possible and throughput sub-optimality is indeed required.

Our model of interest consists of multiple customer classes, indexed by $\mathcal{I}$, and several service pools, indexed by $\mathcal{J}$, each consisting of many i.i.d. exponential servers. The servers rates depend on both the station and the class. A system administrator dynamically controls all scheduling and routing in the system; see Figure 1. The model is considered in the heavy traffic parametric regime, first proposed by Halfin and Whitt [9], in which the number of servers at each station and the arrival rates grow without bound, proportionally to some $n \uparrow \infty$.

Received September 2013; revised December 2013.

[1] Supported in part by the NSERC Grant N: RGPIN 402292-2011.

MSC2010 subject classifications. 60K25, 68M20, 90B22, 90B36, 60F05.

Key words and phrases. Multiclass queueing systems, heavy traffic, scheduling and routing, throughput optimality, null controllability, buffer-station graph, simple paths.

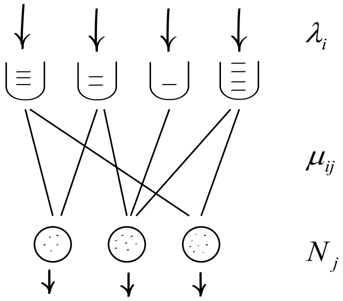

FIG. 1. *A queueing model with four customer classes and three service pools.*

Typically, when analyzing such systems, one looks at the underlying *static fluid model*, obtained in the law of large numbers limit of the processes involved. According to [10, 13], the so-called *static fluid allocation problem* (see Section 2.2 for the details) should then be formulated to determine whether the model (hence, the original system) is under, over or critically loaded; the latter being the proper foundation for the heavy traffic analysis. What one gets is a deterministic matrix ξ^*, where for $(i, j) \in \mathcal{I} \times \mathcal{J}$, the entry ξ^*_{ij} represents the fraction of station-j work dedicated to class-i customers on the fluid level. Consequently, the original network is called critically loaded if the optimized fluid takes 100% of system capacity; that is, $\sum_i \xi^*_{ij} = 1$ for each $j \in \mathcal{J}$. The class-station pairs $(i, j) \in \mathcal{I} \times \mathcal{J}$, along which the service is possible, are called *activities*. The activities (i, j) for which $\xi^*_{ij} > 0$ (resp., $\xi_{ij} = 0$) are regarded as *basic* (resp., *nonbasic*). In both [10, 13] the set ξ^* was assumed to be unique as well as to satisfy the *complete resource pooling* condition, requiring all vertices in the class-station graph to be connected via basic activities. Under the uniqueness assumption, the latter was shown to be equivalent for the graph of basic activities being a tree. The above set of conditions on the underlying fluid model has become standard for considerable amount of works in the conventional (e.g., [10, 11, 13], etc.) and Halfin–Whitt (e.g., [1, 2, 4], etc.) heavy traffic regimes, as well as in the more recent, nondegenerate-slowdown (NDS) regime [3].

With the static fluid model set, an attempt is then made to prove that appropriately scaled (Halfin–Whitt regime) fluctuations of the queueing model about the fluid model converge to a diffusion. Assuming no use of nonbasic activities, the pioneering papers [1, 2, 4] were able to represent the scaled system dynamics as a controlled diffusion with a *drift control*, thus being able to determine asymptotically optimal scheduling policies for the fairly large class of operational costs. It is a general understanding (see Theorem 2.1 in [5] and Theorem 3.3 in this paper) that by including the nonbasic activities one gets additional controls, this time *singular controls*, but the augmented controlled diffusion still remains to be fully analyzed.

Yet, some partial results had been obtained. One of them, called *null controllability* will be the focus of our attention. In particular, in [5], Theorems 2.3, 2.4, it was shown that in the presence of nonbasic activities, some models are prone to a fairly unusual effect when a critically loaded system starts to behave like an underloaded system. More exactly, one can construct a policy, under which for any given $0 < \varepsilon < T < \infty$ all queues in the system are kept empty on the time interval $[\varepsilon, T]$, with probability approaching one (a finiteness of the interval in [5] is crucial, and was later supported by the results from [12] indicating that the phenomenon is not possible in the long run). It is also worth noting that null controllability seems to be the feature of the Halfin–Whitt regime only—by its nature, it cannot happen in the conventional single server asymptotics—and the *conventional-like* NDS regime is not expected to have it either.

The results of [5] were generalized and better explained in [6], attributing the null controllability to what was called *throughput sub-optimality* of the underlying static fluid model, a situation, when (static) resources can be rearranged so that the total service rate becomes greater than the total arrival rate; see Section 2.3 for the exact definition. Throughput sub-optimality, it appears, may occur in wider class of fluid models and, even when the null controllability is impossible, can result in its weaker (though still efficient) version. Namely (Theorem 1 of [6]), for every finite T, the measure of the set of times prior to T, at which at least one customer is in the buffer, converges to zero in probability at the scaling limit.

This brings us to the main objective, to understand the converse relation between throughput sub-optimality and (weak) null controllability (Theorem 2.4). We show that the desired property is rooted in a simply formulated deterministic result (Theorem 3.4) stating that a throughput optimal static fluid model does not become sub-optimal if its fluid amounts are modified along the so-called zero paths, simple paths p from [6] with signed weight $\mu(p) = 0$. This gives a new interesting perspective on zero paths, normally not associated with abrupt changes of fluid material; in contrast with "unwelcome" positive paths $\mu(p) > 0$ that increase the material, or negative paths $\mu(p) < 0$, the existence of which, as shown in Theorem 2 of [6], is equivalent to throughput sub-optimality. Both Theorem 3.4 and its dynamic version Lemma 3.6 are proven for systems with either two customer classes (and possibly more service pools) or vice-versa. Although the full version of Theorem 3.4 still remains unresolved, its simplistic nature [checkable relations (3.13)–(3.14)] allows us to partially generalize the results for arbitrary I and J (Theorem 4.3).

The organization of the paper is rather straightforward, with the main result (Theorem 2.4) followed by its proof (Section 3). Sections 2.1–2.2 provide all the prerequisites, while Section 3.1 is the roadmap for the proof. After that, Theorem 4.3 of Section 4 discusses possible extensions of our findings.

Notation. For a positive integer d and $x \in \mathbb{R}^d$, let $\|x\| = \sum_{i=1}^d |x_i|$. For $v, u \in \mathbb{R}^d$ let $v \cdot u = \sum_{i=1}^d u_i v_i$. The symbols e_i denote the unit coordinate vectors and $e = (1, \dots, 1)$. The dimension of e may change from one expression to another. Thus, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we have $e \cdot x = \sum_{i=1}^d x_i$. For an event A we use

$1\{A\}$ for the indicator of A. Denote by $\mathbb{D}(\mathbb{R}^d)$ the space of all cadlag functions (i.e., right continuous and having finite left limits) from $\mathbb{R}_+$ to $\mathbb{R}^d$. Denote $|X|_t^* = \sup_{0\le u\le t}|X(u)|$ for $X \in \mathbb{D}(\mathbb{R})$, $\|X\|_t^* = \sup_{0\le u\le t}\|X(u)\|$ for $X \in \mathbb{D}(\mathbb{R}^d)$ and $f(t:s) = f(t) - f(s)$.

2. The model and the main result.

2.1. *Original model.* The setting is standard; see, for example, [2, 4–6]. A complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is given, supporting all stochastic processes defined below. There is a sequence of systems indexed by $n \in \mathbb{N}$, each having I customer classes and J service stations. Station j has N_j^n identical servers. The classes are labeled as $1, \dots, I$ and the stations as $I+1, \dots, I+J$. We set $\mathcal{I} = \{1, \dots, I\}$, $\mathcal{J} = \{I+1, \dots, I+J\}$. The arrival and service processes, all mutually independent, are denoted by $\{A_i^n, i \in \mathcal{I}\}$ and $\{S_{ij}^n, (i,j) \in \mathcal{I}\times\mathcal{J}\}$. Each A_i^n is a renewal process whose inter-arrival times have finite second moment and an inverse mean (or rate) equal to $\lambda_i^n > 0$. Each service process S_{ij}^n is a Poisson process with rate $\mu_{ij}^n \ge 0$. We also allow a possibility for $\mu_{ij}^n = 0$, in which case we say that class-i customers cannot be served at station j.

Denote the set of all class-station pairs by $\mathcal{E} := \mathcal{I}\times\mathcal{J}$, let $\mathcal{E}_{\mathrm{a}} = \{(i,j) \in \mathcal{I}\times\mathcal{J} : \mu_{ij}^n > 0\}$, and, throughout, assume that $\mathcal{E}_{\mathrm{a}}$ does not depend on n. A class-station pair $(i,j) \in \mathcal{E}_{\mathrm{a}}$ is said to be an *activity*. The set of class-station pairs that are not activities is denoted by $\mathcal{E}_{\mathrm{a}}^c \equiv \mathcal{E} \setminus \mathcal{E}_{\mathrm{a}}$.

The number of service completions of class-i customers by all servers of station j by time t is therefore (see, e.g., [2, 4–6]), given by $S_{ij}^n(\int_0^t \Psi_{ij}^n(s)\,ds)$, where for every $(i,j) \in \mathcal{E}_{\mathrm{a}}$, we denote by $\Psi_{ij}^n(t)$ the number of class-i customers being served in station j at time t. Denote by $X_i^n(t)$ the number of class-i customers in the system at time t. By definition,

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - \sum_{j\in\mathcal{J}} S_{ij}^n\biggl(\int_0^t \Psi_{ij}^n(s)\,ds\biggr), \qquad i \in \mathcal{I}; \tag{2.1}$$

$$\sum_{j\in\mathcal{J}} \Psi_{ij}^n(t) \le X_i^n(t), \qquad i \in \mathcal{I}; \qquad \sum_{i\in\mathcal{I}} \Psi_{ij}^n(t) \le N_j^n, \qquad j \in \mathcal{J}. \tag{2.2}$$

The processes $\Psi^n = (\Psi_{ij}^n)_{(i,j)\in\mathcal{I}\times\mathcal{J}}$ are regarded as *scheduling control policy* (SCP) and assumed to be right-continuous, taking values in $\mathbb{Z}_+$. Thus

$$\Psi_{ij}^n(t) \ge 0, \qquad (i,j) \in \mathcal{E}_{\mathrm{a}}; \qquad \Psi_{ij}^n(t) = 0, \qquad (i,j) \in \mathcal{E}_{\mathrm{a}}^c. \tag{2.3}$$

Note that the above definition of SCP is very general and does not include the *standard* requirements; see, for example, [2–4, 6].

2.2. *Static fluid model and throughout sub-optimality.* The paper deals with certain properties of an underlying fluid model, to be introduced in this section. We start with the first order approximations of the parameters.

ASSUMPTION 2.1. There exist constants $\lambda_i, \nu_j \in (0, \infty)$, $i \in \mathcal{I}$, $j \in \mathcal{J}$ and $\mu_{ij} \in (0, \infty)$, $(i, j) \in \mathcal{E}_\mathrm{a}$, such that $n^{-1}\lambda_i^n \to \lambda_i, n^{-1}N_j^n \to \nu_j, \mu_{ij}^n \to \mu_{ij}$. Let $\mu_{ij} = 0$ for $(i, j) \in \mathcal{E}_\mathrm{a}^c$.

The above assumption allows one to imagine a model where arrival and service processes are deterministic flows with rates λ_i and μ_{ij}. There are J service stations, processing I classes of incoming fluid. Station j has capacity to hold ν_j units of fluid. Since routing/scheduling is an important part of managing the network, an allocation of work among the stations is pivotal element of the model. The static fluid model uses a fixed allocation for all times (hence "static"). Let Ξ be the set of *allocation* matrices

$$\Xi = \left\{\xi_{ij}, (i, j) \in \mathcal{E}, \text{ such that } \xi_{ij} \geq 0, \text{ and } \sum_{i \in \mathcal{I}} \xi_{ij} \leq 1, \forall j \in \mathcal{J}\right\},$$

where each entry ξ_{ij} represents the fraction of station's j capacity allocated to process class-i. When station j contains $\psi_{ij} := \xi_{ij}\nu_j$ units of class-i fluid, the rate at which this fluid is processed is $\mu_{ij}\psi_{ij} = \bar{\mu}_{ij}\xi_{ij}$, where we set $\bar{\mu}_{ij} = \mu_{ij}\nu_j$. The allocation matrix ξ^* to our model will be chosen according to the following rule.

ASSUMPTION 2.2. Consider the following *s*tatic allocation problem [10]:

$$\min_{\xi \in \Xi, \rho} \rho, \text{ subject to } \sum_{j \in \mathcal{J}} \bar{\mu}_{ij}\xi_{ij} = \lambda_i, \forall i, \sum_{i \in \mathcal{I}} \xi_{ij} \leq \rho, \forall j, \tag{2.4}$$

and assume it has a unique solution (ξ^*, ρ^*), satisfying:

(1) $\rho^* = 1$ and $\sum_{i \in \mathcal{I}} \xi_{ij}^* = 1$ for all $j \in \mathcal{J}$;

(2) the set of activities (edges) $(i, j) \in \mathcal{E}_a$, for which $\xi_{ij}^* > 0$, form a connected tree in a graph with the set of vertices $\mathcal{I} \cup \mathcal{J}$.

For convenience, we choose to keep this *standard* set of assumptions throughout the paper, but, in fact, neither the uniqueness, nor the tree-like structure is crucial. See more explanation in Section 5. For the solution ξ^* from Assumption 2.2 we denote

$$\psi_{ij}^* = \xi_{ij}^*\nu_j, \qquad x_i^* = \sum_j \psi_{ij}^*, \qquad i \in \mathcal{I}, j \in \mathcal{J}. \tag{2.5}$$

Thus x^* represents the mass of material of each class being processed in all service stations. The introduced deterministic model, with parameters $\{\lambda, \nu, \mu\}$ and allocation matrix $\{\psi^*\}$, satisfying Assumption 2.2, will be referred to as the *static fluid model*. Following [10, 13], an activity $(i, j) \in \mathcal{E}_\mathrm{a}$ is said to be *basic* (resp., *nonbasic*) if $\psi_{ij}^* > 0$ (resp., $\psi_{ij}^* = 0$).

Throughput sub-optimality. For $\bar{x} \in \mathbb{R}_+^I$ and $\bar{\nu} \in \mathbb{R}_+^J$, define

$$\Xi(\bar{x}, \bar{\nu}) := \Big\{\psi_{ij}, (i, j) \in \mathcal{E} : \psi_{ij} \geq 0, \sum_{i \in \mathcal{I}} \psi_{ij} \leq \bar{\nu}_j, \forall j \in \mathcal{J} \\ \text{and } \sum_{j \in \mathcal{J}} \psi_{ij} \leq \bar{x}_i, \forall i \in \mathcal{I}\Big\}. \tag{2.6}$$

Note that from (2.5) we have $\psi^* \in \Xi(x^*, \nu)$. Assumption 2.2 expresses the critical load on the system, but does not discard the possibility that the total processing rate can exceed the total arrival rate. For a static fluid model we will say that it is *throughput optimal* if the following holds:

$$\text{Whenever } \psi \in \Xi(x^*, \nu), \text{ one has } \sum_{(i,j) \in \mathcal{E}} \mu_{ij} \psi_{ij} \leq \sum_{i \in \mathcal{I}} \lambda_i. \tag{2.7}$$

The model is said to be *throughput sub-optimal* if it is not throughput optimal.

2.3. *The main result.* The following assumption regards the second order behavior of the parameters and initial condition.

ASSUMPTION 2.3. There exist $c \in (0, \infty)$, independent of n, such that for $n \geq 1$,

$$\|n^{-1}\lambda^n - \lambda\| + \|\mu^n - \mu\| + \|n^{-1}X^n(0) - x^*\| \leq cn^{-1/2}, \\ \|n^{-1}N^n - \nu^n\| \leq (1/2)n^{-1/2}. \tag{2.8}$$

THEOREM 2.4. *Let Assumptions 2.1–2.3 hold. Assume $I = 2$ or $J = 2$. If, for some $T > 0$, there exists a sequence of SCPs, under which*

$$\int_0^T \{e \cdot X^n(s) \geq e \cdot N^n\} \, ds \to 0 \qquad \textit{in probability}, \tag{2.9}$$

then the underlying static fluid model is throughput sub-optimal.

3. Proof of Theorem 2.4.

3.1. *Intuition and preparations.* First, we outline the main ideas of the proof. Fix n. It would be convenient to rescale the system dynamics with respect to the static fluid model. Namely, we rewrite (2.1)–(2.3) in the form

$$\widehat{X}_i^n(t) = \widehat{X}_i^n(0) + \widehat{W}_i^n(t) - \sum_{j \in \mathcal{J}} \mu_{ij} \int_0^t \widehat{\Psi}_{ij}^n(s) \, ds, \qquad i \in \mathcal{I}, \tag{3.1}$$

$$\sum_{j \in \mathcal{J}} \widehat{\Psi}_{ij}^n(t) \leq \widehat{X}_i^n(t), \qquad i \in \mathcal{I}; \qquad \sum_{i \in \mathcal{I}} \widehat{\Psi}_{ij}^n(t) \leq \widehat{N}_j^n, \qquad j \in \mathcal{J}, \tag{3.2}$$

where we use

$$
\begin{aligned}
&\widehat{A}_i^n(t) = n^{-1/2}\bigl(A_i^n(t) - \lambda_i^n t\bigr), \qquad \widehat{S}_{ij}^n(t) = n^{-1/2}\bigl(S_{ij}^n(t) - \mu_{ij}^n t\bigr),\\
&\widehat{X}_i^n(t) = n^{-1/2}\bigl(X_i^n(t) - nx_i^*\bigr), \qquad \widehat{\Psi}_{ij}^n(t) = n^{-1/2}\bigl(\Psi_{ij}^n(t) - n\psi_{ij}^*\bigr),\\
&\widehat{N}_j^n = n^{-1/2}\bigl(N_j^n - n\nu_j\bigr)
\end{aligned}
\tag{3.3}
$$

and

$$
\begin{aligned}
\widehat{W}_i^n(t) = {}& \widehat{A}_i^n(t) - \sum_{j\in\mathcal{J}} \widehat{S}_{ij}^n\biggl(\int_0^t \Psi_{ij}^n(s)\,ds\biggr)\\
&+ n^{-1/2}\bigl(\lambda_i^n - n\lambda_i\bigr)t - n^{-1/2}\sum_{j\in\mathcal{J}}\bigl(\mu_{ij}^n - \mu_{ij}\bigr)\int_0^t \Psi_{ij}^n(s)\,ds.
\end{aligned}
\tag{3.4}
$$

The proof will be completed in several steps. The basic principle would be to show that, once the underlying fluid model is throughput optimal, it is impossible to quickly eliminate a nonnegligible surplus of customers.

• Our main candidate for a fast unloading of the system will be the last term of (3.1), since $\widehat{W}^n$ is well known to be tight; see, for example, [2, 4, 6]. Now, due to throughput optimality (2.6), (2.7), since $\Psi^n \in \Xi(X^n, N^n)$, we have a crude estimate

$$
\sum_{ij}\mu_{ij}\widehat{\Psi}_{ij}^n(t) \le \mu_{\max}\bigl(\|\widehat{X}^n(t)\| + \|\widehat{N}^n\|\bigr), \qquad t \ge 0,
\tag{3.5}
$$

for $\mu_{\max} = \max_{ij}\{\mu_{ij}\}$, which tells us that, in principle, the left-hand side of (3.5) can be made large by *quickly* increasing $\|\widehat{X}^n\|$. Of course, stopping the service (partially or completely) will do the trick, but will not serve our purpose, thus inviting the question whether, and if so, in what directions, $\widehat{X}^n$ can be quickly changed without significant increase of the total mass $e \cdot \widehat{X}^n$.

• To answer the above we would need Theorem 3.3 of Section 3.3, namely, *representation* (3.10), showing that it can be done by using the nonbasic activities along the so-called zero simple paths, the objects first introduced in [6], but with $\mu(p) = 0$. To make this paper self contained we have included Section 3.2, reminiscing about the basic definitions of simple paths from [6] as well as their connection to throughput optimality (Theorem 3.2).

• The representation theorem prompts us back to the static fluid model in an attempt to understand whether one can increase the throughput by inflicting changes along zero paths. The corresponding Theorem 3.4 of Section 3.4 provides the *desired* negative answer and culminates in its dynamic version (Lemma 3.6 of Section 3.5), *essentially* saying that there is *no way* to quickly increase $\sum_{ij}\mu_{ij}\widehat{\Psi}_{ij}^n$ without increasing $e \cdot \widehat{X}^n$, which is quite the opposite of what we are trying to achieve.

• The details are finalized in Section 3.6.

3.2. *Simple paths. Characterization of throughput sub-optimality.* Denote the index set for all customer classes and service stations by $\mathcal{V} := \mathcal{I} \cup \mathcal{J}$. For a nonempty set V and $E \subseteq V \times V$, we write $G = (V, E)$ for the graph with vertex set V and edge set E; see, for example, [8] for standard definitions. A connected graph that does not contain cycles is called a *tree*. We denote $\mathcal{G}_{\mathrm{a}} = (\mathcal{V}, \mathcal{E}_{\mathrm{a}})$ and refer to it as the graph of activities.

Define the graph of basic activities $\mathcal{G}_{ba}$ to be the subgraph of $\mathcal{G}_{\mathrm{a}}$ having $\mathcal{V}$ as a vertex set, and the collection

$$\mathcal{E}_{ba} := \{(i, j) \in \mathcal{E}_{\mathrm{a}} : \xi^*_{ij} > 0\}$$

of basic activities as an edge set. By Assumption 2.2, the graph $\mathcal{G}_{ba}$ is a tree, and by construction of it as a subgraph of $\mathcal{G}_a$, all its edges are of the form (i, j) where $i \in \mathcal{I}$ and $j \in \mathcal{J}$. In the definition below and elsewhere in this section, it will be convenient to identify (i, j) with (j, i) (where $i \in \mathcal{I}$ and $j \in \mathcal{J}$) when referring to an element of the edge set $\mathcal{E}$. Although the notation is abused, there will be no confusion, since $\mathcal{I}$ and $\mathcal{J}$ do not intersect.

DEFINITION 3.1. (i) A subgraph $q = (\mathcal{V}_q, \mathcal{E}_q)$ of $\mathcal{G}_{ba}$ is called a *basic path* if one has $\mathcal{V}_q = \{i_0, j_0, \dots, i_k, j_k\}$ and

$$\mathcal{E}_q = \{(i_0, j_0), (j_0, i_1), \dots, (i_k, j_k)\},$$

where $k \geq 1$ and $i_0, \dots, i_k \in \mathcal{I}$, $j_0, \dots, j_k \in \mathcal{J}$ are $2k + 2$ distinct vertices. Note that every edge of a basic path is a basic activity (i.e., an element of $\mathcal{E}_{ba}$). Basic paths are used to define simple paths, as follows:

(ii) Let the leaves i_0 and j_k of a basic path q be denoted by i^q and, respectively, j^q. The pair (i^q, j^q) could be an activity (an element of $\mathcal{E}_a$), in which case it is necessarily a nonbasic activity (i.e., an element of $\mathcal{E}_a \setminus \mathcal{E}_{ba}$), and we say that the graph $(\mathcal{V}_q, \mathcal{E}_q \cup \{(i^q, j^q)\})$ is a *closed simple path*; otherwise (i^q, j^q) is not an activity (i.e., it is in $\mathcal{E}^c_a$), and we say that q itself is an *open simple path.* We say that p is a *simple path* if it is either a closed or an open simple path. Let *SP* be the set of simple paths.

EXAMPLE. Consider the following static fluid model, with 2 classes of customers and 3 stations:

$$\nu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \lambda = \begin{pmatrix} 8 \\ 4 \end{pmatrix}$$

and

$$\textit{Case A}: \quad \mu = \bar{\mu} = \begin{pmatrix} 3 & 10 & 1 \\ 1 & 4 & 2 \end{pmatrix};$$

$$\textit{Case B}: \quad \mu = \bar{\mu} = \begin{pmatrix} 3 & 10 & 1 \\ 0 & 4 & 2 \end{pmatrix}.$$

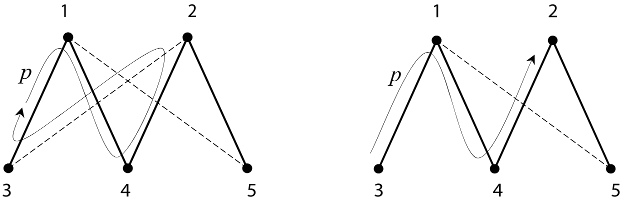

FIG. 2. *Simple paths for cases A and B*: *On the left p is a closed simple path*, *while on the right p is open. For case A*, $\mu_{23} > 0$ *and* $(2, 3)$ *is a nonbasic activity. For case B*, $\mu_{23} = 0$ *and* $(2, 3)$ *is not an activity.*

The resulting optimal static allocation (2.4), (2.5) in both cases is given as

$$\psi^* = \xi^* = \begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{pmatrix} \quad \text{and} \quad x^* = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix},$$

and we can visualize the graph of activities on Figure 2. In both cases we have the same $\mathcal{G}_{ba}$, consisting of vertices $\{1, 2, 3, 4, 5\}$ and edge set $\mathcal{E}_{ba} = \{(1, 3), (1, 4), (2, 4), (2, 5)\}$. Similarly, both cases have two basic paths [recall, we identify (i, j) with (j, i)]

$$q_1 = \{(3, 1), (1, 4), (4, 2)\} \quad \text{and} \quad q_2 = \{(1, 4), (4, 2), (2, 5)\}.$$

The basic path q_1, together with the corresponding leaves 3 and 2, defines a path $p = \{(3, 1), (1, 4), (4, 2), (2, 3)\}$ which will be closed if $\mu_{23} > 0$ [i.e., (2, 3) is an activity, case A] and open otherwise (case B). The only other possible simple path $\{(5, 2), (2, 4), (4, 1), (1, 5)\}$ in both cases will be a closed one.

Next, we associate directions with edges of simple paths. Let p be a simple path, and let $q = q^p = (\mathcal{V}_q, \mathcal{E}_q)$ be the corresponding basic path with $\mathcal{E}_q = \{(i_0, j_0), \ldots, (i_k, j_k)\}$, where $i_0, \ldots, i_k \in \mathcal{I}$ and $j_0, \ldots, j_k \in \mathcal{J}$. The direction that will be associated with the edges in $\mathcal{E}_q$, when considered as edges of p, is as follows: $j_k \to i_k \to j_{k-1} \to i_{k-1} \to \cdots \to j_0 \to i_0$. In the case of an open simple path, this exhausts all edges of p. In the case of a closed simple path, the direction of (i_0, j_k) is $i_0 \to j_k$. We note that an edge (corresponding to a basic activity) may have different directions when considered as an edge of different simple paths. We signify the directions along simple paths by $s(p, i, j)$, defined for $i \in \mathcal{I}$, $j \in \mathcal{J}$, $(i, j) \in \mathcal{E}_p$, $p \in SP$, as

$$s(p, i, j) = \begin{cases} -1, & \text{if } (i, j)\text{, considered as an edge of } p\text{,} \\ & \text{is directed from } i \text{ to } j, \\ 1, & \text{if } (i, j)\text{, considered as an edge of } p\text{,} \\ & \text{is directed from } j \text{ to } i. \end{cases} \tag{3.6}$$

Set

$$m_{i,p} = \sum_{j:(i,j)\in p} s(p, i, j)\mu_{ij}, \qquad i \in \mathcal{I}, \qquad m_p = (m_{i,p}, i \in \mathcal{I}) \tag{3.7}$$

and

$$\mu(p) = \sum_{i:(i,j)\in p} m_{i,p} = \sum_{(i,j)\in\mathcal{E}_p} s(p,i,j)\mu_{ij}, \qquad i \in \mathcal{I}. \tag{3.8}$$

EXAMPLE (cont.). Referring to the simple path p, for case A we have $\mu(p) = -7 + 3 = -4$ since

$$m_{1,p} = s(p,1,3)\mu_{13} + s(p,1,4)\mu_{14} = \mu_{13} - \mu_{14} = -7,$$
$$m_{2,p} = s(p,2,3)\mu_{23} + s(p,2,4)\mu_{24} = -\mu_{23} + \mu_{24} = 3.$$

Similarly, for the case B we have $m_{1,p} = -7$, $m_{2,p} = 4$ (since $\mu_{23} = 0$) and $\mu(p) = -3$.

THEOREM 3.2 (Theorem 2, [6]). *Let Assumptions 2.2 and 2.3 hold. Then the following statements are equivalent*:

(1) *the static fluid model is throughput sub-optimal*;
(2) *there exists a simple path $p \in SP$ such that $\mu(p) < 0$.*

EXAMPLE (cont.). Both cases have a path with $\mu(p) < 0$, hence both are throughput sub-optimal. To see that, for example, the fluid model in case A is throughput sub-optimal, let $\beta > 0$ be sufficiently small, and consider the allocation matrix

$$\widehat{\xi} = \begin{pmatrix} 1-\beta & 0.5+\beta & 0 \\ \beta & 0.5-\beta & 1 \end{pmatrix}.$$

Clearly, we have $\sum_j \widehat{\xi}_{ij}\nu_j = x_i^*$ for every i. However, $\sum_{(i,j)\in\mathcal{E}} \widehat{\xi}_{ij}\bar{\mu}_{ij} > \lambda_1 + \lambda_2$.

3.3. *Representation.*

THEOREM 3.3. *Let X^n and Ψ^n satisfy* (2.1)–(2.3). *Then there exist processes Φ^n, M^n and Υ^n, satisfying*:

(1) $\Phi^n(t) \in \Xi(X^n(t), N^n)$ *for* $t \geq 0$ *and* $\Phi^n_{ij} \equiv 0$ *for* $(i,j) \notin \mathcal{E}_{ba}$;
(2) $\|\widehat{\Phi}^n(t)\| \leq c_F(\|\widehat{X}^n(t)\| + \|\widehat{N}^n\|)$ *for some constant* c_F, *independent of* t, n;
(3) $M^n \in \mathbb{D}(\mathbb{R}^{|SP|})$, $\Upsilon^n \in \mathbb{D}(\mathbb{R}^{|\mathcal{I}|})$, *are component-wise nondecreasing, initially zero, so that the following holds for the scaled processes*[2] *for* $t \geq 0$, $i \in .\mathcal{I}$

$$\begin{aligned}\widehat{X}^n_i(t) = \widehat{X}^n_i(0) + \widehat{W}^n_i(t) - \sum_{j:(i,j)\in\mathcal{E}_a} \mu_{ij} \int_0^t \widehat{\Phi}^n_{ij}(s)\,ds \\ + \sum_{p\in SP} m_{i,p}\widehat{M}^n_p(t) + \widehat{\Upsilon}^n_i(t).\end{aligned} \tag{3.9}$$

[2] We use $\widehat{\Phi}^n_{ij} := n^{-1/2}(\Phi^n_{ij} - n\psi^*_{ij})$, $\widehat{M}^n := n^{-1/2}M^n$ and $\widehat{\Upsilon}^n := n^{-1/2}\Upsilon^n$.

The proof is relegated to the Appendix. Together with inequality (2), which is obviously stronger than (3.5), the theorem indicates that the last two terms of (3.9) are the only possible reasons for the abrupt change of $\|\widehat{X}^n\|$. The summation term is associated with simple paths, while the last term corresponds to direct nonwork conservation; see the proof for more details. The theorem can be viewed as a generalization of Theorem 2.1 from [5] where only closed simple paths (called cycles) were considered.

For a simple path $p \in SP$, we say $p \in \mathcal{P}_0$, (resp., $p \in \mathcal{P}_-$; or $p \in \mathcal{P}_+$) if $\mu(p) = 0$, [resp., $\mu(p) < 0$; or $\mu(p) > 0$]. Depending on the subscript sign of $\mathcal{P}$ the paths will be called, respectively, *zero*, *negative* or *positive* paths.

If the static fluid model is throughput optimal (in which case Theorem 3.2 implies $\mathcal{P}_- = \varnothing$), we rewrite (3.9) as

$$\widehat{X}_i^n(t) = \widehat{X}_i^n(0) + \widehat{W}_i^n(t) - \sum_{j:(i,j)\in\mathcal{E}_{ba}} \mu_{ij} \int_0^t \widehat{\Phi}_{ij}^n(s)\,ds + \widehat{\zeta}_i^n(t) + \widehat{\eta}_i^n(t), \tag{3.10}$$
$$i \in \mathcal{I},$$

where

$$\widehat{\zeta}_i^n(t) = \sum_{p\in\mathcal{P}_0} m_{i,p}\widehat{M}_p^n(t), \qquad \widehat{\eta}_i^n(t) = \sum_{p\in\mathcal{P}_+} m_{i,p}\widehat{M}_p^n(t) + \widehat{\Upsilon}_i^n(t). \tag{3.11}$$

Notice that $\widehat{\zeta}^n$ and $\widehat{\eta}^n$ satisfy [due to (3.8) and nonnegativity of Υ_i^n]

$$e \cdot \widehat{\zeta}^n(t) \equiv 0, \qquad e \cdot \widehat{\eta}^n(t) \geq 0. \tag{3.12}$$

3.4. *Discarding zero paths.* From (3.10)–(3.12) we see that both $\widehat{\zeta}^n$ and $\widehat{\eta}^n$ can lead to abrupt increase of $\|\widehat{X}^n\|$, though only $\widehat{\eta}^n$ that can do such for $e \cdot \widehat{X}^n$. The next deterministic (key!) result, viewed as a prelude to estimate (3.27) of Lemma 3.6, discards any significant influence of zero paths (represented by $\widehat{\zeta}^n$) on system's drift.

THEOREM 3.4. *Assume that the static fluid model*, (*as defined in Section* 2.2), *is throughput optimal. Take an arbitrary vector* $M \in \mathbb{R}_+^{|\mathcal{P}_0|}$, *with* $\|M\|$ *small enough*, *and set*

$$x = x^* + \sum_{p\in\mathcal{P}_0} m_p M_p. \tag{3.13}$$

Then, *if either* $I = 2$ *or* $J = 2$, *the following inequality is true*:

$$\sum_{ij} \mu_{ij}\psi_{ij} \leq \sum_{ij} \mu_{ij}\psi_{ij}^* \tag{3.14}$$

for all $\psi \in \Xi(x, \nu)$.

Before proving the theorem, we point out an important corollary.

COROLLARY 3.5. *Let the static fluid model be throughput optimal. Assume we are given some $x_0 \in \mathbb{R}^I_+$, $\widetilde{\nu} \in \mathbb{R}^J_+$, $\gamma \in \mathbb{R}^I_+$ and a set of numbers $\{M_p \ge 0, p \in \mathcal{P}_0 \cup \mathcal{P}_+\}$ with $\|M\|$ sufficiently small. Define $\widetilde{x} = x_0 + \zeta + \eta$, where*

$$\zeta_i = \sum_{p \in \mathcal{P}_0} m_{i,p} M_p, \qquad \eta_i := \sum_{p \in \mathcal{P}_+} m_{i,p} M_p + \gamma_i \qquad \forall i \in \mathcal{I}. \tag{3.15}$$

Then, if either $I = 2$ or $J = 2$, for all $\psi \in \Xi(\widetilde{x}, \widetilde{\nu})$, we have

$$\sum_{ij} \mu_{ij}(\psi_{ij} - \psi^*_{ij}) \le c_\mu(\|x_0 - x^*\| + \|\widetilde{\nu} - \nu\| + e \cdot \eta), \tag{3.16}$$

where c_μ is a constant, independent of ξ, η, M.

PROOF. Just note that $\widetilde{x} = x^* + \zeta + (x_0 - x^*) + \eta = x + (x_0 - x^*) + \eta$ for x from (3.13), together with (3.14) yielding $\sum_{ij} \mu_{ij}(\psi_{ij} - \psi^*_{ij}) \le \mu_{\max}(\|x_0 - x^*\| + \|\widetilde{\nu} - \nu\| + \|\eta\|)$ and the corollary follows since $\mu(p) > 0$ for each $p \in \mathcal{P}_+$, and

$$\begin{aligned}\mu_{\max}\|\eta\| &\le \mu_{\max}\biggl(\sum_{p \in \mathcal{P}_+} \|m_p\| M_p + e \cdot \gamma_i\biggr) \\ &\le c_\mu\biggl(\sum_{p \in \mathcal{P}_+} \mu(p) M_p + e \cdot \gamma_i\biggr) = c_\mu(e \cdot \eta)\end{aligned} \tag{3.17}$$

for $c_\mu = \mu_{\max}(1 + \min\{c \ge 0 \colon \|m_p\| \le c\mu(p), \text{for all} p \in \mathcal{P}_+\})$. □

PROOF OF THEOREM 3.4. We will start with a basic case when $I = J = 2$ then extend it to more general systems.

Case 1: let $\mathcal{I} = \{1, 2\}$ and $\mathcal{J} = \{3, 4\}$, and assume the (unique) basic path is given as $q = \{(3, 1), (1, 4), (4, 2)\}$ with $(2, 3)$ being either nonbasic activity or not an activity. The corresponding simple path p belongs to $\mathcal{P}_0$, and hence satisfies

$$m_{1,p} + m_{2,p} = (\mu_{13} - \mu_{14}) + (\mu_{24} - \mu_{23}) = 0. \tag{3.18}$$

Take a small enough $M > 0$, set $\Delta := (\mu_{13} - \mu_{14})M = (\mu_{23} - \mu_{24})M$ and define a new $x = (x_1, x_2) = (x^*_1 + \Delta, x^*_2 - \Delta)$. Because of (3.18), an elementary argument implies that any *throughput optimizing* allocation matrix ψ is of the form $\psi = \psi^\gamma$

$$(\psi^\gamma_{13}, \psi^\gamma_{14}, \psi^\gamma_{23}, \psi^\gamma_{24}) = (\psi^*_{13} - \gamma, \psi^*_{14} + \Delta + \gamma, \gamma, \psi^*_{24} - \Delta - \gamma) \tag{3.19}$$

for some $0 \le \gamma \le \min\{\psi^*_{13}, \psi^*_{24} - \Delta\}$, with the total throughput remaining a constant, independent of γ,

$$\sum_{ij} \psi^\gamma_{ij}\mu_{ij} \equiv \psi^*_{13}\mu_{13} + (\psi^*_{14} + \Delta)\mu_{14} + (\psi^*_{24} - \Delta)\mu_{24}. \tag{3.20}$$

Assume, on the contrary, that $\sum_{ij} \psi^\gamma_{ij}\mu_{ij} > \lambda_1 + \lambda_2$. Due to (3.20), the latter inequality will hold for any feasible γ. In particular, take $\gamma_0 = M\mu_{24}$. It is easy to check that with such a choice, we have [recall $\Delta = M(\mu_{23} - \mu_{24})$]

$$\psi^{\gamma_0}_{23}\mu_{23} + \psi^{\gamma_0}_{24}\mu_{24} = \lambda_2. \tag{3.21}$$

Together with $\sum_{ij} \psi^{\gamma_0}_{ij} \mu_{ij} > \lambda_1 + \lambda_2$, it means

$$\psi^{\gamma_0}_{13}\mu_{13} + \psi^{\gamma_0}_{14}\mu_{14} > \lambda_1, \tag{3.22}$$

clearly contradicting the static fluid allocation problem; see Assumption 2.2. Indeed, (3.21)–(3.22) means there is a static fluid allocation $(\widetilde{\psi}_{13}, \psi^{\gamma_0}_{14}, \psi^{\gamma_0}_{23}, \psi^{\gamma_0}_{24})$, with $\widetilde{\psi}_{13} < \psi^{\gamma_0}_{13}$, that *fully serves* each of the two incoming classes without using all the capacity.

Case 2: now consider the case $I = 2$ or $J = 2$. An important property of such systems is that each simple path consists of four vertices and three or four edges, depending whether or not it is open or closed; and the arguments from case 1 will be very helpful. In particular, we argue that the statement of the theorem remains true if *only one* zero path modification is applied, that is, if $x = x^* + m_p M_p$ for some path $p \in \mathcal{P}_0$, then $\sum_{ij} \mu_{ij}\psi_{ij} \le \sum_{ij} \mu_{ij}\psi^*_{ij}$ for any $\psi \in \Xi(x, \nu)$. Indeed, let, on the contrary, there exist a throughput maximizing allocation $\psi \in \Xi(x, \nu)$ satisfying $\sum_{ij} \mu_{ij}\psi_{ij} > \sum_{ij} \mu_{ij}\psi^*_{ij}$. Let $\mathcal{V}_p = \{i_1, i_2, j_1, j_2\}$ with a nonbasic (i_2, j_1). Then, again, due to $\mu_{i_1,j_1} - \mu_{i_1,j_2} + \mu_{i_2,j_2} - \mu_{i_2,j_1} = 0$ [recall (3.8) that $p \in \mathcal{P}_0$], we have that the following allocation:

$$\begin{aligned}
&\psi^o_{ij} = \psi^*_{ij} \qquad \text{for } (i, j) \notin p,\\
&(\psi^o_{i_1,j_1}, \psi^o_{i_1,j_2}, \psi^o_{i_2,j_1}, \psi^o_{i_2,j_2})\\
&\qquad = (\psi^*_{i_1,j_1} - \gamma, \psi^*_{i_1,j_2} + \Delta + \gamma, \gamma, \psi^*_{i_2,j_2} - \Delta - \gamma)
\end{aligned} \tag{3.23}$$

will satisfy $\sum_{ij} \mu_{ij}\psi^o_{ij} = \sum_{ij} \mu_{ij}\psi_{ij} > \sum_{ij} \mu_{ij}\psi^*_{ij}$ for any feasible γ and $\Delta = M_p m_{i_1,p} = -M_p m_{i_2,p}$, bringing us precisely to the first case and, hence, to a contradiction.

Now we extend the latter to several zero paths. Set $k = |\mathcal{P}_0| > 1$. Once again, assume that there exists a throughput maximizing matrix $\psi \in \Xi(x, \nu)$ that satisfies $\sum_{ij} \mu_{ij}\psi_{ij} > \sum_{ij} \mu_{ij}\psi^*_{ij}$. Consider an allocation matrix $\overline{\psi}$ of the form $\overline{\psi}_{ij} = \sum_{p \in \mathcal{P}_0} \psi^{(p)}_{ij}$, where [slightly abusing the notation and denoting $\mathcal{V}_p = \{i^p_1, i^p_2, j^p_1, j^p_2\}$ with a nonbasic (i^p_2, j^p_1) per each path p],

$$\begin{aligned}
&\psi^{(p)}_{ij} = \frac{1}{k}\psi^*_{ij} \qquad \text{for } (i, j) \notin \mathcal{E}_p,\\
&(\psi^{(p)}_{i^p_1,j^p_1}, \psi^{(p)}_{i^p_1,j^p_2}, \psi^{(p)}_{i^p_2,j^p_1}, \psi^{(p)}_{i^p_2,j^p_2})\\
&\qquad = \left(\frac{1}{k}\psi^*_{i^p_1,j^p_1}, \frac{1}{k}\psi^*_{i^p_1,j^p_2} + \Delta^p, 0, \frac{1}{k}\psi^*_{i^p_2,j^p_2} - \Delta^p\right),
\end{aligned} \tag{3.24}$$

with $\Delta^p = M_p m_{i^p_1,p} = -M_p m_{i^p_2,p}$. Once again, since each simple path p belongs to $\mathcal{P}_0$, we have $\sum_{ij} \mu_{ij}\overline{\psi}_{ij} = \sum_{ij} \mu_{ij}\psi_{ij} > \sum_{ij} \mu_{ij}\psi^*_{ij}$. Now consider k *completely separated from each other* systems with identical set $\{\mu_{ij}\}$, but with arrival rates and capacities divided by k. Clearly, the values $\{\frac{1}{k}\psi^*_{ij}\}$ will solve the static fluid allocation problem in the smaller systems. Let each of the smaller systems correspond to each of the possible $p \in \mathcal{P}_0$. To each system apply a modification along the corresponding path

$$x^{(p)} = \frac{1}{k} x^* + m_p M_p. \tag{3.25}$$

The allocation $\{\psi_{ij}^{(p)}\}$ from (3.24) optimizes the throughput in the corresponding small system and satisfies (since we have already treated the case when only one $p \in \mathcal{P}_0$ has been activated)

$$\sum_{ij} \mu_{ij} \psi_{ij}^{(p)} \leq \frac{1}{k} \sum_{ij} \mu_{ij} \psi_{ij}^*, \tag{3.26}$$

implying overall

$$\sum_{ij} \mu_{ij} \overline{\psi}_{ij} = \sum_{ij} \mu_{ij} \Bigl(\sum_{p \in \mathcal{P}_0} \psi_{ij}^{(p)} \Bigr) = \sum_{p \in \mathcal{P}_0} \Bigl(\sum_{ij} \mu_{ij} \psi_{ij}^{(p)} \Bigr) \leq \sum_{ij} \mu_{ij} \psi_{ij}^*,$$

which completes the proof by contradiction. □

3.5. *Important estimate.* Consider the event $\Omega_w^n = \{\|\widehat{A}^n\|_1^* + \|\widehat{S}^n\|_1^* \leq 5\}$.

LEMMA 3.6. *Let Assumptions 2.1–2.3 hold, assume that the static fluid model is throughput optimal, and let $I = 2$ or $J = 2$. Then, on the event Ω_w^n, for any scheduling policy, we have, for $\varepsilon > 0$ small enough and $t \leq 2\varepsilon$,*

$$\sum_{ij} \mu_{ij} \widehat{\Psi}_{ij}^n(t) \leq \varepsilon^{-2/3} \bigl(1 + \bigl|(e \cdot \widehat{X}^n)^+\bigr|_t^*\bigr). \tag{3.27}$$

REMARK 3.7. In fact, the above inequality holds for some constant κ, but for our purposes a crude bound of $\kappa < \varepsilon^{-2/3}$ will be enough as it saves us the trouble of adjusting *essentially irrelevant* constants after each operation.

PROOF OF LEMMA 3.6. We will start by showing the relation

$$\sum_{ij} \mu_{ij} \widehat{\Psi}_{ij}^n(t) \leq \varepsilon^{-1/2} \bigl(1 + \bigl(e \cdot \widehat{X}^n(t)\bigr)^+ + e \cdot \widehat{\eta}^n(t)\bigr), \qquad t \geq 0. \tag{3.28}$$

Recall (3.5)

$$\sum_{ij} \mu_{ij} \widehat{\Psi}_{ij}^n(t) \leq \mu_{\max} \bigl(\|\widehat{X}^n(t)\| + \|\widehat{N}^n\|\bigr), \qquad t \geq 0. \tag{3.29}$$

Due to (2.8) and since $e \cdot \widehat{\eta}^n$ and is a nondecreasing process starting at zero, inequality (3.28) will follow for all t when $\|\widehat{X}^n(t)\| \leq \varepsilon^{-1/3}(1 + e \cdot \widehat{\eta}^n(t))$. Now consider the case when $\|\widehat{X}^n(t)\| > \varepsilon^{-1/3}(1 + e \cdot \widehat{\eta}^n(t))$.

First, assume there is only one class i with $|\widehat{X}^n_i(t)| > \frac{\varepsilon^{-1/3}}{I}(1 + e \cdot \widehat{\eta}^n(t))$. If $\widehat{X}^n_i(t) < 0$, relation (3.28) clearly follows from (3.29) since the left-hand side of (3.28) would only increase if $\widehat{X}^n_i(t)$ is increased to $-\frac{\varepsilon^{-1/3}}{I}(1 + e \cdot \widehat{\eta}^n(t))$. Otherwise, if $\widehat{X}^n_i(t) > 0$, relation (3.28) follows from

$$\bigl(e \cdot \widehat{X}^n(t)\bigr)^+ \geq e \cdot \widehat{X}^n(t) \geq \widehat{X}^n_i(t) - \varepsilon^{-1/3}\bigl(1 + e \cdot \widehat{\eta}^n(t)\bigr),$$

because for all other classes $j \neq i$ we have $\widehat{X}^n_j(t) \geq -|\widehat{X}^n_j(t)| \geq -\frac{\varepsilon^{-1/3}}{I}(1 + e \cdot \widehat{\eta}^n(t))$.

For the rest of the proof assume that $|\widehat{X}^n_i(t)| > \frac{\varepsilon^{-1/3}}{I}(1 + e \cdot \widehat{\eta}^n(t))$ for several different i's. From (3.9)–(3.12) we have (using the fact $t \leq 2\varepsilon$)

$$\begin{aligned} \|\widehat{X}^n\|^*_t \leq{}& \|\widehat{X}^n(0)\| + \|\widehat{W}^n\|^*_t + 2\varepsilon c_F\bigl(\|\widehat{X}^n\|^*_t + \|\widehat{N}^n\|\bigr) + \|\widehat{\zeta}^n(t)\| \\ &+ \|\widehat{\eta}^n(t)\|. \end{aligned} \tag{3.30}$$

Using (3.17), we have $\|\eta^n(t)\| \leq (c_\mu/\mu_{\max})e \cdot \eta^n(t)$. Moreover, due to the lemma's assumptions, we have [see (3.4)] $\|\widehat{X}^n(0)\| + \|\widehat{W}^n\|^*_\varepsilon \leq \varepsilon^{-1/6}$ for ε small enough, altogether implying

$$\|\widehat{X}^n(t)\| \leq \|\widehat{X}^n\|^*_t \leq \varepsilon^{-1/6}\bigl(1 + \|\widehat{\zeta}^n(t)\| + e \cdot \widehat{\eta}^n(t)\bigr). \tag{3.31}$$

Since $\|\widehat{X}^n(t)\| > \varepsilon^{-1/3}(1 + e \cdot \widehat{\eta}^n(t))$, inequality (3.31) would imply

$$\|\widehat{\zeta}^n(t)\| \geq \bigl(\varepsilon^{-1/6} - 1\bigr)\bigl(1 + e \cdot \widehat{\eta}^n(t)\bigr), \tag{3.32}$$

that is, there is at least one large "zero path" (i.e., $p \in \mathcal{P}_0$) activity usage and we are going to apply Corollary 3.5 to "filter out" the effect of such.

First, if $I > 2$, $J = 2$, then all vertices $i \in \mathcal{I}$, except for one (denote it by k), are leaves in the tree of basic activities $\mathcal{G}_{ba}$. For each leaf i_0 there is a unique simple path p, going through i_0 and k.

Consider the following procedure: Let $\mathcal{I}_0 = \mathcal{I}_0(t) = \{i \in \mathcal{I} \setminus \{k\} : |\widehat{X}^n_{i_0}(t)| > \frac{\varepsilon^{-1/3}}{I}(1 + e \cdot \widehat{\eta}^n(t))\}$. For $i \in \mathcal{I}_0(t)$ define $\widehat{x}_i := \frac{\widehat{X}^n_i(t)}{|\widehat{X}^n_i(t)|}\frac{\varepsilon^{-1/3}}{I}(1 + e \cdot \widehat{\eta}^n(t))$, and set $\widehat{x}_k := \widehat{X}^n_k(t) + \sum_{i \in \mathcal{I}_0(t)}(\widehat{X}^n_i(t) - \widehat{x}_i)$. Finally, for $i \notin (\mathcal{I}_0 \cup \{k\})$, set $\widehat{x}_i = \widehat{X}^n_i(t)$. Viewing vector $\widehat{X}^n$ as if it has been obtained from $\widehat{x}$ by applying $|\mathcal{I}_0|$ zero paths to the latter [as (2.8) we obviously have $\|\widehat{X}^n\|^*_\varepsilon \leq \|\widehat{X}^n(0)\| + \|\widehat{A}^n\|^*_1 + 2cn^{1/2}\varepsilon \leq \frac{n^{-1/2}}{|SP|}(n \min_{i,j} \psi^*_{ij})$ on Ω^n_w, so the perturbation is indeed *sufficiently small* when viewed on the fluid level], one can use Corollary 3.5 to get

$$\begin{aligned} \sum_{ij} \mu_{ij} \widehat{\Psi}^n_{ij}(t) &\leq c_\mu\bigl(\|\widehat{N}^n\| + \|\widehat{x}\| + e \cdot \widehat{\eta}^n(t)\bigr) \\ &\leq c_\mu\bigl(\|\widehat{N}^n\| + \widehat{x}^+_k + \varepsilon^{-1/3}\bigl(1 + e \cdot \widehat{\eta}^n(t)\bigr) + e \cdot \widehat{\eta}^n(t)\bigr). \end{aligned} \tag{3.33}$$

In the last inequality we once again use the fact that only strictly positive $\widehat{x}_k$ was worth considering [otherwise the left-hand side of (3.33) would only increase if

$\widehat{x}_k$ is increased to $-\frac{\varepsilon^{-1/3}}{I}(1 + e \cdot \widehat{\eta}^n(t))]$. A crude estimate $\widehat{x}_k^+ \le (e \cdot \widehat{X}^n(t))^+ + \varepsilon^{-1/3}(1 + e \cdot \widehat{\eta}^n(t))$ that follows from the definition of $\widehat{x}_i$ and the relation $\widehat{x}_k = e \cdot \widehat{X}^n(t) - \sum_{i \in \mathcal{I}_0} \widehat{x}_i$ completes the proof of (3.28). If $I = 2$, the same procedure is applied only once, along any of several possible zero paths. This proves (3.28).

To finalize the lemma, note that Φ^n from (3.9)–(3.12) satisfies $\Phi^n(t) \in \Xi(X^n(t), N^n)$ for all t in the given range, hence is subject to (3.28) as well. Using that, we have

$$|(e \cdot \widehat{X}^n)^+|_t^* \ge e \cdot \widehat{X}^n(0) - e \cdot \widehat{W}^n(t) + e \cdot \widehat{\eta}^n(t) \tag{3.34}$$

$$- \varepsilon^{1/2}(1 + |(e \cdot \widehat{X}^n)^+|_t^* + e \cdot \widehat{\eta}^n(t)) \tag{3.35}$$

$$\ge C + \tfrac{1}{2} e \cdot \widehat{\eta}^n(t) - \varepsilon^{1/2}|(e \cdot \widehat{X}^n)^+|_t^* \tag{3.36}$$

implying

$$e \cdot \widehat{\eta}^n(t) \le \varepsilon^{-1/5}(1 + |(e \cdot \widehat{X}^n)^+|_t^*), \tag{3.37}$$

and we complete the proof by substituting (3.37) into (3.28). □

3.6. *Finalizing the proof.* For arbitrary $\varepsilon > 0$, consider the event

$$\begin{aligned} \Omega_1^n = \Omega_1^n(\varepsilon) = \Omega_w^n &\cap \{e \cdot \widehat{X}^n(0) - e \cdot \widehat{N}^n + e \cdot \widehat{A}^n(\varepsilon) \ge 4\} \\ &\cap \{\|\widehat{A}^n(\cdot) - \widehat{A}^n(\varepsilon)\|_{[\varepsilon,2\varepsilon]}^* \le 1/4\} \\ &\cap \{\|\widehat{S}^n\|_1^* \le 1/4\}. \end{aligned} \tag{3.38}$$

It is standard (e.g., Theorem 14.6 in [7]) that component-wise both $\widehat{A}^n$ and $\widehat{S}^n$ converge weakly to *independent* Brownian motion processes. Therefore there exist constants $n_1 = n_1(\varepsilon) \in \mathbb{N}$ and $\delta = \delta(\varepsilon) > 0$, so that $\mathbb{P}(\Omega_1^n) > \delta$ for all $n \ge n_1$.

Fix $\varepsilon > 0$. Theorem 2.4 guarantees that there exists a sequence of SCPs satisfying

$$\lim_{n \to \infty} \mathbb{P}\left(\Omega_1^n \cap \left\{\int_0^T 1\{e \cdot X^n(s) \ge e \cdot N^n\} \, ds > \varepsilon\right\}\right) = 0. \tag{3.39}$$

Let $\Omega^n = \Omega_1^n \cap \{\int_0^T 1\{e \cdot X^n(s) \ge e \cdot N^n\} \, ds \le \varepsilon\}$. Relation (3.39) implies that there exists a constant $n_0(\varepsilon) \in \mathbb{N}$ so that

$$\mathbb{P}(\Omega^n) > \frac{\delta}{2} \qquad \text{for all } n \ge n_0. \tag{3.40}$$

In what follows we assume that the static fluid model is *throughput optimal* and will come to a conclusion that the event Ω^n is impossible (i.e., Ω^n is an empty set) for $n \ge n_0$ and ε small enough, thus contradicting (3.40).

From (3.1), (3.4), (2.8), Lemma 3.6 and (3.38) on the event Ω^n,

$$e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n \ge 7/2 - c\varepsilon - \varepsilon^{1/3}(1 + |(e \cdot \widehat{X}^n)^+|_\varepsilon^*), \tag{3.41}$$

which, for ε small enough, yields

$$e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n + \varepsilon^{1/3} \bigl| (e \cdot \widehat{X}^n)^+ \bigr|_\varepsilon^* \ge 2, \tag{3.42}$$

giving us two possible scenarios: $e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n \ge \varepsilon^{1/3} |(e \cdot \widehat{X}^n)^+|_\varepsilon^*$ and $\varepsilon^{1/3} |(e \cdot \widehat{X}^n)^+|_\varepsilon^* \ge e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n$.

Case 1. Assume $e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n \ge \varepsilon^{1/3} |(e \cdot \widehat{X}^n)^+|_\varepsilon^*$. Together with (3.42), this implies $e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n \ge 1$. Let $\tau_\varepsilon = \inf\{t > \varepsilon : e \cdot \widehat{X}^n(t) = e \cdot \widehat{N}^n\}$. Notice that τ_ε is well defined since the jumps of $e \cdot X^n$ are of size 1 and, moreover, satisfies $\tau_\varepsilon < 2\varepsilon$ on Ω^n, because the total queueing time does not exceed ε. Using $e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n \ge 1$, (3.1), Lemma 3.6 and (3.38) we can write

$$\begin{aligned} 0 &= e \cdot \widehat{X}^n(\tau_\varepsilon) - e \cdot \widehat{N}^n \\ &\ge e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n + e \cdot \widehat{W}^n(\tau_\varepsilon : \varepsilon) - \varepsilon^{1/3}\bigl(1 + \bigl|(e \cdot \widehat{X}^n)^+\bigr|_{\tau_\varepsilon}^*\bigr) \\ &\ge 1/8 - \varepsilon^{1/3} \bigl|(e \cdot \widehat{X}^n)^+\bigr|_{\tau_\varepsilon}^*, \end{aligned}$$

implying

$$\bigl|(e \cdot \widehat{X}^n)^+\bigr|_{\tau_\varepsilon}^* \ge \varepsilon^{-1/4}. \tag{3.43}$$

In other words, a large queue of at least $\varepsilon^{-1/4}$ has to be eliminated before time τ_ε. Let α be the *last* time before τ_ε, satisfying $|(e \cdot \widehat{X}^n)^+|_{\tau_\varepsilon}^* = e \cdot \widehat{X}^n(\alpha) \ge \varepsilon^{-1/4}$. We have

$$\begin{aligned} 0 &= e \cdot \widehat{X}^n(\tau_\varepsilon) - e \cdot \widehat{N}^n \\ &\ge e \cdot \widehat{X}^n(\alpha) - e \cdot \widehat{N}^n + e \cdot \widehat{W}^n(\tau_\varepsilon : \alpha) - \varepsilon^{1/3}\bigl(1 + e \cdot \widehat{X}^n(\alpha)\bigr) \\ &\ge C + \tfrac{1}{2} e \cdot \widehat{X}^n(\alpha) \ge C + (1/2)\varepsilon^{-1/4}, \end{aligned}$$

for some constant C, which is an obvious contradiction for ε small enough.

Case 2. If $\varepsilon^{1/3} |(e \cdot \widehat{X}^n)^+|_\varepsilon^* \ge e \cdot \widehat{X}^n(\varepsilon) - e \cdot \widehat{N}^n$, then $|(e \cdot \widehat{X}^n)^+|_\varepsilon^* \ge \varepsilon^{-1/3}$ by (3.42), and the same considerations as in the previous case can be applied. Let α be the *last* time before ε, satisfying $|(e \cdot \widehat{X}^n)^+|_\varepsilon^* = e \cdot \widehat{X}^n(\alpha) \ge \varepsilon^{-1/3}$, and define $\tau_\alpha = \inf\{t > \alpha : e \cdot \widehat{X}^n(t) = e \cdot \widehat{N}^n\}$. Then

$$\begin{aligned} 0 &= e \cdot \widehat{X}^n(\tau_\alpha) - e \cdot \widehat{N}^n \\ &\ge e \cdot \widehat{X}^n(\alpha) - e \cdot \widehat{N}^n + e \cdot \widehat{W}^n(\tau_\alpha : \alpha) - 2\varepsilon^{1/3}\bigl(1 + e \cdot \widehat{X}^n(\alpha)\bigr) \\ &\ge C + \tfrac{1}{3} e \cdot \widehat{X}^n(\alpha) \ge C + (1/3)\varepsilon^{-1/3}, \end{aligned}$$

for some constant C, giving the contradiction once again. This concludes the proof of Theorem 2.4.

4. General structures. Theorem 2.4 shows that null-controllability is impossible if the underlying fluid model is throughput optimal. The result is valid for the case $\min\{I, J\} = 2$, and the assumption is crucial for both Theorem 3.4 and Lemma 3.6. How can Theorem 2.4 be extended for general I and J, especially, since it is relatively easy to *numerically check* conditions (3.13)–(3.14) (enough to check separately for each zero path)? We give a partial answer.

DEFINITION 4.1. A path $p \in SP$ is called class-dependent if (3.6)–(3.8)

$$\sum_{j:(i,j)\in\mathcal{E}_p} s(p,i,j)\mu_{ij} = 0, \qquad i \in \mathcal{I}. \tag{4.1}$$

There are only two summands for each given i in (4.1). Basically, the definition says that for each $i \in \mathcal{I}$, belonging to p, and two (just these two!) adjacent activities (i, j_1) and (i, j_2) from the very same path p, we must have $\mu_{i,j_1} = \mu_{i,j_2}$. Similarly, we have the following:

DEFINITION 4.2. A path $p \in SP$ is called pool-dependent if

$$\sum_{i:(i,j)\in\mathcal{E}_p} s(p,i,j)\mu_{ij} = 0, \qquad j \in \mathcal{J}. \tag{4.2}$$

From (3.6)–(3.8), each of the above two types must be a zero path, that is, $\mu(p) = 0$.

THEOREM 4.3. *Let Assumptions 2.1–2.3 hold, and let $I, J \geq 1$. Assume that the fluid model is throughput optimal and satisfies one of the following*:

(1) *has no zero paths, that is, $\mathcal{P}_0 = \varnothing$*;
(2) *each $p \in \mathcal{P}_0$ is either class- or pool-dependent; or, for small $\kappa > 0$,*

$$\sum_{ij} \mu_{ij}(\psi_{ij} - \psi^*_{ij}) < 0 \qquad \textit{whenever } \psi \in \Xi(x^* + m_p\kappa, \nu). \tag{4.3}$$

Then it is impossible to find $T > 0$ and a sequence of SCPs, satisfying (2.9); *that is*, (*weak*) *null controllability is impossible.*

REMARK 4.4. Currently this is as close as we can get to the conclusion that, in the general case, (3.13)–(3.14) prescinds null controllability (for throughput optimal fluid models). Apparently, more work is required when (4.3) results in equality, with path being neither class- nor pool-dependent. We feel, however, that such situations are very rare, maybe even impossible (and may as well contradict to uniqueness of the underlying fluid model; see Assumption 2.2).

REMARK 4.5. Theorem 4.3 trivially implies that null-controllability is also impossible for either one of the following types of the fluid model:

(1) the service rates depend only on the class type (*class-dependent*),

$$\mu_{ij} = \mu_i, \qquad i \in \mathcal{I}, j \in \mathcal{J}; \tag{4.4}$$

(2) the service rates depend only on the station type (*pool-dependent*),

$$\mu_{ij} = \mu_j, \qquad i \in \mathcal{I}, j \in \mathcal{J}. \tag{4.5}$$

Indeed, in both cases the fluid model is throughput optimal, and all paths are either class- or pool-dependent.

PROOF OF THEOREM 4.3. It will be enough to show that relation (3.28) of Lemma 3.6 remains intact, as no other part of the proof of Theorem 2.4 has any structure constraints.

Case 1. Relation (3.28) trivially follows from the current proof of Lemma 3.6.

Case 2. The argument goes exactly as in the proof of Lemma 3.6, until it gets to (3.32), stating that at least one "large" zero path has been activated. Some extra work has to be done at this point.

A. *All zero paths are class-dependent*: in such case [see (3.6)–(3.8)] we have $\|m_p\| = 0$ for each zero path; hence none of these paths has *any* effect on the system. Once again, (3.28) follows trivially.

B. *All zero paths are pool-dependent*: assume that all zero paths satisfy (4.2). We start with the case when there is only one zero path p. The following estimate will be useful. From the exact structure of Φ^n from Theorem 3.3, and (3.10), we have a crude estimate

$$\left|\widehat{X}_i^n(t)\right| \le \varepsilon^{-1/3}\bigl(1 + \bigl(e \cdot \widehat{X}^n(t)\bigr)^+ + e \cdot \widehat{\eta}^n(t)\bigr), \qquad i \notin \mathcal{V}_p, \tag{4.6}$$

since no zero paths have been applied to such classes i. Now, for any feasible allocation $\Psi^n(t)$ consider a unique, *standard*, [1, 2] allocation $\phi^n(t) \in \Xi(X^n(t), N^n)$ that is zero for nonbasic activities and is work conserving: $\min\{(e \cdot \widehat{X}^n(t) - e \cdot \widehat{N}^n), (e \cdot \widehat{N}^n - \sum_{ij} \widehat{\phi}_{ij}^n(t))\} = 0$. By throughput optimality of the fluid model [i.e., $\mathcal{P}_- = \varnothing$; see the definitions before (3.10)], we must have $\sum_{ij} \mu_{ij} \widehat{\Psi}_{ij}^n(t) \le \sum_{ij} \mu_{ij} \widehat{\phi}_{ij}^n(t)$. Using (4.6) and the structure of ϕ^n,

$$\left|\widehat{\phi}_{ij}^n(t)\right| \le \varepsilon^{-1/3}\bigl(1 + \bigl(e \cdot \widehat{X}^n(t)\bigr)^+ + e \cdot \widehat{\eta}^n(t)\bigr), \qquad (i,j) \notin \mathcal{E}_p, \tag{4.7}$$

as well as

$$\left|\widehat{\phi}_{i_0,j_0}^n(t)\right| \le \varepsilon^{-1/3}\bigl(1 + \bigl(e \cdot \widehat{X}^n(t)\bigr)^+ + e \cdot \widehat{\eta}^n(t)\bigr) \tag{4.8}$$

for the leaf (i_0, j_0) of the basic simple path, corresponding to p. Together, (4.7), (4.8) imply that for each station $j \in \mathcal{J} \cap \mathcal{V}_p$, connecting exactly two path edges $(i_1, j) \in \mathcal{E}_p$ and $(i_2, j) \in \mathcal{E}_p$ [this excludes the leaf from (4.8)], we also have

$$\widehat{\phi}_{i_1,j}^n(t) + \widehat{\phi}_{i_2,j}^n(t) \le \varepsilon^{-1/3}\bigl(1 + \bigl(e \cdot \widehat{X}^n(t)\bigr)^+ + e \cdot \widehat{\eta}^n(t)\bigr) \tag{4.9}$$

and due to pool-dependence along the path (4.2) (otherwise it will not hold!), we get

$$\begin{aligned}\mu_{i_1,j}\widehat{\phi}^n_{i_1,j}(t) + \mu_{i_2,j}\widehat{\phi}^n_{i_2,j}(t) &= \mu_{i_1,j}\bigl(\widehat{\phi}^n_{i_1,j}(t) + \widehat{\phi}^n_{i_2,j}(t)\bigr)\\ &\le \mu_{i_1,j}\varepsilon^{-1/3}\bigl(1 + \bigl(e\cdot\widehat{X}^n(t)\bigr)^+ + e\cdot\widehat{\eta}^n(t)\bigr).\end{aligned}\tag{4.10}$$

This proves (3.28). The extension to several pool-dependent paths is straightforward—the only difference being the inclusion of all pool-dependent activities (possibly more than two), connected by station j, into the left-hand side of (4.10). The right-hand side of (4.10) will remain the same.

C. *All zero paths satisfy* (4.3): by linearity, there exists a constant $c > 0$, such that for any feasible set $\{M_p \ge 0, p \in \mathcal{P}_0\}$ and $\psi \in \Xi(x^* + \sum_{p\in\mathcal{P}_0} m_p M_p, \nu)$

$$\sum_{ij}\mu_{ij}\bigl(\psi_{ij} - \psi^*_{ij}\bigr) \le -c\sum_{p\in\mathcal{P}_0} M_p \le -\varepsilon^{1/2}\sum_{p\in\mathcal{P}_0}\|m_p\|M_p.\tag{4.11}$$

Note that the first inequality in (4.11) becomes an equality if and only if each of the M_p is zero. Applying (4.11) to processes from (3.10) and using Theorem 3.3(2), we get for any feasible allocation $\Psi^n(t)$,

$$\begin{aligned}\sum_{ij}\mu_{ij}\widehat{\Psi}^n_{ij}(t) &\le -\varepsilon^{1/2}\|\widehat{\zeta}^n(t)\| + \varepsilon c_F\bigl(\|\widehat{X}^n\|^*_t + \|\widehat{N}^n\|\bigr)\\ &\quad + \varepsilon^{-1/3}\bigl(1 + e\cdot\widehat{\eta}^n(t)\bigr).\end{aligned}\tag{4.12}$$

Using relation (3.31), we continue

$$\begin{aligned}\sum_{ij}\mu_{ij}\widehat{\Psi}^n_{ij}(t) &\le -\varepsilon^{1/2}\|\widehat{\zeta}^n(t)\| + \varepsilon^{5/6}\bigl(1 + \|\widehat{\zeta}^n(t)\| + e\cdot\widehat{\eta}^n(t)\bigr)\\ &\quad + \varepsilon^{-1/3}\bigl(1 + e\cdot\widehat{\eta}^n(t)\bigr) \le \varepsilon^{-1/3}\bigl(1 + e\cdot\widehat{\eta}^n(t)\bigr),\end{aligned}\tag{4.13}$$

where we used $\|\widehat{\zeta}^n(t)\|(-\varepsilon^{1/2} + \varepsilon^{5/6} + \varepsilon^{5/6}\varepsilon^{1/6}) \le 0$ [again, the strict inequality in (4.3) is crucial for the existence of the "$-\varepsilon^{1/2}$" term!]. And (3.28) follows.

D. *Finalizing*: we use B and C to complete the theorem. In particular, once again, the (*work-conserving, no nonbasic activities*) allocation ϕ^n will be introduced. After that, the sum $\sum_{ij}\mu_{ij}\widehat{\phi}^n_{ij}(t)$ will be decomposed into two different sums: one will contain all the terms satisfying (4.7)–(4.10) [this will also include the possible intersections of pool-dependent paths and paths, satisfying (4.3)]; another part of the summation will satisfy (4.13). This completes the proof.

5. Final remarks.

The text is an attempt to understand whether a *given* static fluid model is throughput optimal; and some words need to be said regarding Assumption 2.2, in particular, (a) the treelike structure and (b) the uniqueness of the solution to (2.4) in general.

(a) To start with, null controllability is clearly impossible if the solution to (2.4) does not contain a basic path of the length at least 3 (we need to have at least two stations and two classes to be connected together by basic activities), so some kind of connectivity should be assumed. When a connected component contains cycles fully composed by basic activities, one may have trouble defining weights/directions along simple paths (as was done in Section 3.2), yet one thing will remain true: a throughput optimal model can only have cycles with weight zero. Otherwise, a positive path can become negative if applied in the other direction, and vice versa. This brings us back to the very same zero paths, main ingredients of the current paper.

(b) A mass vector x^*, coming from the solution of (2.4) is a key (!) element due to Assumption 2.3 about the initial condition. This invites a reasonable question. What if there is another optimal solution, with the same vector x^*, but a different graph structure? This is clearly feasible, although both (or, infinite number, in that case) possible static fluid models would still be either all throughput optimal, or sub-optimal altogether, since definition (2.7) does not require any special graph structure.

Now, what if the other solution has a *different* mass vector, say, x^{**}. Is it possible that x^*-solution is throughput optimal, while x^{**} is not? We claim it is not feasible, at least in the case $I = 2$ or $J = 2$, with arguments similar to ones in the proof of Theorem 3.4 (since $e \cdot x^* = e \cdot x^{**}$). The more general structure is still to be resolved....

APPENDIX: SKETCH OF THE PROOF OF THEOREM 3.3

Using the scaling $\widehat{f} = n^{-1/2} f$, introduce auxiliary processes $\widehat{Y}_i^n(t)$, representing the scaled number of class-i customers that are in the queue (and not being served) at time t, and $\widehat{Z}_j^n(t)$—the scaled number of servers at station j that are idle at time t. Clearly, we have the following relations:

$$\widehat{Y}_i^n(t) + \sum_{j \in \mathcal{J}} \widehat{\Psi}_{ij}^n(t) = \widehat{X}_i^n(t), \qquad i \in \mathcal{I}, \tag{A.1}$$

$$\widehat{Z}_j^n(t) + \sum_{i \in \mathcal{I}} \widehat{\Psi}_{ij}^n(t) = \widehat{N}_j^n, \qquad j \in \mathcal{J}. \tag{A.2}$$

The proof can be viewed as generalization of Theorem 2.1 from Atar, Mandelbaum and Shaikhet [5], whose decomposition used only closed simple paths (called cycles). In particular the set $\{\widehat{\Psi}_{ij}^n, (i,j) \in \mathcal{E}_a\}$ was decomposed into basic and nonbasic activities, respectively, $\{\widehat{\Psi}_{ij}^n, (i,j) \in \mathcal{E}_{ba}\}$ and $\{\widehat{\Psi}_{ij}^n, (i,j) \in \mathcal{E}_{ba}^c\}$, turning (3.1) into (see Section 2.3 in [5])

$$\begin{aligned}\widehat{X}_i^n(t) = \widehat{X}_i^n(0) + \widehat{W}_i^n(t) - \sum_{j \in \mathcal{J}} \mu_{ij} \int_0^t G_{ij}\big(\widehat{X}^n(s) - \widehat{Y}^n(s), \widehat{N}^n - \widehat{Z}^n(s)\big)\, ds \\ + \sum_{p: p\text{-closed simple path}} m_{i,p} \int_0^t \widehat{\Psi}_p^n(s)\, ds,\end{aligned}$$

where $\widehat{\Psi}_p^n$ corresponds to a unique nonbasic activity, associated with simple path p and the function G, introduced in [1].

For our purposes, however, that is not enough, since we want to single out all the terms that can cause an abrupt change of $\widehat{X}^n$, and nonwork conservation is exactly what we are looking for, since, a priori we do not have the relation $e \cdot \widehat{Y}^n \wedge e \cdot \widehat{Z}^n = 0$.

The rest of the proof follows the lines of Theorem 2.1 in [5], with an additional requirement to cover direct nonwork conservation, that is, situations when $\widehat{Y}_i^n \wedge \widehat{Z}_j^n > 0$ while $\mu_{ij} > 0$, as well as the open simple paths, corresponding to what we call an indirect nonwork conservation, that is, situations when $\widehat{Y}_i^n \wedge \widehat{Z}_j^n > 0$ while $\mu_{ij} = 0$. We leave out the details.

Acknowledgements. The author wants to thank the anonymous referees of [6] for suggesting the research topic. The interpretation (or misinterpretation) of their question, resulting in the current work, is all mine.

REFERENCES

[1] ATAR, R. (2005). A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Probab.* **15** 820–852. MR2114991

[2] ATAR, R. (2005). Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15** 2606–2650. MR2187306

[3] ATAR, R. and GURVICH, I. (2014). Scheduling parallel servers in the nondegenerate slowdown diffusion regime: Asymptotic optimality results. *Ann. Appl. Probab.* **24** 760–810. MR3178497

[4] ATAR, R., MANDELBAUM, A. and REIMAN, M. I. (2004). Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **14** 1084–1134. MR2071417

[5] ATAR, R., MANDELBAUM, A. and SHAIKHET, G. (2006). Queueing systems with many servers: Null controllability in heavy traffic. *Ann. Appl. Probab.* **16** 1764–1804. MR2288704

[6] ATAR, R. and SHAIKHET, G. (2009). Critically loaded queueing models that are throughput suboptimal. *Ann. Appl. Probab.* **19** 521–555. MR2521878

[7] BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed. Wiley, New York. MR1700749

[8] DIESTEL, R. (2000). *Graph Theory*, 2nd ed. *Graduate Texts in Mathematics* **173**. Springer, New York. MR1743598

[9] HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588. MR0629195

[10] HARRISON, J. M. and LÓPEZ, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Syst.* **33** 339–368. MR1742575

[11] MANDELBAUM, A. and STOLYAR, A. L. (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52** 836–855. MR2104141

[12] STOLYAR, A. L. and TEZCAN, T. (2010). Control of systems with flexible multi-server pools: A shadow routing approach. *Queueing Syst.* **66** 1–51. MR2674107

[13] WILLIAMS, R. J. (2000). On dynamic scheduling of a parallel server system with complete resource pooling. In *Analysis of Communication Networks*: *Call Centres*, *Traffic and Performance* (*Toronto*, *ON*, 1998). *Fields Inst*. *Commun*. **28** 49–71. Amer. Math. Soc., Providence, RI. MR1788708

DEPARTMENT OF MATHEMATICS AND STATISTICS
CARLETON UNIVERSITY
OTTAWA, ONTARIO K1S 5B6
CANADA
E-MAIL: gennady@math.carleton.ca