Control of Fork-Join Networks in Heavy Traffic

Rami Atar
Faculty of Electrical Engineering
Technion
Haifa 32000, Israel
Email: atar@ee.technion.ac.il

Avishai Mandelbaum
Faculty of Industrial Engineering
and Management
Technion
Haifa 32000, Israel
Email: avim@ie.technion.ac.il

Asaf Zviran
Faculty of Industrial Engineering
and Management
Technion
Haifa 32000, Israel
Email: azviran@gmail.com

***Abstract*—A Fork-Join Network (FJN) is a natural model for a queueing system in which customers, or rather tasks associated with customers, are processed both sequentially and in parallel. In this paper we analyze a network that, in addition, accommodates feedback of tasks. An example of a FJN is an assembly operation, where parts are first produced and then assembled to ultimately create a final product. Another example is an emergency department, where a patient "forks" into, say, a blood test and an X-ray, which must then "join" the patient as a prerequisite for a doctor examination. There is a fundamental difference between the dynamics of these two examples: In an assembly network, parts are exchangeable while, in an emergency department, tasks are associated uniquely with patients. They are thus nonexchangeable in the sense that one cannot combine/join tasks associated with different customers.**

In single-server feed-forward FJNs, FCFS processing maintains a fully synchronized flow of tasks. Probabilistic feedback, however, introduces flow disruptions that give rise to task delays and ultimately a decrease in throughput rate. Nevertheless, we show that a simple flow control of tasks can render this decrease of performance asymptotically negligible (though it is not absolutely negligible). More specifically, we analyze a concrete FJN, with nonexchangeable tasks and Markovian feedback, in the conventional heavy-traffic (diffusion) regime. We prove asymptotic equivalence between this network and its corresponding assembly network (exchangeable tasks), thus establishing asymptotic throughput-optimality of our control. The analysis also reveals further interesting properties, such as state-space collapse of synchronization queues.

I. Introduction

There are many examples of processing systems where arriving jobs fork (split) into several tasks, which are then independently processed along parallel routes, and ultimately joined (matched) to create a final product. Indeed, the idea of breaking complex jobs into simpler multiple tasks, which are then performed simultaneously and sequentially by specialized servers, is fundamental. A natural model that captures the complex dynamics thus described is a *Fork-Join* (also called Split-Match) queueing network. Such networks have found applications in a wide variety of domains, for example multi-project environments (Cohen, Mandelbaum and Shtub [7]), parallel communication networks (Hoekstra, Van Der Mei and Bhulai [8]), the justice system (Larson, Cahn and Shell [11]), distributed data-bases (Avi-Itzhak and Halfin [1]) and, finally, health care systems, which have motivated the present study.

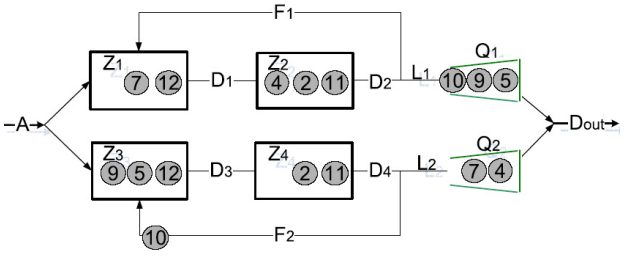

Fig. 1. Fork-Join network with probabilistic feedback

The main model considered in this paper is depicted in Fig. 1. In this network, a sequence of i.i.d. customers (jobs) arrive to the system (process A). Each customer "forks" into two tasks; these tasks are then processed simultaneously and independently along two parallel routes ($Z_1 \to Z_2$ and $Z_3 \to Z_4$), each route consisting of two service stations in tandem; after the second station, some of the tasks are completed (L_i), and the others feed back (F_i) to the first station of their route (Z_1 or Z_3). Each first-completed task joins its *Synchronization Queue* (Q_1 or Q_2), waiting there until its partner task, originating from the *same* customer, is completed. Both tasks are then joined, at which time the corresponding customer departs from the system (D_{out}).

As noted, we are motivated by health care systems, where the fork-join construct is prevalent. We thus assume that tasks are *nonexchangeable*, in the sense that a task associated with a specific customer (e.g. a blood test of a patient) cannot join a task originating in another customer (e.g. an x-ray of another patient). Note that this same Fig. 1 can also model an *assembly* network. Here, the arrival process corresponds to product orders, and service stations manufacture parts from which, ultimately, products are assembled, namely tasks in Q_1 and Q_2 are joined. In such assembly networks, the tasks (parts) are naturally *exchangeable*, which is in contrast to FJNs that arise in health care. We shall refer to the latter simply as *Fork-Join networks*: Each such network, after relaxing its nonexchangeability constraints, turns into a corresponding *assembly network*. This distinction and association between Fork-Join and assembly networks is central for what is to come.

II. On Exchangeability

The feedback in Fig. 1 is random (Markovian). Consequently, the order of the arriving customers (according to process A) is disrupted: It thus differs from the order of departing tasks (in L_i, $i = 1, 2$). This reshuffled order forces tasks (e.g. task 5 in Q_1 and task 7 in Q_2) to wait for their partner, which would not have happened had tasks been exchangeable (5 and 7 would then join and leave as a final product). An (exchangeable) assembly network can thus be characterized by the following **Complementarity Condition**:

$$Q_1(t) \wedge Q_2(t) = 0, \quad \forall t \geq 0; \tag{1}$$

in other words, one of the synchronization queues must be empty at all times (which is clearly not the case in Fig. 1). Our main result (Theorem 1) shows that a simple flow control gives rise to **asymptotic equivalence** between the FJN in Fig. 1 and its corresponding assembly network. More precisely, the above complementarity condition holds *asymptotically*, in (conventional) heavy-traffic, if one always gives preemptive priority to incomplete tasks whose partnering task is already waiting in its synchronization queue. Such a strategy will prove to be asymptotically optimal in the sense of maximum throughput, but *not* optimal.

The technical challenge in the proof of our main result stems from the dependencies caused by the abrupt information exchange between routes: A task that reaches a synchronization queue of one route is immediately changing the priority of its partnering task on the other route. This complex dynamics renders inapplicable the standard method of fluid- followed by diffusion-approximation. Instead, we develope estimates on down-crossing probabilities to deduce tightness of the queue-length processes, which turns out to be sufficient for the purpose of (1). By-products of the proof are some dynamical properties of our FJN under heavy traffic, specifically state-space collapse (to one dimension) of the synchronization queues, and asymptotic equivalence between Fork-Join and assembly networks.

III. Simpler Models

While we do not consider in this paper more general models, it is natural to regard Fig. 1 as a special case of a model with several processing routes, each containing a critically loaded Jackson network. The model that we do solve is the simplest nontrivial representative of this class. The following progression of models and results clarify this (with further details provided in [17]):

- *Single-server feedforward FJNs* [13]: Here a first-come-first-serve (FCFS) discipline at all stations maximizes throughput at all times. Indeed, under such FCFS, the order by which customers arrive is the exact order by which their corresponding tasks are processed, hence the Complementarity Condition (1) must prevail.
- Single-server Fork-Join queues with *feedback*, as in Fig. 1, but with a *single* station per route: An exhaustive-service control serves a task until it reaches a synchronization queue, thus maximizing throughput as it reduces the model to the above feedforward case.
- *Multi-server feedforward FJNs* [17]: In this case, FCFS is only asymptotically optimal. Specifically, task-ordering is disrupted by the parallel processing of many-server stations, but this disruption is proved to be negligible in conventional heavy traffic.

IV. Our Fork-Join Network

All our stochastic processes are defined on a common given probability space. We first introduce the building blocks for external arrivals, service times, departures and feedback, followed by the corresponding processes. Let a sequence of i.i.d. random variables, $\{u(m),\ m \geq 1\}$, be given, where $u(m)$ are strictly positive with unit mean; set $u(0) = 0$. Denote by λ the average arrival rate. Then the *external-arrivals* process is $A = \{A(t),\ t \geq 0\}$, where

$$A(t) \equiv max\{k :\ \textstyle\sum_{m=0}^{k} \lambda^{-1} u(m) \leq t\},\ t \geq 0.$$

We also have four unit-rate Poisson processes $S_j = \{S_j(t),\ t \geq 0\}$, $j = 1, \ldots, 4$, and corresponding service-rates μ_j, where each pair (S_j, μ_j) is associated with a station j in Fig. 1. The *departure process* for station j, $D_j = \{D_j(t),\ t \geq 0\}$, is given by

$$\begin{cases} D_j(t) = S_j(\mu_j B_j(t)); \\ \\ B_j(t) = \int_0^t \mathbb{I}_{\{Z_j(s)>0\}} ds; \\ \\ I_j(t) = t - B_j(t) = \int_0^t \mathbb{I}_{\{Z_j(s)=0\}} ds; \end{cases}$$

here B_j and I_j are, respectively, the Busyness and Idleness processes associated with station j, and each Z_j, to be formally introduced momentarily, is the number of tasks in station j, either served or waiting in the *resource* queue preceding server j.

Remark on Work-Conservation: Following standard terminology, a control is *work-conserving* if it does not idle servers that are faced with waiting customers. Formally, work-conservation is precisely the above defining relation of the B_j's in terms of Z_j's. We are thus restricting attention to work-conserving controls which, in fact, turns out to be without loss of generality. Indeed, as will be formalized in the sequel, we shall be concerned with maximizing system output (the processes D_j's and D_{out}). Turning a control into work-conserving can only increase its "Busyness" process and hence its output. □

Next, consider two sequences of i.i.d. indicators $\xi^i = \{\xi^i_k,\ k \in \mathbb{N}\}$, $i = 1, 2$. Each r.v. ξ^i_k is $\{0, 1\}$-valued, and is equal to 1 to indicate that task k at route i is fed back to re-initiate the service process, after completing service at route i. The probability of feedback on route i is given by p_i, $i = 1, 2$. The *feedback* building block for route i, F_i, is then defined as $F_i(d) = \sum_{k=1}^{d} \xi^i_k$, $d = 0, 1, \ldots$, with $F_i(0) = 0$.

Note that the customer population is assumed to be homogeneous, in the sense that all customers have the same precedence constraints, interarrival time distributions, service

time distributions and feedback probability. It is further assumed that all building blocks A, S_j, F_i are mutually independent.

Finally, we state basic relations among the stochastic processes. These are given, for $t \geq 0$, by

$$\begin{cases} Z_1(t) = A(t) - D_1(t) + F_1(D_2(t)); \\ Z_2(t) = D_1(t) - D_2(t); \\ Z_3(t) = A(t) - D_3(t) + F_2(D_4(t)); \\ Z_4(t) = D_3(t) - D_4(t); \\ L_1(t) = D_2(t) - F_1(D_2(t)); \\ L_2(t) = D_4(t) - F_2(D_4(t)); \\ Q_i(t) = L_i(t) - D_{out}(t),\ i = 1, 2; \\ N(t) = A(t) - D_{out}(t). \end{cases} \tag{2}$$

The interpretations of the various processes in (2) are as follows:

- $D_{out}(t)$ - Throughput process: cumulative number of departures, namely customers that completed their services, up until time t;
- $L_i(t)$ - Route departure process: cumulative number of departures, namely tasks that completed their services on route i, till time t;
- $D_j(t)$ - Station departure process: cumulative number of departures, namely tasks that completed their processing at station j, till time t;
- $Z_j(t)$ - Number of customers in station j, either served or waiting for service at the *resource queue* of station j, at time t;
- $Q_i(t)$ - Number of customers in the *synchronization queue* at the end of route i, at time t;
- $N(t)$ - Total number of customers within the network, at time t; (Note that the number of tasks in both routes at all times is identical, hence equals $N(t)$.)

V. Related Work

Exact analysis of Fork-Join models is extremely hard, hence their analysis has traditionally focused on bounds and approximations. Baccelli and Makowski [2] and Baccelli, Makowski and Shwartz [3] obtained bounds via stochastic ordering (association of random variables). Squillante et al. [16] deployed approximate matrix-analytic methods, for the analysis of general parallel-server Fork-Join queues with dynamic task scheduling. Boxma, Koole and Liu [5] reviewed various solution methods for models of parallel and distributed systems. And lastly, most relevant to the present work, Nguyen [13] analyzed single-server feedforward FCFS FJNs, working within the framework of *conventional heavy traffic* and deriving Brownian approximations. Since [13], to the best of our knowledge, there as been little if any research progress on fork-join control in heavy-traffic. An explanation can be found in [14], specifically in its title and the fact that the paper is restricted to FCFS. (For example, redoing [14] with static priorities would turn the model tractable - see Section XI). Our hope is that the present paper will change this state of affairs. Specifically, our paper continues [13], and it is based on [17]. We analyze the FJN in Fig. 1, assuming nonexchangeable tasks. This is more general than the models in [13] in that it allows *feedback*. The latter causes throughput degradation, which we overcome asymptotically in heavy traffic.

VI. Heavy Traffic

Our FJN will be analyzed in *Heavy Traffic*, as we now define precisely. Consider a sequence of FJNs, each as in Fig. 1, which are indexed by n. The following relations are assumed to hold, as $n \uparrow \infty$:

- Arrival rates: $\lambda^n = \lambda \cdot n + \hat{\lambda} \cdot \sqrt{n} + o(\sqrt{n})$;
- Service rates: $\mu_j^n = \mu_j \cdot n + \hat{\mu}_j \cdot \sqrt{n} + o(\sqrt{n})$;
- **Heavy Traffic**: $\lim_{n\to\infty} n^{\frac{1}{2}}(\rho_j^n - 1) = \theta_j$, where ρ_j^n is the *traffic intensity* of station j.

These traffic intensities of the stations are given by:
$\rho_1^n = \frac{\lambda^n}{\mu_1^n \cdot (1-p_1)}$, $\rho_2^n = \frac{\lambda^n}{\mu_2^n \cdot (1-p_1)}$,
$\rho_3^n = \frac{\lambda^n}{\mu_3^n \cdot (1-p_2)}$ and $\rho_4^n = \frac{\lambda^n}{\mu_4^n \cdot (1-p_2)}$.
Note that $\frac{1}{1-p_i}$ is the average number of times that a task "visits" route i (following a Geometric distribution with success probability $1 - p_i$, $i \in \{1, 2\}$).
In the above, the following scalars are apriori given and assumed finite: $\lambda > 0$, $\mu_j > 0$, $\hat{\lambda}$, $\hat{\mu}_j \in (-\infty, \infty)$, $p_i \geq 0$, $\theta_j \leq 0$. A simple sufficient condition for Heavy Traffic is

$$\lambda = \mu_1 \cdot (1 - p_1) = \mu_3 \cdot (1 - p_2);\ \mu_2 = \mu_1,\ \mu_4 = \mu_3.$$

The following notation for scaled (and possibly centered) stochastic processes is standard: resource-queue length $\hat{Z}_j^n(t) = \frac{Z_j^n(t)}{\sqrt{n}}$, synchronization-queue length $\hat{Q}_i^n(t) = \frac{Q_i^n(t)}{\sqrt{n}}$, potential service $\hat{S}_j^n(t) = \frac{S_j^n(\mu_j^n t) - \mu_j^n t}{\sqrt{n}}$ and idleness process $\hat{I}_j^n(t) = \frac{I_j^n(t)}{\sqrt{n}}$, all defined for $t \geq 0$.

VII. Optimal Control

As already discussed, the order of customer departures becomes unsynchronized due to the random feedback at the end of each route. This phenomena causes a delay in the join process and hence reduces throughput. Heuristically, the optimal performance (maximal throughput) is that of a corresponding assembly network, with exchangeable tasks. Thus, one could attempt to characterize optimal performance by the *Complementarity Condition* (1). We shall now rigorously formulate our control problem, and demonstrate that (1) is indeed a sufficient condition for optimal performance, in the sense of maximum throughput.

A. Optimality Criteria

The identity of the job being processed at time t by server j, for every t and j, is regarded as the *control process*. We use α to denote a generic control processes. A rigorous definition requires significant additional notation, which we prefer to avoid and thus settle for the above verbal description. Similarly, we do not provide a complete definition of the term state process, but only a verbal description. The *state process* is defined as the information on the identity of all jobs at each station at each time. Note that this does not include

information on which jobs are being *processed* at each time, and so thus the control, in general, cannot be reconstructed from the state. We will regard a control *admissible* if it is adapted to the filtration of the state process. Note, in particular, that such a control is nonanticipating, in the sense that it does not obtain information from future events.

Denote by $\mathbb{A}$ the set of admissible controls.

Definition 1. *Optimality is defined in terms of maximal achievable number of departures over a finite time-horizon, namely maximal throughput. More precisely,*

- Exact Optimality*: Control $\gamma \in \mathbb{A}$ is optimal if, for all $T > 0$, γ attains $esssup_{\alpha\in\mathbb{A}}(D^{\alpha}_{out}(T))$. (Here, and in the sequel, a control is appended as a superscript of a process (e.g. D^{α}_{out}) to indicate that this process evolves under that control.)*
- Asymptotic Optimality*: Control $\gamma \in \mathbb{A}$ is asymptotically optimal if for any other control $\alpha \in \mathbb{A}$ and for all $T > 0$,*

$$\hat{D}^{n,\gamma}_{out}(T) \geq \hat{D}^{n,\alpha}_{out}(T) - \epsilon_n(T), \text{ with } \epsilon_n(T) \to 0,$$

where the convergence of $\epsilon_n(\cdot)$ is uniformly on compacts (u.o.c.), in probability. (Here, and in the sequel, a superscript n of a stochastic process (e.g. $\hat{D}^{n,\gamma}_{out}$) indicates that this process arises from the n^{th} network in the heavy-traffic sequence.)

Proposition 1. *Each of the following conditions implies its corresponding Definition 1:*

- Exact Optimality*: $Q_1(T) \wedge Q_2(T) = 0, \ \forall T > 0$;*
- Asymptotic Optimality*: $\hat{Q}^n_1(\cdot) \wedge \hat{Q}^n_2(\cdot) \xrightarrow{p}_n 0$, where $\xrightarrow{p}_n$ denotes convergence u.o.c., in probability.*

Proof of Proposition 1: The proof is a direct consequence of three sample-path properties of our system. These will now be formulated and their proofs outlined. (For the full details, here and in the sequel, readers are referred to [17].)

Property 1: Restricting to work-conserving controls, the processes Z_j, D_j and L_i do not depend on the control, for all routes i and stations j. This independence is due to servers time being server dependent, as opposed to customer dependent. (Alternatively, only the following are control-dependent processes: Q_i, D_{out} and N; this dependence is due to the departure synchronization constraints associated with nonexchangeable tasks.)

Property 2: $argmax_{\alpha\in\mathbb{A}}(D^{\alpha}_{out}(T)) =$
$= argmin_{\alpha\in\mathbb{A}}[\sum_{j=1}^{4} Z_j(T) + Q^{\alpha}_1(T) + Q^{\alpha}_2(T)]$

Property 3: $Q^{\alpha}_1(T) + Q^{\alpha}_2(T) =$

$$= |L_1(T) - L_2(T)| + 2 \cdot Q^{\alpha}_1(T) \wedge Q^{\alpha}_2(T), \ \forall \alpha \in \mathbb{A}. \quad (3)$$

Proof of Property 1: By tracking sample-path evolutions (according to our system equations), it does indeed follow that as long as there is no forced idleness of servers, task-counts (Z_j's) and flows prior to the synchronization queues (A, D_j, L_i and $F_1(D_2), F_2(D_4)$) are control-independent. The network output (D_{out}), on the other hand, generally does depend on the control, which affects the Q_i's and N as well. □

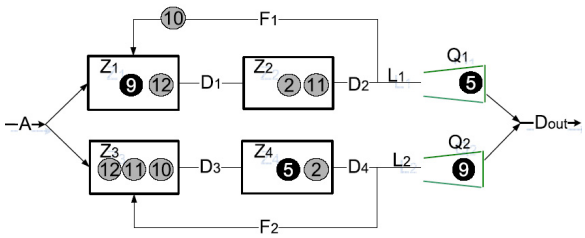

Fig. 2. Customer classes under the γ-control

Proof of Property 2: The total number of customers within each route is the same at each given time, since a customer joins and departs all routes simultaneously. It follows that $2 \cdot N^{\alpha}(T) = \sum_{j=1}^{4} Z_j(T) + Q^{\alpha}_1(T) + Q^{\alpha}_2(T)$, for any control α. Now, the external arrival process A is primitive and thus uncontrollable. Recalling the definition of D_{out} in (2) now yields Property 2. □

Proof of Property 3: We omit the α for notational convenience. Using the relation $N(T) = A(T) - L_1(T) + Q_1(T) = A(T) - L_2(T) + Q_2(T)$, we deduce that $|Q_1(T) - Q_2(T)| = |L_1(T) - L_2(T)|$, for all $T \geq 0$. However, $Q_1(T) + Q_2(T) = Q_1(T) \vee Q_2(T) - Q_1(T) \wedge Q_2(T) + 2 \cdot Q_1(T) \wedge Q_2(T)$. Therefore, for all $T \geq 0$,

$$Q_1(T) + Q_2(T) = |L_1(T) - L_2(T)| + 2 \cdot Q_1(T) \wedge Q_2(T).$$

□

In view of the above three properties, the Exact Optimality condition in Proposition 1 implies Definition 1. We now continue with asymptotic optimality. For any control $\alpha \in \mathbb{A}$ and $T > 0$, one has $Q^{\alpha}_1(T) + Q^{\alpha}_2(T) \geq |L_1(T) - L_2(T)|$. Hence, $Q^{\gamma}_1(T) + Q^{\gamma}_2(T) \leq Q^{\alpha}_1(T) + Q^{\alpha}_2(T) + 2 \cdot Q^{\gamma}_1(T) \wedge Q^{\gamma}_2(T)$, for any α, γ. Taking γ in Proposition 1, and letting $\epsilon_n(\cdot) = \hat{Q}^{n,\gamma}_1(\cdot) \wedge \hat{Q}^{n,\gamma}_2(\cdot) \xrightarrow{p}_n 0$, one deduces that

$$\hat{Q}^{n,\gamma}_1(T) + \hat{Q}^{n,\gamma}_2(T) \leq \hat{Q}^{n,\alpha}_1(T) + \hat{Q}^{n,\alpha}_2(T) + \epsilon_n(T),$$

for all $T > 0$ and any $\alpha \in \mathbb{A}$. Property 2 now enables one to translate this last inequality for Q's to an inequality for D's, as in Definition 1. □

B. Control Policy

The exact control problem for the model in Fig. 1 seems intractable. We now propose a control that, while not optimal, will be proved asymptotically optimal.

Proposed Control (referred to as Cronyism- or γ-control): Within each route, assign preemptive priority to tasks of customers whose service was completed in the other route. (Preemptive priority entails interrupting and resuming a task at a later time.)

The Cronyism-control creates a natural division of all customers into two *classes*:

- LP (Low Priority) Customers: Customers whose service is still incomplete in *both* routes; e.g., gray customers in Fig. 2.

- HP (High Priority) Customers: Customers whose service is completed in one of the routes but is still incomplete in the other; e.g., black customers in Fig. 2.

Finally, assume FCFS within each class, which now fully characterizes the control. (The control is adaptive in the sense that decisions depended solely on immediate system state.) Note that the γ-control requires information exchange between routes, which creates dependencies between routes. This dependency, on one hand, is the reason for asymptotic optimality but, on the other, is the main technical challenge in establishing it.

In the sequel, we use the following notation for a generic process G_j: $G_j^H(G_j^L)$ is the process associated with the High-Priority (Low-Priority) tasks, respectively, and G_j^T is associated with all customers. We then have

$$Z_{1,2}^H = Q_2 \quad \text{and} \quad Z_{3,4}^H = Q_1, \tag{4}$$

where $Z_{1,2}^H = Z_1^H + Z_2^H$ and $Z_{3,4}^H = Z_3^H + Z_4^H$. Hence, minimizing synchronization queues is equivalent to minimizing the number of tasks within the resource queues that are associated with HP customers.

Denote by $A_j^H = \{A_j^H(t),\, t \geq 0\}$ the "Birth" ("arrival") process of HP customers in Station j. For example, assume a departure of a task from Route 2, associated with an LP customer; then this departure causes a priority change (to HP) of that customer, as well as a count increase in A_j^H, if j is the station on Route 1 where the partner task is then present.

VIII. Asymptotic Optimality

The following is our main result. Its proof is outlined in Section X.

Theorem 1 (Asymptotic Optimality). *For any fixed interval* $[0, T]$ *and any* $\epsilon > 0$,

$$\mathrm{P}(max_{t\in[0,T]}[\hat{Z}_{1,2}^{n,H}(t) \wedge \hat{Z}_{3,4}^{n,H}(t)] > \epsilon) \longrightarrow_n 0, \tag{5}$$

where $\hat{Z}_{1,2}^{n,H} = \hat{Z}_1^{n,H} + \hat{Z}_2^{n,H}$, *and* $\hat{Z}_{3,4}^{n,H} = \hat{Z}_3^{n,H} + \hat{Z}_4^{n,H}$. □

From (4), we now conclude that $\hat{Q}_1^n(\cdot) \wedge \hat{Q}_2^n(\cdot) \xrightarrow{p}_n 0$.

A central part of the proof is to establish tight estimates on the number of HP customer at the various stations. The challenge stems from the dynamics of HP customers being coupled with that of LP customers. Specifically, the departure of a task associated with an LP customer at a given route, joining the corresponding synchronization queue, triggers a priority change of that customer (its task in the other route), from LP to HP. The dynamics of LP customers, in turn, is constrained by the presence of HP customers. These HP "Birth" processes (A_j^H) are far from standard models of arrival processes: their precise analysis would entail tracking, for every individual customer in the system, the precise station where each of its tasks is located, which would give rise to an intractable state-description. Instead, rather than making an attempt to characterize A_j^H, our approach is to develop estimates (Lemma 3) that are *uniform* over all birth processes with intensity that is not too large. Since these HP births are caused by departures of LP customers from the other route, showing that the birth intensity is indeed not too large amounts to bounding the intensity of LP departure (Lemma 2).

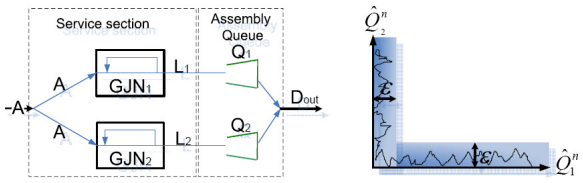

Fig. 3. System Dynamics in Heavy Traffic

Asymptotically Optimal but Not Optimal. Our γ-control will be proved asymptotically optimal, but it is not optimal. To see that, consider the following γ'-*control*: In each route, assign preemptive priority to tasks of customers whose service was completed in the other route, as before, but also to tasks of customers whose service was initiated in the departure station (Station 2 or Station 4) in the other route; again, assume FCFS within each priority class. Consider now the realization in Fig. 2, for this extended γ'-control: the task to be scheduled for service in Station 3 (bottom-left station) is associated with customer 11; in contrast, the previous γ-control would have served Customer 10, adhering to FCFS of LP customers. We now explain that

$$\exists T \;\; s.t. \;\; \mathrm{P}(D_{out}^{\gamma'}(T) > D_{out}^{\gamma}(T)) > 0,$$

assuming, without loss, the scenario in Fig. 2 at some time T_0.

Denote by T_l^j the departure time of the l-th customer from station j ($l = 1, 2, \ldots,$). Consider the event $\{T_1^3 < T_1^2 < T_1^4 < T_2^4 < T_2^3 < T_3^4\}$; for this event, and for the above departures from Stations 2 and 4, we further assume that these departures directly join the corresponding synchronization queue (no feedbacks). Thus, under the γ'-control, one has $D_{out}^{\gamma'}(T_2^4) - D_{out}^{\gamma'}(T_0) = 2$ (departures of Customers 5 and 11); this is in contrast to the γ-control, where $D_{out}^{\gamma}(T_3^4-) - D_{out}^{\gamma}(T_0) = 1$ (departure of Customer 5 only). The above event has a positive probability, hence there must be a deterministic time T for which $\mathrm{P}(T \in [T_2^4, T_3^4],\, D_{out}^{\gamma'}(T) > D_{out}^{\gamma}(T)) > 0$. □

IX. System Dynamics in Heavy Traffic

Theorem 1 and Properties 1-3 reveal asymptotic equivalence between our FJN under γ-control and a corresponding assembly network: same topology, arrivals and services; exchangeable tasks with FCFS control. To see that, note the following relations for an assembly network, at all $T > 0$, each of which reflects exchangeability in the assembly dynamics: $Q_1(T) \wedge Q_2(T) \equiv 0$, $Q_1(T) \vee Q_2(T) \equiv |L_1(T) - L_2(T)|$, $D_{out}(T) \equiv L_1(T) \wedge L_2(T)$ (the latter following from the general relation $D_{out}(T) \equiv L_1(T) \vee L_2(T) - Q_1(T) \vee Q_2(T)$). The equivalence alluded to amounts to having these last relations hold also for the FJN with non-exchangeable

tasks, though only asymptotically after rescaling, as we now explain.

State-space Collapse of Synchronization Queues. The relation $\hat{Q}_1^n \wedge \hat{Q}_2^n \overset{p}{\approx} 0$ (Theorem 1) implies that the 2-dimensional stochastic process $\hat{Q}_1^n, \hat{Q}_2^n$ collapses to 1-dimension, being restricted (with high probability) to an ϵ-environment of the axes ($\epsilon > 0$ arbitrarily small); see Fig. 3, the right graph.

Throughput Equivalence, or **Asymptotic Exchangeability**. For the two networks (fork-join under γ-control and assembly under FCFS control), the processes Z_j, D_j and L_i have identical sample paths, for all routes i and stations j. To see that, first consider both networks with FCFS control, in which case the considered sample paths are clearly equal. Then, according to Property 1, the processes Z_j, D_j and L_i do not depend on the control.

Theorem 1 is one way of expressing asymptotic exchangeability under γ-control. Together with Property 3, it also implies that $\hat{Q}_1^n(\cdot) \vee \hat{Q}_2^n(\cdot) \overset{p}{\approx} \frac{1}{\sqrt{n}}|L_1^n(\cdot) - L_2^n(\cdot)|$, from which we deduce: $D_{out}^{n,\gamma}(\cdot) \equiv L_1^n(\cdot) \vee L_2^n(\cdot) - Q_1^{n,\gamma}(\cdot) \vee Q_2^{n,\gamma}(\cdot) \overset{p}{\approx} L_1^n(\cdot) \wedge L_2^n(\cdot)$.

The significance of asymptotic exchangeability is that, in heavy-traffic, applying γ-control to our FJN yields a throughput process D_{out} that has approximately the same distribution as that of an assembly network under FCFS control. The latter is the minimum of the L_i's, each of which is the exogenous output process of a 2-station Jackson network (with feedback); as such, each L_i is a Poisson process with rate λ, though the L_i's are dependent (emanating from the same exogenous input A). The distribution of $D_{out} = L_1 \wedge L_2$ is thus tractable, in principle, following from the joint distribution of exogenous output processes from a Generalized Jackson Network [6].

X. Proof Outline for Theorem 1

We now outline the proof of Theorem 1. (More details appear in [17].) Fix any interval $[0, T]$ and $\epsilon > 0$.

Proof of Theorem 1: Introduce
$E_{n,T} = \{max_{t\in[0,T]}\{Z_{1,2}^{n,H}(t) \wedge Z_{3,4}^{n,H}(t)\} > \epsilon\sqrt{n}\}$.
Then define

$$\sigma = inf\{t \ : \ Z_{1,2}^{n,H}(t) \wedge Z_{3,4}^{n,H}(t) > \epsilon\sqrt{n}\};$$

$$\tau = sup\{t < \sigma \ : \ Z_{1,2}^{n,H}(t) \wedge Z_{3,4}^{n,H}(t) \leq \tfrac{\epsilon}{3}\sqrt{n}\}.$$

Now let $E_n = E_{n,T} \cap \{Z_{3,4}^{n,H}(\tau) \leq Z_{1,2}^{n,H}(\tau)\}$. Then on E_n and during the time-interval (τ, σ) (Fig. 4):

- Both processing routes contain more than $\frac{\epsilon}{3}\sqrt{n}$ HP customers;
- The number of HP customers in $Z_{3,4}^{n,H}$ increases by more than $\frac{\epsilon}{2}\sqrt{n}$.

We prove Theorem 1 by showing that $\mathrm{P}(E_n) \longrightarrow_n 0$. The proof that $\mathrm{P}(\tilde{E}_n) \longrightarrow_n 0$, where $\tilde{E}_n = E_{n,T} \cap \{Z_{3,4}^{n,H}(\tau) > Z_{1,2}^{n,H}(\tau)\}$, is completely analogous. The proof of the former is based on the following three lemmas; their proofs are given subsequently.

For any process, say X, and random times $a \leq b$, we write $X[a, b]$ for $X(a) - X(b)$.

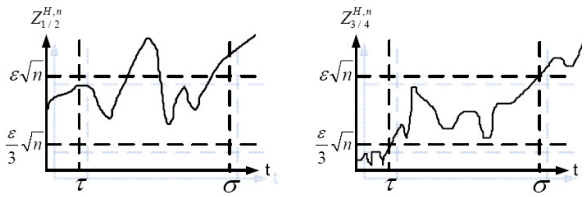

Fig. 4. Example of possible sample-path for event En

Lemma 1. *(*Bounding HP idleness*): Fix $\delta \in (0, 1/4)$. Then*

$$\begin{aligned} &\mathrm{P}(I_2^{n,H}[\tau, \sigma] > n^{-\frac{1}{2}+\delta}) \longrightarrow_n 0; \\ &\mathrm{P}(I_4^{n,H}[\tau, \sigma] > n^{-\frac{1}{2}+\delta}) \longrightarrow_n 0. \end{aligned} \tag{6}$$

Lemma 2. *(*Bounding number of LP departures*): Fix $\delta \in (0, 1/4)$. Then*

$$\mathrm{P}(D_2^{n,L}(\sigma) - D_2^{n,L}(\tau) > n^{\frac{1}{2}+\delta}) \longrightarrow_n 0. \tag{7}$$

Lemma 3. *(*Bounding $|\sigma - \tau|$*): Fix $\delta \in (0, 1/4)$. Then*

$$\mathrm{P}(|\sigma - \tau| < n^{-\delta}, \ \ A_{3,4}^{n,H}[\tau, \sigma] \geq \frac{\epsilon}{2} \cdot \sqrt{n}) \longrightarrow_n 0. \tag{8}$$

Now consider the event
$H_n = \{\exists \sigma, \tau \in [0, T] \text{ s.t. } A_{3,4}^{n,H}[\tau, \sigma] \leq n^{\delta+\frac{1}{2}}, \ B_4^{n,H}[\tau, \sigma] \geq \frac{n^{-\delta}}{2}, \ Z_{3,4}^{n,H}[\tau, \sigma] > \frac{\epsilon}{2}\sqrt{n}\}$.
Using Lemmas 1-3, system's equations (Section IV) and some computation, one can verify that

$$\mathrm{P}(E_n) \leq \mathrm{P}(H_n) + \alpha_n, \ \ \alpha_n \longrightarrow_n 0.$$

Hence it is enough to prove that $\mathrm{P}(H_n) \longrightarrow_n 0$. To this end, divide $[0, nT]$ into K intervals with fixed interval-length $|J_k| \approx n^{1-\delta}$. On the event H_n there is at least one interval on which $\Delta_k S_4 < n^{\delta+\frac{1}{2}}$, where S_4 is the unit-rate Poisson process defined above (Section IV). It follows that
$\mathrm{P}(H_n) \leq \mathrm{P}(\exists k \in \{1, \ldots, K\} \text{ s.t. } \Delta_k S < n^{\delta+\frac{1}{2}}) \longrightarrow_n 0$, where $\Delta_k S$ is the increment of S over the interval J_k. This completes the proof of Theorem 1. ■

Proof of Lemma 1: We shall prove the claim for $I_2^{n,H}[\tau, \sigma]$; the proof for $I_4^{n,H}[\tau, \sigma]$ is similar.
For a fixed $\delta \in (0, 1/4)$, define the event $\hat{E}_n = E_n \cap \{I_2^{n,H}[\tau, \sigma] > n^{-\frac{1}{2}+\delta}\}$. Recall that, for any $t > 0$, $Z_2^{n,H}(t) = Z_2^{n,H}(0) + A_2^{n,H}(t) + S_1^H(\mu_1^n B_1^{n,H}(t)) - S_2^H(\mu_2^n B_2^{n,H}(t))$, where $A_2^{n,H}$ denotes the birth process of HP tasks in Station 2. Note that $A_2^{n,H}$ is a positive increasing process. Scale by $\sqrt{n}$ and define

$$\hat{X}_2^n(t) = \hat{S}_1^{n,H}(B_1^H(t)) - \hat{S}_2^{n,H}(B_2^H(t)) + (\hat{\mu}_1 - \hat{\mu}_2) \cdot t. \tag{9}$$

Then the following relation holds on the event E_n:

$$\begin{aligned} &\hat{Z}_2^{n,H}(t) = \hat{Z}_2^{n,H}(0) + \text{ increasing process } + \hat{X}_2^n(t) + \\ &\qquad \mu_2^n \cdot \hat{I}_2^{n,H}(t) - \mu_1^n \cdot \hat{I}_1^{n,H}(t); \\ &\begin{cases} \int_0^t \mathbb{I}_{\{\hat{Z}_2^{n,H}(s) > 0\}} d\hat{I}_2^{n,H} = 0; \\ \int_0^t \mathbb{I}_{\{\hat{Z}_2^{n,H}(s) < \frac{\epsilon}{4}\}} d\hat{I}_1^{n,H} = 0; \end{cases} \end{aligned} \tag{10}$$

Hence the measures induced by the increasing processes $\hat{I}_1^{n,H}, \hat{I}_2^{n,H}$ do not charge the set of times t where $\hat{Z}_2^{n,H}(t) \in$

$(0, \frac{\epsilon}{4})$. Define the following random times, e.g., Fig. 5 (where the infimum over the empty set is $+\infty$).

$$\begin{cases} A_1 = \inf\{\tau \le t \le \sigma : \hat{Z}_2^{n,H}(s) = 0\}; \\ B_1 = \inf\{A_1 < t \le \sigma : \hat{Z}_2^{n,H}(s) \ge \frac{\epsilon}{4}\}; \\ \text{continue in an inductive manner, for } i = 1, 2, \ldots : \\ A_{i+1} = \inf\{B_i < t \le \sigma : \hat{Z}_2^{n,H}(s) = 0\}; \\ B_{i+1} = \inf\{A_{i+1} < t \le \sigma : \hat{Z}_2^{n,H}(s) \ge \frac{\epsilon}{4}\}. \end{cases}$$

For every interval $[B_i, A_{i+1})$ contained in (τ, σ), define $C_i = \sup\{t \in [B_i, A_{i+1}) : \hat{Z}_2^{n,H}(s) \ge \frac{\epsilon}{4}\}$. By the definitions above, one sees that, on the intervals $[C_i, A_{i+1})$, $\hat{Z}_2^{n,H}$ starts at $\frac{\epsilon}{4}$ and ends at zero without exiting $(0, \frac{\epsilon}{4})$. We shall refer to $[C_i, A_{i+1})$, for which $A_{i+1} < \infty$, as ***Down Crossing*** intervals.

Claim. Denote by $R^n[\tau, \sigma]$ the number of *down crossings* on $[\tau, \sigma]$. Then $R^n[\tau, \sigma] 1_{\hat{E}_n}$ are tight r.v.s.

Proof of Claim. Define $H_K = \hat{E}_n \cap \{R^n[\tau, \sigma] > K\}$. Using (10), positivity of the arrival process and the fact that $[\tau, \sigma] \subseteq [0, T]$, one can verify that

$$H_K \subseteq \{\exists 0 \le s_1 \le t_1 \le s_2 \le \ldots \le t_K \le T \text{ s.t. } |\hat{X}_2^n[s_i, t_i]| \ge \tfrac{\epsilon}{4}, \ \forall i \in \{1, \ldots, K\}\}.$$

Thus

$$\mathrm{P}(H_K) \le \mathrm{P}(\exists i \text{ s.t. } |\hat{X}_2^n[s_i, t_i]| \ge \tfrac{\epsilon}{4}, \ 0 \le t_i - s_i \le \tfrac{T}{K}) \le \mathrm{P}(mod_T(\hat{X}_2^n, \tfrac{T}{K}) \ge \tfrac{\epsilon}{4}).$$

Here $mod_T(X, \delta) = \sup_{0 \le s \le t \le T, t-s \le \delta} |X(t) - X(s)|$.

The processes $\hat{S}_j^{n,H}$ are centered, scaled Poisson processes, which converge weakly to a *Brownian Motion* (BM) process. In particular, they are C-tight, that is, tight in the Skorohod J_1 topology, and having a.s. continuous sample paths for every subsequential limit. Since B_j^H have sample paths that are Lipschitz with constant 1, the processes $\hat{X}_2^n$ are also C-tight. Thus $\forall \eta > 0 \ \exists K \in \mathbb{N}$ s.t. $\mathrm{P}(H_K) \le \eta$, as follows from Proposition VI.3.26 of [10], which characterizes C-tightness. This proves the claim. □

We now return to the proof of Lemma 1. Given η let K be so large that $P(H_K) < \eta/2$. Let us analyze the event $\hat{E}_n \cap H_K^c$. On this event one has less than $K + 1$ intervals of $[A_i, B_i)$. Note that $\hat{I}_2^{n,H}$ does not increase outside these intervals. Therefore, there exists an interval j on which $\hat{I}_2^{n,H}[A_j, B_j)$ increases by $\frac{n^{\delta-1}}{K+1}$. Using (10) and positivity of the arrival process, on this event one must have $|\hat{X}_2^n[A_j, B_j)| > |\frac{n^\delta}{K'}|$, for some constant K' (that depends on K). By the C-tightness of the processes $\hat{X}^n$, the probability for such an event converges to 0, as $n \longrightarrow \infty$.

As a result, for all large n, $\mathrm{P}(\hat{E}_n) \le \eta$; since η is arbitrary, we obtain $\mathrm{P}(\hat{E}_n) \longrightarrow_n 0$.

This completes the proof of Lemma 1. ■

Proof of Lemma 2: Fixing $\delta \in (0, 1/4)$, consider $H_n = \{D_2^{n,L}(\sigma) - D_2^{n,L}(\tau) \ge n^{\frac{1}{2}+\delta}\}$. Define

$$\alpha = \inf\{\tau \le t \le \sigma \ : \ D_2^{n,L}(t) - D_2^{n,L}(\tau) \ge \tfrac{1}{3} \cdot n^{\frac{1}{2}+\delta}\};$$
$$\beta = \inf\{\tau \le t \le \sigma \ : \ D_2^{n,L}(t) - D_2^{n,L}(\tau) \ge n^{\frac{1}{2}+\delta}\}.$$

On H_n α, β are finite. With $\delta' = \frac{\delta}{2}$,

$\mathrm{P}(H_n) = \mathrm{P}(H_n, I_2^{n,H}[\alpha, \beta] > n^{\delta' - \frac{1}{2}}) + \mathrm{P}(H_n, I_2^{n,H}[\alpha, \beta] \le n^{\delta' - \frac{1}{2}})$.

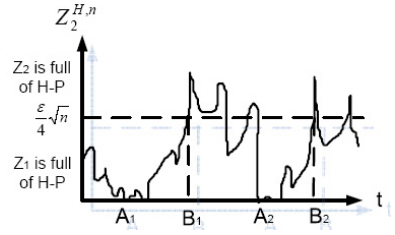

Fig. 5. Illustration of Z_2^H sample-path

On the event $H_n \cap \{I_2^{n,H}[\alpha, \beta] > n^{\delta' - \frac{1}{2}}\}$, the following holds: $\mathrm{P}(H_n, \ I_2^{n,H}[\alpha, \beta] > n^{\delta' - \frac{1}{2}}) \le \mathrm{P}(I_2^{n,H}[\alpha, \beta] > n^{\delta' - \frac{1}{2}}) \le \mathrm{P}(I_2^{n,H}[\tau, \sigma) > n^{\delta' - \frac{1}{2}}) \longrightarrow_n 0$, according to Lemma 1.

On the event $H_n \cap \{I_2^{n,H}[\alpha, \beta] \le n^{\delta' - \frac{1}{2}}\}$, notice that $B_2^{n,L}[\alpha, \beta] \le I_2^{n,H}[\alpha, \beta]$, since $I_2^{n,H}[\alpha, \beta] = I_2^{n,T}[\alpha, \beta] + B_2^{n,L}[\alpha, \beta]$. Therefore on the event considered, the following relations hold:

- Averaged LP departure rate: $\mu_2^n B_2^{n,L}[\alpha, \beta] \le n^{\frac{1}{2}+\frac{\delta}{2}}$;
- LP cumulative departures: $D_2^{n,L}[\alpha, \beta] \ge \frac{1}{2} n^{\frac{1}{2}+\delta}$.

One can verify that the probability for such an event converges to zero, as $n \to \infty$. This completes the proof of Lemma 2. ■

Proof of Lemma 3: Note that $D_2^{n,L}[\tau, \sigma] \ge A_{3,4}^{n,H}[\tau, \sigma] \ge \frac{\epsilon}{2}\sqrt{n}$, since $A_{3,4}^{n,H}(t) = D_2^{n,L}(t) - F_1(D_2^{n,L}(t))$. For any fixed $\delta \in (0, 1/4)$, we now define the event $H_n = \{|\sigma - \tau| < n^{-\delta}, \ Z_{1,2}^{n,H}(s) \ge \frac{\epsilon}{3}\sqrt{n}, \ \forall s \in [\tau, \sigma], \ D_2^{n,L}[\tau, \sigma] \ge \frac{\epsilon}{2}\sqrt{n}\}$. Let

$$\alpha = \inf\{t \ge \tau \ : \ Z_2^{n,H}(t) \ge \tfrac{\epsilon}{4}\sqrt{n}\};$$
$$\sigma' = \inf\{\tau \le t \le \sigma \ : \ D_2^{n,L}[\tau, t] \ge \tfrac{\epsilon}{2}\sqrt{n}\};$$

and represent $\mathrm{P}(H_n) = \mathrm{P}(H_n, \ \alpha \le \sigma') + \mathrm{P}(H_n, \ \alpha > \sigma')$.

On the event $H_n \cap \{\alpha \le \sigma'\}$, $Z_2^{n,H}$ must perform at least one *Down Crossing* from $\frac{\epsilon}{4}\sqrt{n}$ to 0, before the completion of $\Delta D_2^{n,L} \ge \frac{\epsilon}{2}\sqrt{n}$. i.e., server 2 will not serve LP tasks unless the number of HP tasks decreases to 0. Hence, the probability for such an event is less than $\mathrm{P}(mod_T(\hat{X}_2^n, n^{-\delta}) \ge \frac{\epsilon}{4}) \longrightarrow 0$, where the latter convergence is due to C-tightness.

On the event $H_n \cap \{\alpha > \sigma'\}$, Station 2 serves more than $\frac{\epsilon}{2}\sqrt{n}$ LP tasks on interval $[\tau, \sigma']$, while Server 1 is continuously busy ($Z_{1,2}^{n,H}(s) \ge \frac{\epsilon}{3}\sqrt{n}$, hence $Z_1^{n,H}(s) \ge \frac{\epsilon}{12}\sqrt{n}$, $\forall s \in [\tau, \sigma]$) with HP tasks, which are served and depart to Server 2. Note that

$$D_2^{n,T}[\tau, \sigma'] = D_2^{n,L}[\tau, \sigma'] + D_2^{n,H}[\tau, \sigma']. \tag{11}$$

On the event $H_n \cap \{\alpha > \sigma'\}$, one has

$$\begin{cases} \Delta D_2^{n,T}[\tau, \sigma'] = S_2^T(\mu_2^n B_2^{n,T}(\sigma')) - S_2^T(\mu_2^n B_2^{n,T}(\tau)); \\ \Delta D_2^{n,L}[\tau, \sigma'] \ge \frac{\epsilon}{2} \cdot \sqrt{n}; \\ \Delta D_2^{n,H}[\tau, \sigma'] \ge S_1^H(\mu_1^n B_1^{n,H}(\sigma')) - S_1^H(\mu_1^n B_1^{n,H}(\tau)). \end{cases}$$

The last inequality is due to the preemptive control, i.e., all HP tasks must be served before LP service can begin. Therefore, on the event considered, the following relations hold:

- $\Delta D_2^{n,T}[\tau,\sigma^{'}] - \Delta D_1^{n,H}[\tau,\sigma^{'}] \geq \frac{\epsilon}{2}\sqrt{n}$;
- $|\sigma^{'} - \tau| \leq |\sigma - \tau| < n^{-\delta}$;
- $B_1^{n,H}[\tau,\sigma^{'}] = |\sigma^{'} - \tau| \geq B_2^{n,T}[\tau,\sigma^{'}]$, since server 1 is **always busy** with HP customers;
- Recall also that $\mu_2 = \mu_1$ (Section VI).

One can verify that the probability for such an event converges to zero, as $n \to \infty$. This completes the proof of Lemma 3. ■

XI. Generalization and Extensions

We have not calculated the limiting distribution of the throughput process (Section IX). Our derivations are also restricted to preemptive controls and exponential service durations. We believe, however, that ideas from the proof (e.g., C-tightness, down-crossing considerations) may be used in far greater generality.

The model considered can be extended in various ways. We now describe some that are especially relevant to our healthcare motivation (in an increasing order of difficulty).

Multiple processing routes. Consider M parallel routes, rather that 2 as in Fig. 1.

Optimality conditions (maximizing throughput):

- Exact Optimality: $\bigwedge_{i\in\{1,..,M\}}(Q_i(T)) = 0, \forall T > 0$;
- Asymptotic Optimality: $\bigwedge_{i\in\{1,..,M\}}(\hat{Q}_i^n(\cdot)) \to 0$, u.o.c., in probability.

Optimal control: At each route, assign preemptive priority to tasks of customers whose service is completed at *all other* routes. Optimality is based on the following analogue of Property 3 in Section VII: $\sum_{i=1}^M Q_i(T) = \sum_{i=1}^M (L_i(T) - \bigwedge_{i\in\{1,..,M\}}(L_i(T))) + M \cdot \bigwedge_{i\in\{1,..,M\}}(Q_i(T))$, for all $T > 0$.

The Customer View, or the Snapshot Principle. The principle asserts [15], [13] that the *"state"* (e.g. queue-lengths) which a customer "sees" upon arrival does not change (in diffusion scale) during that customer's sojourn within the system. The validity of this principle thus dramatically simplifies the prediction of customer sojourn times, a problem that is important in our motivating service (healthcare) applications. However, our asymptotically optimal γ-control creates a volatile environment of priority switches (LP to HP). This renders challenging even the precise articulation of the snapshot principle, which we thus leave as a natural significant direction for future research.

Heterogeneous customer population. Consider a FJN with several customer classes: each class has its own precedence constraints, interarrival-time and service-time distributions. This is the model in Nguyen [14] where, in addition, a FCFS discipline was enforced at each station. The heavy traffic limits in [14] turn out intractable, which is due to task disordering in view of ample overtaking. However, we conjecture that applying static priority among customer classes (the same priority uniformly across stations, with FCFS within a class), will reduce in heavy traffic to a fork-join critical single-class FCFS network. This is a consequence of the collapse of high-priority processes [15] that "see" a network in light-traffic, which leaves only the lowest priority class in heavy-traffic. A far more challenging question is (asymptotically) optimal control of such heterogenous networks (eg. mixing global γ-control with station-level $Gc\mu$).

Many-server FJNs. Consider a FJN in which the number of servers per station is very large. Due to a high level of parallel processing, the phenomena of customer overtaking becomes both uncontrollable (as long as parallel servers operate independently) and non-negligible (if the number of servers scales up sufficiently fast). This is in contrast to multi-servers in conventional heavy-traffic [17]. The question that now arises is whether there exists a control under which Fork-Join and assembly networks are asymptotic equivalent. An interesting scaling to contemplate is the one in Halfin and Whitt [9]. Such many-server networks seem important since they naturally arise in intelligence or biological networks.

References

[1] B. Avi-Itzhak and S. Halfin, "Non-preemptive priorities in simple fork-join queues," in *Queueing, Performance and Control in ATM (ITC-13)*, I. W. Cohen and C. D. Pack, Eds. Elsevier Science Publishers B.V. (North-Holland), 1991.

[2] F. Baccelli and A. M. Makowski, "Queueing models for systems with synchronization constraints," in *Proceedings of the IEEE*, vol. 77, pp. 138–161, 1989.

[3] F. Baccelli, A. M. Makowski and A. Shwartz, "The fork-join queue and related systems with synchronizaton constraints: Stochastic ordering, approximations and computable bounds," *J. Adv. Probab.* vol. 21, pp. 629–660, 1989.

[4] P. Billingsley, "Convergence of Probability Measures," *Wiley Series in Probability and Mathematical Statistics*, 1968.

[5] O. J. Boxma, G. Koole and Z. Liu, "Queueing-Theoretic Solution Methods for Models of Parallel and Distributed Systems," *QMIPS*, 1996.

[6] H. Chen and D. Yao, "Fundamentals of Queueing Networks," *Springer-Verlag*, 2001.

[7] I. Cohen, A. Mandelbaum and A. Shtub, "Multi-project scheduling and control: A process-based comparative study of the critical chain methodology and some alternatives," *Project Management Journal*, vol. 35, pp. 39–50, 2004.

[8] G.J. Hoekstra, R.D. van der Mei and S. Bhulai, "Optimal job splitting in parallel processor sharing queues," *Stochastic Models*, vol. 28, pp. 144–166, 2012.

[9] S. Halfin and W. Whitt, "Heavy traffic limits for queues with many exponential servers," *Oper. Res.*, vol. 29, pp. 567–588, 1981.

[10] J. Jacod and A.N. Shiryaev, "Limit Theorems for Stochastic Processes," *Springer*, 1987.

[11] R. Larson, M. Cahn and M. Shell, "Improving the N.Y.C A-to-A system," *Interfaces*, vol. 23, pp. 76–96, 1993.

[12] R. Nelson, D. Towsley and A. N. Tantawi, "Performance analysis of parallel processing systems," *IEEE Trans. on Parallel and Software Engineering*, vol. 14, no. 4, pp. 532–540, 1988.

[13] V. Nguyen, "Heavy traffic analysis of processing networks with parallel and sequential tasks," *The Annals of Applied Probability*, vol. 3, pp. 28–55, 1993.

[14] V. Nguyen, "The troubles with diversity: Fork-join networks with heterogeneous customer population," *The Annals of Applied Probability*, vol. 4, pp. 1–25, 1994.

[15] M.I. Reiman and B. Simon, "A Network of Proiority Queues in Heavy Traffic: One Bottleneck Station," *Queueing Systems*, vol. 6, pp. 33–58, 1990.

[16] M. S. Squillante, Y. Zhang, A. Sivasubramaniam and N. Gautam, "Generalized parallel-server fork-join queues with dynamic task scheduling," *Ann. Oper. Res.*, vol. 160, pp. 227–255, 2008.

[17] A. Zviran, "Fork-join networks in heavy traffic: Diffusion approximation and control [Online]," Technion, 2011. Available: http://ie.technion.ac.il/serveng/References/Thesis-paper.pdf