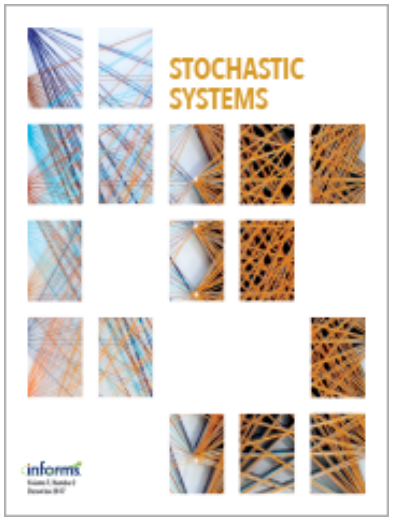

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

On the Power of (Even a Little) Resource Pooling

John N. Tsitsiklis, Kuang Xu

To cite this article:
John N. Tsitsiklis, Kuang Xu (2012) On the Power of (Even a Little) Resource Pooling. Stochastic Systems 2(1):1-66.
https://doi.org/10.1287/11-SSY033

Stochastic Systems
2012, Vol. 2, No. 1, 1–66
DOI: 10.1214/11-SSY033

ON THE POWER OF (EVEN A LITTLE) RESOURCE POOLING*†

By John N. Tsitsiklis and Kuang Xu

Massachusetts Institute of Technology

We propose and analyze a multi-server model that captures a performance trade-off between centralized and distributed processing. In our model, a fraction p of an available resource is deployed in a centralized manner (e.g., to serve a most-loaded station) while the remaining fraction $1-p$ is allocated to local servers that can only serve requests addressed specifically to their respective stations.

Using a fluid model approach, we demonstrate a surprising *phase transition* in the *steady-state delay scaling*, as p changes: in the limit of a large number of stations, and when *any amount* of centralization is available ($p>0$), the average queue length in steady state scales as $\log_{\frac{1}{1-p}} \frac{1}{1-\lambda}$ when the traffic intensity λ goes to 1. This is *exponentially smaller* than the usual $M/M/1$-queue delay scaling of $\frac{1}{1-\lambda}$, obtained when all resources are fully allocated to local stations ($p=0$). This indicates a strong qualitative impact of even a small degree of resource pooling.

We prove convergence to a fluid limit, and characterize both the transient and steady-state behavior of the actual system, in the limit as the number of stations N goes to infinity. We show that the sequence of queue-length processes converges to a *unique* fluid trajectory (over any finite time interval, as $N\to\infty$), and that this fluid trajectory converges to a unique invariant state $\mathbf{v}^I$, for which a simple closed-form expression is obtained. We also show that the steady-state distribution of the N-server system concentrates on $\mathbf{v}^I$ as N goes to infinity.

Received August 2011.

*Research supported in part by an MIT Jacobs Presidential Fellowship, an MIT-Xerox Fellowship, a Siebel Scholarship, and NSF grant CCF-0728554. The authors are grateful to the anonymous reviewers and Yuan Zhong for a careful reading of the manuscript and constructive feedback, and to Nicolas Gast for helpful discussions on the subject.

†Revised March 2012. A preliminary version of this paper appeared in the Proceedings of the 2011 ACM Sigmetrics Conference.

AMS 2000 subject classifications: Primary 60K25; secondary 60K30, 60F17, 90B15, 90B22, 37C10.

Keywords and phrases: Queueing, service flexibility, resource pooling, asymptotics, fluid approximation.

CONTENTS

1. Introduction. The tension between *distributed* and *centralized* processing seems to have existed ever since the inception of computer networks. Distributed processing allows for simple implementation and robustness, while a centralized scheme guarantees optimal pooling of processing resources, at the cost of implementation complexity and communication overhead. A natural question is how performance varies with the *degree of centralization*, or *resource pooling*. Such understanding is of great interest in the context of, for example, infrastructure planning (static) or task scheduling (dynamic) in large server farms or cloud computing clusters, and can provide insights on the trade-off between performance (e.g., delay) and cost (e.g., communication infrastructure, energy consumption, etc.).

It is well known that resource pooling can drastically improve performance, as exemplified by the comparison of $M/M/1$ and $M/M/n$ queueing systems with the same total arrival and service rates. The main message of this paper is that even a small degree of resource pooling can deliver significant benefits. We capture this effect by formulating and analyzing a multi-server model with a *limited* level of centralization. We begin by describing informally two motivating applications.

1.1. *Primary Motivation: Server farm with local and central servers.* Consider a server farm consisting of N stations, depicted in Figure 1. Each station is fed with an independent stream of tasks, arriving at a rate of λ tasks per second, with $0 < \lambda < 1$,[1] and is equipped with a *local server* with identical performance; these servers are local in the sense that each one can only serve its own station. All stations are also connected to a single *centralized server* that always serves a task (if one exists) at a station with the longest queue.

We consider an N-station system. The system designer is granted a total amount N of divisible *computing resources* (e.g., a collection of processors). In a loose sense (to be formally defined in Section 2.1), this means that the system is capable of processing N tasks per second when fully loaded. The system designer allocates computing resources to local and central servers. Specifically, for some $p \in (0,1)$, each of the N local servers is able to process tasks at a maximum rate of $1-p$ tasks per second, while the centralized server, equipped with the remaining computing power, is capable of processing tasks at a maximum rate of pN tasks per second. The parameter p captures the amount of centralization in the system. Note that since the total arrival rate is λN, with $0 < \lambda < 1$, the system is underloaded for any value $p \in (0,1)$.

[1] Without loss of generality, we normalize so that the largest possible arrival rate is 1.

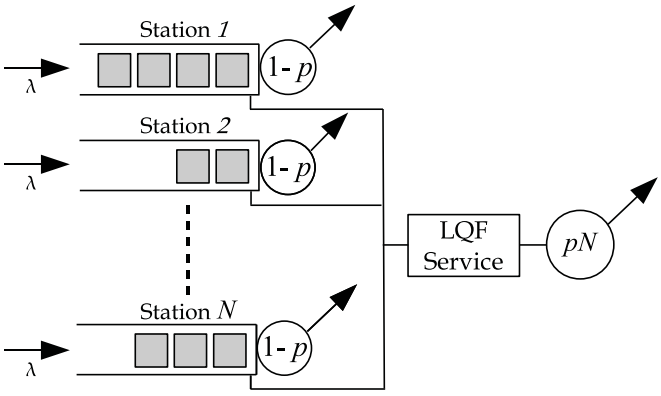

FIG 1. *Server farm with local and central servers.*

When the arrival processes and task processing times are random, there will be times when some stations are empty while others are loaded. Since a local server cannot help another station process tasks, the total computational resources will be better utilized if a larger fraction is allocated to the central server. However, a greater degree of centralization (corresponding to a larger value of p) entails more frequent communications and data transfers between the local stations and the central server, resulting in higher infrastructure and energy costs.

How should the system designer choose the coefficient p? Alternatively, we can ask an even more fundamental question: is there a significant difference between having a small amount of centralization (a small but positive value of p), and complete decentralization (no central server and $p = 0$)?

1.2. *Secondary Motivation: Partially centralized scheduling.* Consider a system with N stations, depicted in Figure 2. The arrival assumptions are the same as in Section 1.1. However, there is no local server associated with a station; all stations are served by a single central server. Whenever the central server becomes free, it chooses a task to serve as follows. With probability p, it processes a task from a most loaded station, with an arbitrary tie-breaking rule. Otherwise, it processes a task from a station selected uniformly at random; if the randomly chosen station has an empty queue, the current round is in some sense wasted (to be formalized in Section 2.1).

This second interpretation is intended to model a scenario where resource allocation decisions are made at a centralized location on a *dynamic* basis, but *communications* between the decision maker (central server) and local stations are costly or simply unavailable from time to time. While it is intu-

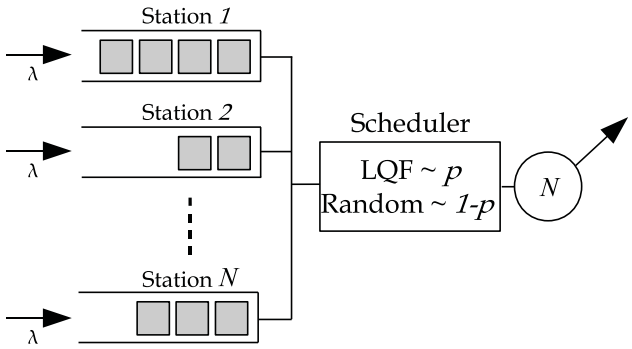

FIG 2. *Centralized scheduling with communication constraints.*

itively obvious that longest-queue-first (LQF) scheduling is more desirable, up-to-date state information (i.e., queue lengths at all stations) may not always be available to the central server. Thus, the central server may be forced to allocate its service blindly. In this setting, a system designer is interested in a judicious choice of the *frequency* (p) at which global state information is collected, so as to balance performance and communication costs.

As we will see in the sequel, the system dynamics in the two applications are captured by the *same* mathematical structure under appropriate stochastic assumptions on task arrivals and processing times, and hence will be addressed jointly in this paper.

1.3. *Overview of main contributions.* We provide here an overview of the main contributions. Exact statements of the results will be provided in Section 3 after the necessary terminology has been introduced.

Our goal is to study the performance implications of varying degrees of centralization (or resource pooling), as expressed by the coefficient p. To accomplish this, we use a so-called *fluid approximation*, whereby the queue length dynamics at the local stations are approximated, as $N \to \infty$, by a deterministic *fluid model*, governed by a system of ordinary differential equations (ODEs).

Fluid approximations typically involve results of two flavors: qualitative results derived from the fluid model that give insights into the performance of the original finite stochastic system, and technical convergence results (often mathematically involved) that justify the use of such approximations. We summarize our contributions along these two dimensions:

1. On the **qualitative end**, we derive an exact expression for the invariant state of the fluid model, for any given traffic intensity λ and

centralization coefficient p, thus characterizing the steady-state distribution of the queue lengths in the system as $N \to \infty$. This enables a system designer to use any performance metric and analyze its sensitivity with respect to p. In particular, we show a surprising *exponential phase transition* in the scaling of average system delay as the load approaches capacity ($\lambda \to 1$) (Corollary 3 in Section 3.2): when an *arbitrarily small* amount of centralized computation is applied ($p > 0$), the average queue length in the system scales as[2]

$$\mathbb{E}(Q) \sim \log_{\frac{1}{1-p}} \frac{1}{1-\lambda}, \tag{1}$$

as the traffic intensity λ approaches 1. This is *drastically smaller* than the $\frac{1}{1-\lambda}$ scaling obtained if there is no centralization ($p = 0$).[3] This suggests that for large systems, even a small degree of resource pooling provides significant improvements in the system's delay performance, in the heavy traffic regime.

2. On the **technical end**, we show that:
   (a) Given any finite initial queue sizes, and with high probability, the evolution of the queue length process can be approximated (over any finite time interval, and as $N \to \infty$) by the unique solution to a fluid model.
   (b) All solutions to the fluid model converge to an invariant state, as $t \to \infty$, which is the same for all finite initial conditions (uniqueness and global stability).
   (c) The steady-state distribution of the finite system converges to the invariant state of the fluid model as $N \to \infty$.

   The most notable technical challenge comes from the fact that the longest-queue-first policy used by the centralized server causes discontinuities in the drift in the fluid model (see Section 3.1 for details). In particular, the classical approximation results for Markov processes (see, e.g., [2]), which rely on a Lipschitz-continuous drift in the fluid model, are hard to apply. Thus, in order to establish the finite-horizon approximation result (a), we employ a sample-path based approach: we prove tightness of sample paths of the queue length process and characterize their limit points. Establishing the convergence of steady-state

[2] The $\sim$ notation used in this paper is to be understood as *asymptotic closeness* in the following sense: $[f(x) \sim g(x), \text{ as } x \to 1] \Leftrightarrow \lim_{x\to 1} \frac{f(x)}{g(x)} = 1$.

[3] When $p = 0$, the system degenerates into N independent queues. The $\frac{1}{1-\lambda}$ scaling comes from the mean queue length expression for $M/M/1$ queues.

distributions in (c) also becomes non-trivial due to the presence of discontinuous drifts. To derive this result, we first establish the uniqueness of solutions to the fluid model and a uniform (over a compact set of initial conditions) speed of convergence of stochastic sample paths to the solution of the fluid model.

1.4. *Related work.* To the best of our knowledge, the proposed model for the splitting of processing resources between local and central servers has not been studied before. However, the fluid model approach used in this paper is closely related to, and partially motivated by, the so-called supermarket model of randomized load-balancing. In that literature, it is shown that by routing tasks to the shorter queue among a small number ($d \geq 2$) of randomly chosen queues, the probability that a typical queue has at least i tasks (denoted by $\mathbf{s}_i$) decays as $\lambda^{\frac{d^i-1}{d-1}}$ (super-geometrically), as $i \to \infty$ ([3, 4]); see also the survey paper [8] and references therein. However, this sampling approach to load-balancing seems to offer little improvement when adapted to scheduling. In [5], a variant of the randomized load-balancing policy was applied to a scheduling setting with channel uncertainties, where the server always schedules a task from a longest queue among a finite number of randomly selected queues. It was observed that $\mathbf{s}_i$ no longer exhibits super-geometric decay and only moderate performance gain can be harnessed from sampling more than one queue.

In our setting, the system dynamics causing the exponential phase transition in the average queue length scaling are significantly different from those for the randomized load-balancing scenario. In particular, for any $p > 0$, the steady-state tail probabilities $\mathbf{s}_i$ become zero for sufficiently large finite i, which is markedly faster than the super-geometric decay in the supermarket model.

On the technical side, arrivals and processing times used in supermarket models are often memoryless (Poisson or Bernoulli) and the drifts in the fluid model are typically continuous with respect to the underlying system state. Hence convergence results can be established by invoking classical approximation results, based on the convergence of the generators of the associated Markov processes. An exception is [7], where the authors generalize the supermarket model to arrival and processing times with general distributions. Since the queue length process is no longer Markov, the authors rely on an asymptotic independence property of the limiting system and use tools from statistical physics to establish convergence.

Our system is Markov with respect to the queue lengths, but a significant technical difference from the supermarket model lies in the fact that the

longest-queue-first service policy introduces *discontinuities* in the drifts. For this reason, we need to use a more elaborate set of techniques to establish the connection between stochastic sample paths and the fluid model. Moreover, the presence of discontinuities in the drifts creates challenges even for proving the uniqueness of solutions for the deterministic fluid model. (Such uniqueness is used to establish convergence of steady-state distributions.) Our approach is based on a state representation that is different from the one used in the popular supermarket models, and which turns out to be surprisingly more convenient to work with for establishing the uniqueness of solutions to the fluid model.

Besides the queueing-theoretic literature, similar fluid model approaches have been used in many other contexts to study systems with large populations. Recent results in [6] establish convergence for finite-dimensional symmetric dynamical systems with drift discontinuities, using a more probabilistic (as opposed to sample path) analysis, carried out in terms of certain conditional expectations. We believe that it is possible to prove our results using the methods in [6], with additional work. However, the coupling approach used in this paper provides strong physical intuition on the system dynamics, and avoids the need for additional technicalities from the theory of multi-valued differential inclusions.

Resource pooling is known to improve performance [15–18], but much less is known on the impact of various degrees of pooling, or about scaling behaviors in large-system limits. Some recent work in this area [19] that studies limited pooling in a large-system limit is closer to our work in spirit, but still differs significantly in terms of critical modeling assumptions and dynamics. The notion of limited flexibility has also been studied in manufacturing systems, such as the celebrated Long Chain design [9, 10] and its variants [11–14]. However, models considered in this literature are typically applied to static allocation problems (with a single or small number of stages), whereas our system involves non-trivial queueing dynamics, where resource allocation decisions have to be made repeatedly overtime.

Finally, there has been some work on the impact of service flexibility in routing problems, motivated by applications such as multilingual call centers. These date back to the seminal work of [20], with a more recent numerical study in [21]. These results show that the ability to route a portion of customers to a least-loaded station can lead to a constant-factor improvement in average delay under diffusion scaling. This line of work is very different from ours, but in a broader sense, both are trying to capture the notion that system performance in a random environment can benefit significantly from even a small amount of centralized coordination.

1.5. *Organization of the paper.* Section 2 introduces the precise model to be studied, our assumptions, and the notation to be used throughout. The main results are summarized in Section 3, where we also discuss their implications along with some numerical results. The remainder of the paper is devoted to establishing the technical results, and the reader is referred to Section 4.1 for an overview of the proofs. The steps of some of the more technical proofs are outlined in the main text, with complete proofs relegated to Appendix A. The procedure and parameters used for numerical simulations are described in Appendix D.

2. Model and notation. This section covers the modeling assumptions, system state representations, and mathematical notation that will be used throughout the paper. We provide some intuition behind our modeling choices and assumptions whenever possible, but if the ideas involved cannot be made transparent at this stage, we point the reader to explanations that will appear later in the paper.

2.1. *Model.* We present our model using terminology that corresponds to the server farm application in Section 1.1. Time is assumed to be continuous.

1. **System.** The system consists of N parallel stations. Each station is associated with a queue which stores the tasks to be processed. The queue length (i.e., number of tasks) at station n at time t is denoted by $Q_n(t)$, $n \in \{1, 2, \ldots, N\}$, $t \geq 0$.
2. **Arrivals.** Stations receive streams of incoming tasks according to independent Poisson processes with a common rate $\lambda \in [0, 1)$.
3. **Task Processing.** We fix a centralization coefficient $p \in [0, 1]$.
   (a) **Local Servers**. The local server at station n is modeled by an independent Poisson clock with rate $1-p$ (i.e., the times between two clock ticks are independent and exponentially distributed with mean $\frac{1}{1-p}$). If the clock at station n ticks at time t, we say that a **local service token** is generated at station n. If $Q_n(t) \neq 0$, exactly one task from station n "consumes" the service token and leaves the system immediately. Otherwise, the local service token is "wasted" and has no impact on the future evolution of the system.
   (b) **Central Server**. The central server is modeled by an independent Poisson clock with rate Np. If the clock ticks at time t at the central server, we say that a **central service token** is generated. If the system is non-empty at time t (i.e., if $\sum_{n=1}^{N} Q_n(t) > 0$),

exactly one task from some station n, chosen uniformly at random out of the stations with a *longest queue* at time t, consumes the service token and leaves the system immediately. If the whole system is empty, the central service token is wasted.

Physical interpretation of service tokens. We interpret $Q_n(t)$ as the number of tasks whose service has not yet started. For example, if there are four tasks at station n, one being served and three that are waiting, then $Q_n(t) = 3$. The use of *local service tokens* can be thought of as an approximation to a *work-conserving*[4] server with exponential service time distribution in the following sense. Let t_k be the kth tick of the Poisson clock at the server associated with station n. If $Q_n(t_k-) > 0$,[5] the ticking of the clock can be thought of as the completion of a previous task, so that the server "fetches" a new task from the queue to process, hence decreasing the queue length by 1. Therefore, as long as the queue remains non-empty, the time between two consecutive clock ticks can be interpreted as the service time for a task. On the other hand, if the local queue is currently empty, i.e., $Q_n(t_k-) = 0$, then our modeling assumption implies that the local server does nothing until the next clock tick at t_{k+1}, even if some task arrives during the period (t_k, t_{k+1}). Alternatively, this can be thought of as the server creating a "virtual task" whenever it sees an empty queue, and pretending to be serving the virtual task until the next clock tick. In contrast, a work-conserving server would start serving the next task immediately upon its arrival. We have chosen to use the service token setup, mainly because it simplifies analysis, and because it can also be justified in the following ways.

1. Because of the use of virtual tasks, one would expect the resulting queue length process under our setup to provide an *upper bound* on the queue length process under a work-conserving server. We do not formally prove such a dominance relation in this paper, but note that a similar dominance result in $GI/GI/n$ queues was proved recently (Proposition 1 of [26]).
2. Since the discrepancy between the two setups only occurs when the server sees an empty queue, one would also expect that the queue length processes under the two cases become comparable as the traffic intensity λ approaches 1, in which case the queue at a local server will be non-empty most of the time.

The same physical interpretation applies to the central service tokens.

[4] A server is work-conserving if it is never idle when the queue is non-empty.

[5] Throughout the paper, we use the short-hand notation $f(t-)$ to denote the left limit $\lim_{s \uparrow t} f(s)$.

Mathematical equivalence between the two motivating applications. We note here that the scheduling application in Section 1.2 corresponds to the same mathematical model. The arrival statistics to the stations are obviously identical in both models. For task processing, note that we can equally imagine all service tokens as being generated from a single Poisson clock with rate N. Upon the generation of a service token, a coin is flipped to decide whether the token will be directed to fetch a task from a random station for processing (corresponding to a *local service token*), or from a station with a longest queue (corresponding to a *central service token*). Due to the Poisson splitting property, this produces identical statistics for the generation of local and central service tokens as for the server farm application.

2.2. *System state.* Let us fix N. Since all events (arrivals of tasks and service tokens) are generated according to independent Poisson processes, the queue length vector at time t, $(Q_1(t), Q_2(t), \ldots, Q_N(t))$, is Markov. Moreover, the system is fully symmetric, in the sense that all queues have identical and independent statistics for the arrivals and local service tokens, and the assignment of central service tokens does not depend on the specific identity of stations besides their queue lengths. Hence we can use a Markov process $\left\{\mathbf{S}_i^N(t)\right\}_{i=0}^{\infty}$ to describe the evolution of a system with N stations, where

$$\mathbf{S}_i^N(t) \triangleq \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}_{[i,\infty)}\left(Q_n(t)\right), \quad i \geq 0. \tag{2}$$

Each coordinate $\mathbf{S}_i^N(t)$ represents the fraction of queues with at least i tasks. Note that $\mathbf{S}_0^N(t) = 1$, for all t and N, according to this definition. We call $\mathbf{S}^N(t)$ the **normalized queue length process**. We also define the **aggregate queue length process** as

$$\mathbf{V}_i^N(t) \triangleq \sum_{j=i}^{\infty} \mathbf{S}_j^N(t), \quad i \geq 0. \tag{3}$$

Note that

$$\mathbf{S}_i^N(t) = \mathbf{V}_i^N(t) - \mathbf{V}_{i+1}^N(t). \tag{4}$$

In particular, this means that $\mathbf{V}_0^N(t) - \mathbf{V}_1^N(t) = \mathbf{S}_0^N(t) = 1$. Note also that

$$\mathbf{V}_1^N(t) = \sum_{j=1}^{\infty} \mathbf{S}_j^N(t) \tag{5}$$

is equal to the *average queue length* in the system at time t. More generally, $\mathbf{V}_i^N(t)$ can be interpreted as the *average of excess queue lengths above* $i-1$ at time t, in the sense that[6]

$$\mathbf{V}_i^N(t) = \mathbb{E}\left((X_t - i + 1)^+\right), \quad \forall i \geq 1, \tag{6}$$

where X_t is a random variable distributed according to the empirical queue length distribution in the system at time t: $\mathbb{P}(X_t \geq i) = \mathbf{S}_i^N(t)$, for all $i \geq 0$.

When the total number of tasks in the system is finite (hence all coordinates of $\mathbf{V}^N$ are finite), there is a straightforward bijection between $\mathbf{S}^N$ and $\mathbf{V}^N$. Hence $\mathbf{V}^N(t)$ is Markov and also serves as a valid representation of the system state. While the $\mathbf{S}^N$ representation admits a more intuitive interpretation as the "tail" probability of a typical station having at least i tasks, it turns out the $\mathbf{V}^N$ representation is significantly more convenient to work with, especially in proving uniqueness of solutions to the associated fluid model; the detailed reasons will become clear in the sequel (see Appendix B for an extensive discussion on this topic). For this reason, we will be working mostly with the $\mathbf{V}^N$ representation, but will in some places state results in terms of $\mathbf{S}^N$, if doing so provides a better physical intuition.

2.3. *Notation.* Let $\mathbb{Z}_+$ be the set of non-negative integers. The following sets will be used throughout the paper (where M is a positive integer):

$$\mathcal{S} \triangleq \left\{\mathbf{s} \in [0,1]^{\mathbb{Z}_+} : 1 = \mathbf{s}_0 \geq \mathbf{s}_1 \geq \cdots \geq 0\right\}, \tag{7}$$

$$\overline{\mathcal{S}}^M \triangleq \left\{\mathbf{s} \in \mathcal{S} : \sum_{i=1}^{\infty} \mathbf{s}_i \leq M\right\}, \quad \overline{\mathcal{S}}^\infty \triangleq \left\{\mathbf{s} \in \mathcal{S} : \sum_{i=1}^{\infty} \mathbf{s}_i < \infty\right\}, \tag{8}$$

$$\overline{\mathcal{V}}^M \triangleq \left\{\mathbf{v} : \mathbf{v}_i = \sum_{j=i}^{\infty} \mathbf{s}_j, \text{ for some } \mathbf{s} \in \overline{\mathcal{S}}^M\right\}, \tag{9}$$

$$\overline{\mathcal{V}}^\infty \triangleq \left\{\mathbf{v} : \mathbf{v}_i = \sum_{j=i}^{\infty} \mathbf{s}_j, \text{ for some } \mathbf{s} \in \overline{\mathcal{S}}^\infty\right\}, \tag{10}$$

$$\mathcal{Q}^N \triangleq \left\{\mathbf{x} \in \mathbb{R}^{\mathbb{Z}_+} : \mathbf{x}_i = \frac{K_i}{N}, \text{ for some } K_i \in \mathbb{Z}_+, \forall i\right\}. \tag{11}$$

We define the weighted L_2 norm $\|\cdot\|_w$ on $\mathbb{R}^{\mathbb{Z}_+}$ as

$$\|\mathbf{x} - \mathbf{y}\|_w^2 = \sum_{i=0}^{\infty} \frac{|\mathbf{x}_i - \mathbf{y}_i|^2}{2^i}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathbb{Z}_+}. \tag{12}$$

[6] $(x)^+ \triangleq \max\{0, x\}$.

In general, we will be using bold letters to denote vectors and ordinary letters for scalars, with the exception that a bold letter with a subscript (e.g., $\mathbf{v}_i$) is understood as a (scalar-valued) component of a vector. Upper-case letters are generally reserved for random variables (e.g., $\mathbf{V}^{(0,N)}$) or stochastic processes (e.g., $\mathbf{V}^N(t)$), and lower-case letters are used for constants (e.g., $\mathbf{v}^0$) and deterministic functions (e.g., $\mathbf{v}(t)$). Finally, a function is in general denoted by $x(\cdot)$, but is sometimes written as $x(t)$ to emphasize the type of its argument.

3. Summary of main results. In this section, we provide the exact statements of our main results. The main approach of our work is to first derive key performance guarantees using a simpler fluid model, and then apply probabilistic arguments (e.g., Functional Laws of Large Numbers) to formally justify that such guarantees also carry over to sufficiently large finite stochastic systems. Section 3.1 gives a formal definition of the core fluid model used in this paper, along with its physical interpretation. Section 3.2 contains results that are derived by analyzing the dynamics of the fluid model, and Section 3.3 contains the more technical convergence theorems that justify the accuracy of approximating a finite system using the fluid model approach. The proofs for the theorems stated here will be developed in later sections.

3.1. *Definition of fluid model.*

Definition 1 (Fluid Model). Given an initial condition $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$, a function $\mathbf{v}(t) : [0,\infty) \to \overline{\mathcal{V}}^\infty$ is said to be a **solution to the fluid model** (or **fluid solution** for short) if:

(a) $\mathbf{v}(0) = \mathbf{v}^0$;

(b) for all $t \geq 0$,

$$\mathbf{v}_0(t) - \mathbf{v}_1(t) = 1, \tag{13}$$

$$\text{and } 1 \geq \mathbf{v}_i(t) - \mathbf{v}_{i+1}(t) \geq \mathbf{v}_{i+1}(t) - \mathbf{v}_{i+2}(t) \geq 0, \quad \forall i \geq 0; \tag{14}$$

(c) for almost all $t \in [0,\infty)$, and for every $i \geq 1$, $\mathbf{v}_i(t)$ is differentiable and satisfies

$$\dot{\mathbf{v}}_i(t) = \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i) - (1-p)(\mathbf{v}_i - \mathbf{v}_{i+1}) - g_i(\mathbf{v}), \tag{15}$$

where

$$g_i(\mathbf{v}) = \begin{cases} p, & \mathbf{v}_i > 0, \\ \min\{\lambda \mathbf{v}_{i-1}, p\}, & \mathbf{v}_i = 0, \mathbf{v}_{i-1} > 0, \\ 0, & \mathbf{v}_i = 0, \mathbf{v}_{i-1} = 0. \end{cases} \tag{16}$$

We can write Eq. (15) more compactly as

$$\dot{\mathbf{v}}(t) = \mathbf{F}(\mathbf{v}), \tag{17}$$

where

$$\mathbf{F}_i(\mathbf{v}) \triangleq \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i) - (1-p)(\mathbf{v}_i - \mathbf{v}_{i+1}) - g_i(\mathbf{v}), \quad i \geq 1. \tag{18}$$

We call $\mathbf{F}(\mathbf{v})$ the **drift at point v**.

Interpretation of the fluid model. A solution to the fluid model, $\mathbf{v}(t)$, can be thought of as a deterministic approximation to the sample paths of $\mathbf{V}^N(t)$ for large values of N. Conditions (a) and (b) in Definition 1 correspond to initial and boundary conditions, respectively. The boundary conditions reflect the *physical constraints* of the finite system. For example, the condition that $\mathbf{v}_0(t) - \mathbf{v}_1(t) = 1$ corresponds to the fact that

$$\mathbf{V}_0^N(t) - \mathbf{V}_1^N(t) \triangleq \mathbf{S}_0^N(t) = 1, \tag{19}$$

where $\mathbf{S}_0^N(t)$ is the fraction of queues with a non-negative queue length, which is, by definition, 1. Similarly, the condition that

$$\mathbf{v}_i(t) - \mathbf{v}_{i+1}(t) \geq \mathbf{v}_{i+1}(t) - \mathbf{v}_{i+2}(t), \quad \forall i \geq 0, \tag{20}$$

is a consequence of

$$\left(\mathbf{V}_i^N(t) - \mathbf{V}_{i+1}^N(t)\right) - \left(\mathbf{V}_{i+1}^N(t) - \mathbf{V}_{i+2}^N(t)\right) \triangleq \mathbf{S}_i^N(t) - \mathbf{S}_{i+1}^N(t) \in [0,1], \tag{21}$$

where $\mathbf{S}_i^N(t) - \mathbf{S}_{i+1}^N(t)$ is the fraction of queues at time t with exactly i tasks, a number between 0 and 1.

We now provide some intuition for each of the drift terms in Eq. (15):

I. $\lambda(\mathbf{v}_{i-1} - \mathbf{v}_i)$: This term corresponds to arrivals. When a task arrives at a station with $i-1$ tasks, the system has one more queue with i tasks, and $\mathbf{S}_i^N$ increases by $\frac{1}{N}$. However, the number of queues with at least j tasks, for $j \neq i$, does not change. Thus, $\mathbf{S}_i^N$ is the only component of $\mathbf{S}^N$ that gets incremented. Since $\mathbf{V}_i^N \triangleq \sum_{k=i}^{\infty} \mathbf{S}_k^N$, this implies that $\mathbf{V}_i^N$ is increased by $\frac{1}{N}$ *if and only if* a task arrives at a queue with at least $i-1$ tasks. Since all stations have an identical arrival rate λ, the probability of $\mathbf{V}_i^N$ being incremented upon an arrival to the system is equal to the fraction of queues with at least $i-1$ tasks, which is $\mathbf{V}_{i-1}^N(t) - \mathbf{V}_i^N(t)$. We take the limit as $N \to \infty$, multiply by the total arrival rate, $N\lambda$, and then multiply by the increment due to each arrival, $\frac{1}{N}$, to obtain the term $\lambda(\mathbf{v}_{i-1} - \mathbf{v}_i)$.

II. $(1-p)(\mathbf{v}_i - \mathbf{v}_{i+1})$: This term corresponds to the completion of tasks due to *local* service tokens. The argument is similar to that for the first term.

III. $g_i(\mathbf{v})$: This term corresponds to the completion of tasks due to *central* service tokens.

1. $g_i(\mathbf{v}) = p$, if $\mathbf{v}_i > 0$. If $i > 0$ and $\mathbf{v}_i > 0$, then there is a positive fraction of queues with at least i tasks. Hence the central server is working at full capacity, and the rate of decrease in $\mathbf{v}_i$ due to central service tokens is equal to the (normalized) maximum rate of the central server, namely p.
2. $g_i(\mathbf{v}) = \min\{\lambda \mathbf{v}_{i-1}, p\}$, if $\mathbf{v}_i = 0, \mathbf{v}_{i-1} > 0$. This case is more subtle. Note that since $\mathbf{v}_i = 0$, the term $\lambda \mathbf{v}_{i-1}$ is equal to $\lambda(\mathbf{v}_{i-1} - \mathbf{v}_i)$, which is the rate at which $\mathbf{v}_i$ increases due to arrivals. Here the central server serves queues with at least i tasks whenever such queues arise, trying to keep $\mathbf{v}_i$ at zero. Thus, the total rate of central service tokens dedicated to $\mathbf{v}_i$ tries to match the rate of increase of $\mathbf{v}_i$ due to arrivals.[7]
3. $g_i(\mathbf{v}) = 0$, if $\mathbf{v}_i = \mathbf{v}_{i-1} = 0$. Here, both $\mathbf{v}_i$ and $\mathbf{v}_{i-1}$ are zero and there are no queues with $i-1$ or more tasks. Hence there is no positive rate of increase in $\mathbf{v}_i$ due to arrivals. Accordingly, the rate at which central service tokens are used to serve stations with at least i tasks is zero.

Note that, as was mentioned in the introduction, the discontinuities in the fluid model come from the term $g(\mathbf{v})$, which reflects the presence of a central server.

3.2. *Analysis of the fluid model and exponential phase transition.* The following theorem characterizes the invariant state for the fluid model. It will be used to demonstrate an *exponential improvement* in the rate of growth of the average queue length as $\lambda \to 1$ (Corollary 3).

THEOREM 2. *The drift* $\mathbf{F}(\cdot)$ *in the fluid model admits a unique invariant state* $\mathbf{v}^I$ *(i.e., a state that satisfies* $\mathbf{F}(\mathbf{v}^I) = 0$*). Letting* $\mathbf{s}_i^I \triangleq \mathbf{v}_i^I - \mathbf{v}_{i+1}^I$ *for all* $i \geq 0$*, the exact expressions for the invariant state are as follows:*

(1) If $p = 0$*, then* $\mathbf{s}_i^I = \lambda^i$, $\forall i \geq 1$.
(2) If $p \geq \lambda$*, then* $\mathbf{s}_i^I = 0$, $\forall i \geq 1$.

[7] Technically, the minimization involving p is not necessary: if $\lambda \mathbf{v}_{i-1}(t) > p$, then $\mathbf{v}_i(t)$ cannot stay at zero and will immediately increase after t. We keep the minimization just to emphasize that the maximum rate of increase in $\mathbf{v}_i$ due to central service tokens cannot exceed the central service capacity p.

(3) If $0 < p < \lambda$ *and* $\lambda = 1 - p$, *then*[8]

$$\mathbf{s}_i^I = \begin{cases} 1 - \left(\frac{p}{1-p}\right) i, & 1 \le i \le \tilde{i}^*(p,\lambda), \\ 0, & i > \tilde{i}^*(p,\lambda), \end{cases}$$

where $\tilde{i}^*(p,\lambda) \triangleq \left\lfloor \frac{1-p}{p} \right\rfloor$.

(4) If $0 < p < \lambda$ *and* $\lambda \neq 1 - p$, *then*

$$\mathbf{s}_i^I = \begin{cases} \frac{1-\lambda}{1-(p+\lambda)} \left(\frac{\lambda}{1-p}\right)^i - \frac{p}{1-(p+\lambda)}, & 1 \le i \le i^*(p,\lambda), \\ 0, & i > i^*(p,\lambda), \end{cases}$$

where

$$i^*(p,\lambda) \triangleq \left\lfloor \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda} \right\rfloor, \tag{22}$$

Proof. The proof consists of simple algebra to compute the solution to $\mathbf{F}(\mathbf{v}^I) = 0$. The proof is given in Section 6.1. □

Case (4) in the above theorem is particularly interesting, as it reflects the system's heavy-traffic performance (λ close to 1) for any given value of p. Note that since $\mathbf{s}_1^I$ represents the probability of a typical queue having at least i tasks, the quantity

$$\mathbf{v}_1^I \triangleq \sum_{i=1}^{\infty} \mathbf{s}_i^I \tag{23}$$

represents the *average queue length*. The following corollary, which characterizes the average queue length in the invariant state for the fluid model, follows from Case (4) in Theorem 2 by some straightforward algebra.

Corollary 3 (Phase Transition in Average Queue Length Scaling). *If* $0 < p < \lambda$ *and* $\lambda \neq 1 - p$, *then*

$$\mathbf{v}_1^I \triangleq \sum_{i=1}^{\infty} \mathbf{s}_i^I = \frac{(1-p)(1-\lambda)}{(1-p-\lambda)^2} \left[1 - \left(\frac{\lambda}{1-p}\right)^{i^*(p,\lambda)}\right] - \frac{p}{1-p-\lambda} i^*(p,\lambda), \tag{24}$$

with $i^*(p,\lambda) = \left\lfloor \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda} \right\rfloor$. *In particular, for any fixed* $p > 0$, $\mathbf{v}_1^I$ *scales as*

$$\mathbf{v}_1^I \sim i^*(p,\lambda) \sim \log_{\frac{1}{1-p}} \frac{1}{1-\lambda}, \quad \text{as } \lambda \to 1. \tag{25}$$

The scaling of the average queue length in Eq. (25) with respect to the arrival rate λ is contrasted with (and is *exponentially better* than) the familiar $\frac{1}{1-\lambda}$ scaling when no centralized resource is available ($p = 0$).

[8]Here $\lfloor x \rfloor$ is defined as the largest integer that is less than or equal to x.

Intuition for exponential phase transition. Taking a closer look at the expressions for $\mathbf{s}^I$, we notice that that for any $p > 0$, the tail probabilities $\mathbf{s}^I$ have a **finite support**: $\mathbf{s}_i^I$ "dips" down to 0 as i increases to $i^*(p,\lambda)$, which is even faster than a super-geometric decay. Since $0 \le \mathbf{s}_i^I \le 1$ for all i, it is then intuitive that $\mathbf{v}_1^I = \sum_{i=1}^{i^*(p,\lambda)} \mathbf{s}_i^I$ is upper-bounded by $i^*(p,\lambda)$, which scales as $\log_{\frac{1}{1-p}} \frac{1}{1-\lambda}$ as $\lambda \to 1$. Note that such tail probabilities with finite-support imply that the fraction of stations with more than $i^*(p,\lambda)$ tasks *decreases to zero* as $N \to \infty$. For example, we may have a strictly positive fraction of stations with, say, 10 tasks, but stations with more than 10 tasks hardly exist. While this may appear counterintuitive, it is a direct consequence of centralization in the resource allocation schemes. Since a fraction p of the total resource is constantly going after the longest queues, it is able to prevent long queues (i.e., queues with more than $i^*(p,\lambda)$ tasks) from even appearing. The thresholds $i^*(p,\lambda)$ increasing to infinity as $\lambda \to 1$ reflects the fact that the central server's ability to annihilate long queues is compromised by the heavier traffic loads; our result essentially shows that the increase in $i^*(\lambda,p)$ is surprisingly slow.

Numerical results. Figure 3 compares the invariant state vectors for the case $p = 0$ (stars) and $p = 0.05$ (diamonds). When $p = 0$, $\mathbf{s}_i^I$ decays exponentially as λ^i, while when $p = 0.05$, $\mathbf{s}_i^I$ decays much faster, and reaches zero at around $i = 40$. Figure 4 demonstrates the exponential phase transition in the average queue length as the traffic intensity approaches 1, where the solid curve, corresponding to a positive p, increases significantly slower than the usual $\frac{1}{1-\lambda}$ delay scaling (dotted curve). Simulations show that the

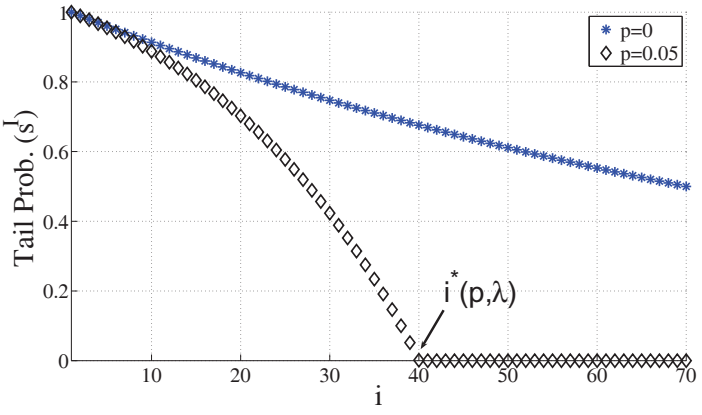

FIG 3. *Values of* $\mathbf{s}_i^I$*, as a function of* i*, for* $p = 0$ *and* $p = 0.05$*, with traffic intensity* $\lambda = 0.99$.

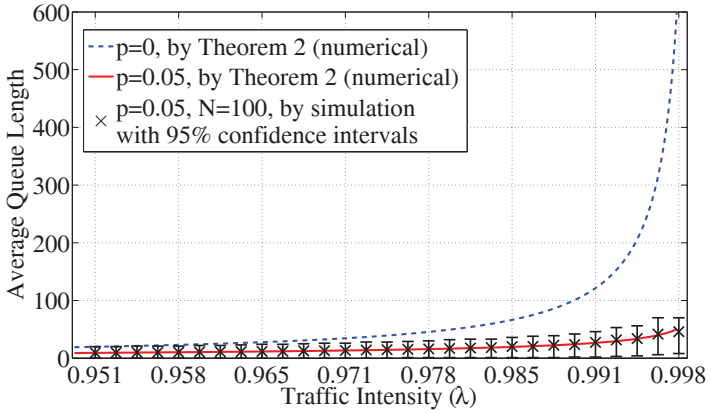

FIG 4. *Illustration of the exponential improvement in the average queue length, from* $O(\frac{1}{1-\lambda})$ *to* $O(\log \frac{1}{1-\lambda})$*, as* $\lambda \to 1$*, when we compare* $p = 0$ *to* $p = 0.05$.

TABLE 1
Values of $i^*(p,\lambda)$ *for various combinations of* (p,λ)

$p = \backslash\ \lambda =$	0.1	0.6	0.9	0.99	0.999
0.002	2	10	37	199	692
0.02	1	6	18	68	156
0.2	0	2	5	14	23
0.5	0	1	2	5	8
0.8	0	0	1	2	4

theoretical model offers good predictions for even a moderate number of servers ($N = 100$). The detailed simulation setup can be found in Appendix B. Table 1 gives examples of the values for $i^*(p,\lambda)$; note that these values in some sense correspond to the *maximum delay* an average customer could experience in the system.

Theorem 2 characterizes the invariant state of the fluid model, without revealing whether and how a solution of the fluid model reaches it. The next two results state that given any finite initial condition, the solution to the fluid model is unique and converges to the unique invariant state as time goes to infinity.

THEOREM 4 (Uniqueness of Solutions to Fluid Model). *Given any initial condition* $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$*, the fluid model has a unique solution* $\mathbf{v}(\mathbf{v}^0, t)$*,* $t \in [0,\infty)$.

PROOF. See Section 6.2. □

THEOREM 5 (Global Stability of Fluid Solutions). *Given any initial condition* $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$, *and with* $\mathbf{v}(\mathbf{v}^0, t)$ *the unique solution to the fluid model, we have*

$$\lim_{t\to\infty} \left\| \mathbf{v}\left(\mathbf{v}^0, t\right) - \mathbf{v}^I \right\|_w = 0, \tag{26}$$

where $\mathbf{v}^I$ *is the unique invariant state of the fluid model given in Theorem 2.*

PROOF. See Section 6.4. □

3.3. *Convergence to a fluid solution - finite horizon and steady state.* The two theorems in this section justify the use of the fluid model as an approximation to the finite stochastic system. The first theorem states that as $N \to \infty$ and with high probability, the evolution of the aggregate queue length process $\mathbf{V}^N(t)$ is uniformly close, over any finite time horizon $[0, T]$, to the unique solution of the fluid model.

THEOREM 6 (Convergence to Fluid Solutions over a Finite Horizon). *Consider a sequence of systems, with the number of servers N increasing to infinity. Fix any* $T > 0$. *If for some* $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$,

$$\lim_{N\to\infty} \mathbb{P}\left(\left\| \mathbf{V}^N(0) - \mathbf{v}^0 \right\|_w > \gamma \right) = 0, \quad \forall \gamma > 0, \tag{27}$$

then

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \left\| \mathbf{V}^N(t) - \mathbf{v}\left(\mathbf{v}^0, t\right) \right\|_w > \gamma \right) = 0, \quad \forall \gamma > 0. \tag{28}$$

where $\mathbf{v}\left(\mathbf{v}^0, t\right)$ *is the unique solution to the fluid model given initial condition* $\mathbf{v}^0$.

PROOF. See Section 6.3. □

Note that if we combine Theorem 6 with the convergence of $\mathbf{v}(t)$ to $\mathbf{v}^I$ in Theorem 5, we see that the finite system $(\mathbf{V}^N)$ is approximated by the invariant state of the fluid model $\mathbf{v}^I$ after a fixed time period. In other words, we now have

$$\lim_{t\to\infty} \lim_{N\to\infty} \mathbf{V}^N(t) = \mathbf{v}^I, \text{ in distribution.} \tag{29}$$

If we switch the order in which the limits over t and N are taken in Eq. (29), we are then dealing with the limiting behavior of the *sequence of steady-state*

distributions (if they exist) as the system size grows large. Indeed, in practice it is often of great interest to obtain a performance guarantee for the steady state of the system, if it were to run for a long period of time. In light of Eq. (29), we may expect that

$$\lim_{N\to\infty}\lim_{t\to\infty}\mathbf{V}^N(t)=\mathbf{v}^I, \text{ in distribution.} \tag{30}$$

The following theorem shows that this is indeed the case, i.e., that a unique steady-state distribution of $\mathbf{v}^N(t)$ (denoted by π^N) exists for all N, and that the sequence π^N concentrates on the invariant state of the fluid model ($\mathbf{v}^I$) as N grows large.

THEOREM 7 (Convergence of Steady-state Distributions to $\mathbf{v}^I$). *For any N, the process $\mathbf{V}^N(t)$ is positive recurrent and admits a unique steady-state distribution π^N.*[9] *Moreover,*

$$\lim_{N\to\infty}\pi^N=\delta_{\mathbf{v}^I}, \quad \textit{in distribution,} \tag{31}$$

where $\delta_{\mathbf{v}^I}$ is the Dirac measure concentrated on $\mathbf{v}^I$.

PROOF. The proof is based on the tightness of the sequence of steady-state distributions π^N, and a uniform rate of convergence of $\mathbf{V}^N(t)$ to $\mathbf{v}(t)$ over any compact set of initial conditions. The proof is given in Section 7. □

Figure 5 summarizes the relationships between the convergence to the solution of the fluid model over a finite time horizon (Theorem 6), the convergence of the fluid solution to the invariant state (Theorem 5), and the convergence of the sequence of steady-state distributions (Theorem 7).

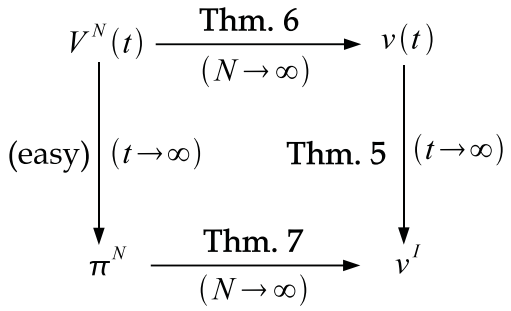

FIG 5. *Relationships between convergence results.*

[9]This probability distribution is defined on the Borel sets of $\overline{\mathcal{V}}^\infty$ under the topology induced by the metric $\|\cdot\|_w$.

4. Probability space and coupling. Starting from this section, the remainder of the paper will be devoted to proving the results summarized in Section 3. We begin by giving an outline of the main proof techniques, as well as the relationships among them, in Section 4.1. The remainder of the current section focuses on constructing probability spaces and appropriate couplings of stochastic sample paths, which will serve as the foundation for later analysis.

4.1. *Overview of technical approach.* We begin by coupling the sample paths of processes of interest (e.g., $\mathbf{V}^N(\cdot)$) with those of two fundamental processes that drive the system dynamics (Section 4.2). This approach allows us to link deterministically the convergence properties of the sample paths of interest to those of the fundamental processes, on which probabilistic arguments are easier to apply (such as the Functional Law of Large Numbers). Using this coupling framework, we show in Section 5 that almost all sample paths of $\mathbf{V}^N(\cdot)$ are "tight" in the sense that, as $N \to \infty$, they are uniformly approximated by a set of Lipschitz-continuous trajectories, which we refer to as the fluid limits, and that all such fluid limits are valid solutions to the fluid model. This result connects the finite stochastic system with the deterministic fluid solutions. Section 6 studies the properties of the fluid model, and provides proofs for Theorem 4 and 5. Note that Theorem 6 (convergence of $\mathbf{V}^N(\cdot)$ to the unique fluid solution, over a finite time horizon) now follows from the tightness results in Section 5 and the uniqueness of fluid solutions (Theorem 4). The proof of Theorem 2 stands alone, and will be given in Section 6.1. Finally, the proof of Theorem 7 (convergence of steady state distributions to $\mathbf{v}^I$) is given in Section 7.

The goal of the current section is to formally define the probability spaces and stochastic processes that we will be working with in the rest of the paper. Specifically, we begin by introducing two *fundamental processes*, from which all other processes of interest (e.g., $\mathbf{V}^N(\cdot)$) can be derived on a per sample path basis.

4.2. *Definition of probability space.*

Definition 8 (Fundamental Processes and Initial Conditions).

(1) **The Total Event Process**, $\{W(t)\}_{t\geq 0}$, defined on a probability space $(\Omega_W, \mathcal{F}_W, \mathbb{P}_W)$, is a Poisson process with rate $\lambda + 1$, where each jump marks the *time* when an "event" takes place in the system.
(2) **The Selection Process**, $\{U(n)\}_{n\in\mathbb{Z}_+}$, defined on a probability space $(\Omega_U, \mathcal{F}_U, \mathbb{P}_U)$, is a discrete-time process, where each $U(n)$ is independent

and uniformly distributed in $[0,1]$. This process, along with the current system state, determines the *type* of each event (i.e., whether it is an arrival, a local token generation, or a central token generation).

(3) **The (Finite) Initial Conditions**, $\{\mathbf{V}^{(0,N)}\}_{N\in\mathbb{N}}$, is a sequence of random variables defined on a common probability space $(\Omega_0,\mathcal{F}_0,\mathbb{P}_0)$, with $\mathbf{V}^{(0,N)}$ taking values[10] in $\overline{\mathcal{V}}^{\infty}\cap\mathcal{Q}^N$. Here, $\mathbf{V}^{(0,N)}$ represents the initial queue length distribution.

For the rest of the paper, we will be working with the product space

$$(\Omega,\mathcal{F},\mathbb{P}) \triangleq (\Omega_W\times\Omega_U\times\Omega_0,\mathcal{F}_W\times\mathcal{F}_U\times\mathcal{F}_0,\mathbb{P}_W\times\mathbb{P}_U\times\mathbb{P}_0). \tag{32}$$

With a slight abuse of notation, we use the same symbols $W(t)$, $U(n)$ and $\mathbf{V}^{(0,N)}$ for their corresponding *extensions* on Ω, i.e., $W(\omega,t)\triangleq W(\omega_W,t)$, where $\omega\in\Omega$ and $\omega=(\omega_W,\omega_U,\omega_0)$. The same holds for U and $\mathbf{V}^{(0,N)}$.

4.3. *A coupled construction of sample paths.* Recall the interpretation of the fluid model drift terms in Section 3.1. Mimicking the expression for $\dot{\mathbf{v}}_i(t)$ in Eq. (15), we would like to decompose $\mathbf{V}_i^N(t)$ into three non-decreasing right-continuous processes,

$$\mathbf{V}_i^N(t)=\mathbf{V}_i^N(0)+\mathbf{A}_i^N(t)-\mathbf{L}_i^N(t)-\mathbf{C}_i^N(t),\quad i\geq 1, \tag{33}$$

so that $\mathbf{A}_i^N(t)$, $\mathbf{L}_i^N(t)$, and $\mathbf{C}_i^N(t)$ correspond to the *cumulative changes* in $\mathbf{V}_i^N$ due to arrivals, local service tokens, and central service tokens, respectively. We will define processes $\mathbf{A}^N(t),\mathbf{L}^N(t)$, $\mathbf{C}^N(t)$, and $\mathbf{V}^N(t)$ on the common probability space $(\Omega,\mathcal{F},\mathbb{P})$, and *couple* them with the sample paths of the fundamental processes $W(t)$ and $U(n)$, and the value of $\mathbf{V}^{(0,N)}$, for each sample $\omega\in\Omega$. First, note that since the N-station system has N independent Poisson arrival streams, each with rate λ, and an exponential server with rate N, the total event process for this system is a Poisson process with rate $N(1+\lambda)$. Hence, we define $W^N(\omega,t)$, the Nth *normalized event process*, as

$$W^N(\omega,t)\triangleq\frac{1}{N}W(\omega,Nt),\quad \forall t\geq 0,\omega\in\Omega. \tag{34}$$

Note that $W^N(\omega,t)$ is normalized so that all of its jumps have a magnitude of $\frac{1}{N}$.

[10] For a finite system of N stations, the measure induced by $\mathbf{V}_i^N(t)$ is discrete and takes positive values only in the set of rational numbers with denominator N.

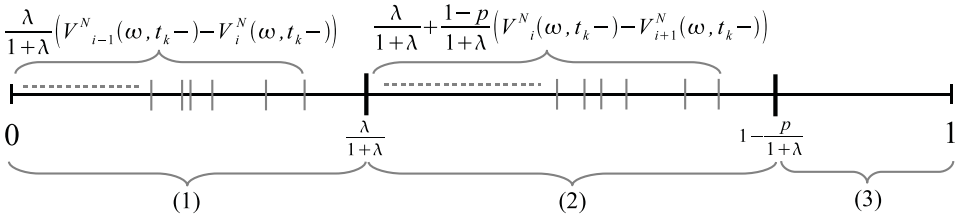

FIG 6. *Illustration of the partition of* $[0,1]$ *for constructing* $\mathbf{V}^N(\omega,\cdot)$.

The coupled construction is intuitive: whenever there is a jump in $W^N(\omega,\cdot)$, we decide the type of event by looking at the value of the corresponding selection variable $U(\omega,n)$ and the current state of the system $\mathbf{V}^N(\omega,t)$. Fix ω in Ω, and let $t_k, k \geq 1$, denote the time of the kth jump in $W^N(\omega,\cdot)$.

We first set all of $\mathbf{A}^N$, $\mathbf{L}^N$, and $\mathbf{C}^N$ to zero for $t \in [0,t_1)$. Starting from $k=1$, repeat the following steps for increasing values of k. The partition of the interval $[0,1]$ used in the procedure is illustrated in Figure 6.

(1) If $U(\omega,k) \in \frac{\lambda}{1+\lambda}\left[0, \mathbf{V}^N_{i-1}(\omega,t_k-) - \mathbf{V}^N_i(\omega,t_k-)\right)$ for some $i \geq 1$, the event corresponds to an **arrival** to a station with at least $i-1$ tasks. Hence we increase $\mathbf{A}^N_i(\omega,t)$ by $\frac{1}{N}$ at all such i.

(2) If $U(\omega,k) \in \frac{\lambda}{1+\lambda} + \frac{1-p}{1+\lambda}\left[0, \mathbf{V}^N_i(\omega,t_k-) - \mathbf{V}^N_{i+1}(\omega,t_k-)\right)$ for some $i \geq 1$, the event corresponds to the **completion** of a task at a station with at least i tasks due to a **local service token**. We increase $\mathbf{L}^N_i(\omega,t)$ by $\frac{1}{N}$ at all such i. Note that $i=0$ is *not* included here, reflecting the fact that if a local service token is generated at an empty station, it is immediately wasted and has no impact on the system.

(3) Finally, if $U(\omega,k) \in \frac{\lambda}{1+\lambda} + \frac{1-p}{1+\lambda} + \left[0, \frac{p}{1+\lambda}\right) = \left[1-\frac{p}{1+\lambda}, 1\right)$, the event corresponds to the generation of a **central service token**. Since the central service token is alway sent to a station with the longest queue length, we will have a task completion at a most-loaded station, unless the system is empty. Let $i^*(t)$ be the last positive coordinate of $\mathbf{V}^N(\omega,t-)$, i.e., $i^*(t) = \sup\{i : \mathbf{V}^N_i(\omega,t-) > 0\}$. We increase $\mathbf{C}^N_j(\omega,t)$ by $\frac{1}{N}$ for all j such that $1 \leq j \leq i^*(t_k)$.

To finish, we set $\mathbf{V}^N(\omega,t)$ according to Eq. (33), and keep the values of all processes unchanged between t_k and t_{k+1}. We set $\mathbf{V}^N_0 \triangleq \mathbf{V}^N_1 + 1$, so as to stay consistent with the definition of $\mathbf{V}^N_0$.

5. Fluid limits of stochastic sample paths. In this section, we establish the connections between the stochastic sample paths $(\mathbf{V}^N(\cdot))$ and the solutions to the fluid model $(\mathbf{v}(\cdot))$. Through two important technical results (Propositions 11 and 12), we show that, as $N \to \infty$ and almost surely,

any subsequence of $\{\mathbf{V}^N(\cdot)\}_{N\geq 1}$ contains a further subsequence that convergences uniformly to a solution of the fluid model, over any finite horizon $[0,T]$. However, note that the results presented in this section do not imply the converse, that *any* solution to the fluid model corresponds to a limit point of some sequence of stochastic sample paths. This issue will be resolved in the next section where we show the uniqueness of fluid solutions, which, together with the results in this section, establishes that the fluid solutions fully characterize the *transient behavior* of $\mathbf{V}^N(\cdot)$, for sufficiently large N, over any finite time horizon $[0,T]$.

In the sample-path wise construction in Section 4.3, all randomness is attributed to the initial condition $\mathbf{V}^{(0,N)}$ and the two fundamental processes $W(\cdot)$ and $U(\cdot)$. Everything else, including the system state $\mathbf{V}^N(\cdot)$ that we are interested in, can be derived from a deterministic mapping, given a particular realization of $\mathbf{V}^{(0,N)}$, $W(\cdot)$, and $U(\cdot)$. With this in mind, the approach we will take to prove convergence to a fluid limit (i.e., a limit point of $\{\mathbf{V}^N(\cdot)\}_{N\geq 1}$), over a finite time interval $[0,T]$, can be summarized as follows.

(1) (Lemma 9) We define a subset $\mathcal{C}$ of the sample space Ω, such that $\mathbb{P}(\mathcal{C}) = 1$ and the sample paths of W and U are sufficiently "nice" for every $\omega \in \mathcal{C}$.
(2) (Proposition 11) We show that for all ω in this nice set, the derived sample paths $\mathbf{V}^N(\cdot)$ are also "nice", and contain a subsequence converging to a Lipschitz-continuous trajectory $\mathbf{v}(\cdot)$, as $N \to \infty$.
(3) (Proposition 12) We characterize the derivative at any regular point[11] of $\mathbf{v}(\cdot)$ and show that it is identical to the drift in the fluid model. Hence $\mathbf{v}(\cdot)$ is a solution to the fluid model.

The proofs will be presented according to the above order.

5.1. *Tightness of sample paths over a nice set.* We begin by proving the following lemma which characterizes a "nice" set $\mathcal{C} \subset \Omega$ whose elements have desirable convergence properties.

LEMMA 9. *Fix $T > 0$. There exists a measurable set $\mathcal{C} \subset \Omega$ such that $\mathbb{P}(\mathcal{C}) = 1$ and for all $\omega \in \mathcal{C}$,*

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \left| W^N(\omega,t) - (1+\lambda)t \right| = 0, \tag{35}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{[a,b)}(U(\omega,i)) = b-a, \quad \text{if } a<b \text{ and } [a,b) \subset [0,1]. \tag{36}$$

[11] Regular points are points where derivative exists along *all coordinates* of the trajectory. Since the trajectory is Lipschitz-continuous along every coordinate, almost all points are regular.

PROOF. Based on the Functional Law of Large Numbers for Poisson processes, we can find $\mathcal{C}_W \subset \Omega_W$, with $\mathbb{P}_W(\mathcal{C}_W) = 1$, over which Eq. (35) holds. For Eq. (36), we invoke the Glivenko-Cantelli lemma[12], which states that the empirical measures of a sequence of i.i.d. random variables converge *uniformly* almost surely, i.e.,

$$\lim_{N\to\infty} \sup_{x\in[0,1]} \left| \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{[0,x)}(U(i)) - x \right| = 0, \quad \text{almost surely.} \tag{37}$$

This implies the existence of some $\mathcal{C}_U \subset \Omega_U$, with $\mathbb{P}_U(\mathcal{C}_U) = 1$, over which Eq. (36) holds. (This is stronger than the ordinary Strong Law of Large Numbers for i.i.d. uniform random variables on $[0,1]$, which states convergence for a *fixed* set $[0,x)$.) We finish the proof by taking $\mathcal{C} = \mathcal{C}_W \times \mathcal{C}_U \times \Omega_0$. □

DEFINITION 10. We call the 4-tuple, $\mathbf{X}^N \triangleq (\mathbf{V}^N, \mathbf{A}^N, \mathbf{L}^N, \mathbf{C}^N)$, the **$N$th system**. Note that all four components are infinite-dimensional processes.[13]

Consider the space of functions from $[0,T]$ to $\mathbb{R}$ that are right-continuous-with-left-limits (RCLL), denoted by $D[0,T]$, and let it be equipped with the uniform metric, $d(\cdot,\cdot)$:

$$d(x,y) \triangleq \sup_{t\in[0,T]} |x(t) - y(t)|, \quad x,y \in D[0,T]. \tag{38}$$

Denote by $D^\infty[0,T]$ the set of functions from $[0,T]$ to $\mathbb{R}^{\mathbb{Z}_+}$ that are RCLL on every coordinate. Let $d^{\mathbb{Z}_+}(\cdot,\cdot)$ denote the uniform metric on $D^\infty[0,T]$:

$$d^{\mathbb{Z}_+}(\mathbf{x},\mathbf{y}) \triangleq \sup_{t\in[0,T]} \|\mathbf{x}(t) - \mathbf{y}(t)\|_w, \quad \mathbf{x},\mathbf{y} \in D^{\mathbb{Z}_+}[0,T], \tag{39}$$

with $\|\cdot\|_w$ defined in Eq. (12).

The following proposition is the main result of this section. It shows that for sufficiently large N, the sample paths are sufficiently close to some absolutely continuous trajectory.

PROPOSITION 11. *Fix $T > 0$. Assume that there exists some $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$ such that*

$$\lim_{N\to\infty} \|\mathbf{V}^N(\omega,0) - \mathbf{v}^0\|_w = 0, \tag{40}$$

[12]For an introduction to the Glivenko-Cantelli lemma, see [27] and references therein.

[13]If necessary, $\mathbf{X}^N$ can be enumerated by writing it explicitly as $\mathbf{X}^N = (\mathbf{V}_0^N, \mathbf{A}_0^N, \mathbf{L}_0^N, \mathbf{C}_0^N, \mathbf{V}_1^N, \mathbf{A}_1^N, \ldots)$.

for all $\omega \in \mathcal{C}$*. Then for all* $\omega \in \mathcal{C}$*, any subsequence of* $\left\{\mathbf{X}^N(\omega,\cdot)\right\}$ *contains a further subsequence,* $\left\{\mathbf{X}^{N_i}(\omega,\cdot)\right\}$*, that converges to some coordinate-wise Lipschitz-continuous function* $\mathbf{x}(t) = (\mathbf{v}(t), \mathbf{a}(t), \mathbf{l}(t), \mathbf{c}(t))$*, with* $\mathbf{v}(0) = \mathbf{v}^0$*,* $\mathbf{a}(0) = \mathbf{l}(0) = \mathbf{c}(0) = 0$ *and*

$$|\mathbf{x}_i(a) - \mathbf{x}_i(b)| \le L|a-b|, \quad \forall a,b \in [0,T], \quad i \in \mathbb{Z}_+, \tag{41}$$

where $L > 0$ *is a universal constant, independent of the choice of* ω*,* $\mathbf{x}$*, and* T*. Here the convergence refers to* $d^{\mathbb{Z}_+}(\mathbf{V}^{N_i}, \mathbf{v})$*,* $d^{\mathbb{Z}_+}(\mathbf{A}^{N_i}, \mathbf{a})$*,* $d^{\mathbb{Z}_+}(\mathbf{L}^{N_i}, \mathbf{l})$*, and* $d^{\mathbb{Z}_+}(\mathbf{C}^{N_i}, \mathbf{c})$ *all converging to* 0*, as* $i \to \infty$*.*

For the rest of the paper, we will refer to such a limit point $\mathbf{x}$, or any subset of its coordinates, as a **fluid limit**.

Proof outline. Here we outline the main steps of the proof; interested readers are referred to Appendix A.1 for a complete proof. We first show that for all $\omega \in \mathcal{C}$, and for every coordinate i, any subsequence of $\left\{X_i^N(\omega,\cdot)\right\}$ has a convergent subsequence with a Lipschitz-continuous limit. We then use the coordinate-wise limit to construct a limit point in the space $D^{\mathbb{Z}_+}$. To establish coordinate-wise convergence, we use a tightness technique previously used in the literature on multiclass queueing networks (see, e.g., [1]). A key realization in this case, is that the total number of jumps in any derived process $\mathbf{A}^N$, $\mathbf{L}^N$, and $\mathbf{C}^N$ *cannot* exceed that of the event process $W^N(t)$ for any particular sample path. Since $\mathbf{A}^N$, $\mathbf{L}^N$, and $\mathbf{C}^N$ are non-decreasing, we expect their sample paths to be "smooth" for large N, due to the fact that the sample path of $W^N(t)$ does become "smooth" for large N, for all $\omega \in \mathcal{C}$ (Lemma 9). More formally, it can be shown that for all $\omega \in \mathcal{C}$ and $T > 0$, there exist diminishing positive sequences $M_N \downarrow 0$ and $\gamma_N \downarrow 0$, such that the sample path along any coordinate of $\mathbf{X}^N$ is γ_N-approximately-Lipschitz continuous with a uniformly bounded initial condition, i.e., for all i,

$$\left|\mathbf{X}_i^N(\omega,0) - x_i^0\right| \le M_N,$$
$$\text{and } \left|\mathbf{X}_i^N(\omega,a) - \mathbf{X}_i^N(\omega,b)\right| \le L|a-b| + \gamma_N, \quad \forall a,b \in [0,T],$$

where L is the Lipschitz constant, and $T < \infty$ is a fixed time horizon. Using a linear interpolation argument, we then show that sample paths of the above form can be uniformly approximated by a set of L-Lipschitz-continuous function on $[0,T]$. We finish by using the Arzela-Ascoli theorem (sequential compactness) along with closedness of this set, to establish the existence of a convergent further subsequence along any subsequence (compactness) and that any limit point must also be L-Lipschitz-continuous (closedness). This

completes the proof for coordinate-wise convergence. With the coordinate-wise limit points, we then use a diagonal argument involving nested subsequences to construct the limit points of $\mathbf{X}^N$ in the space $D^{\mathbb{Z}_+}[0,T]$, and this completes the proof. □

5.2. *Derivatives of the fluid limits.* The previous section established that any sequence of "good" sample paths ($\{\mathbf{X}^N(\omega,\cdot)\}$ with $\omega \in \mathcal{C}$) eventually stays close to some Lipschitz-continuous, and therefore absolutely continuous, trajectory. In this section, we will characterize the derivatives of $\mathbf{v}(\cdot)$ at all regular (differentiable) points of such limiting trajectories. We will show, as we expect, that they are the same as the drift terms in the fluid model (Definition 1). This means that all fluid limits of $\mathbf{V}^N(\cdot)$ are in fact solutions to the fluid model.

Proposition 12 (Fluid Limits and Fluid Model). *Fix $\omega \in \mathcal{C}$ and $T > 0$. Let $\mathbf{x}$ be a limit point of some subsequence of $\mathbf{X}^N(\omega,\cdot)$, as in Proposition 11. Let t be a point of differentiability of all coordinates of $\mathbf{x}$. Then, for all $i \in \mathbb{N}$,*

$$\dot{\mathbf{a}}_i(t) = \lambda\left(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)\right), \tag{42}$$

$$\dot{\mathbf{l}}_i(t) = (1-p)\left(\mathbf{v}_i(t) - \mathbf{v}_{i+1}(t)\right), \tag{43}$$

$$\dot{\mathbf{c}}_i(t) = g_i\left(\mathbf{v}(t)\right), \tag{44}$$

where g was defined in Eq. (16), with the initial condition $\mathbf{v}(0) = \mathbf{v}^0$ and boundary condition $\mathbf{v}_0(t) - \mathbf{v}_1(t) = 1, \forall t \in [0,T]$. In other words, all fluid limits of $\mathbf{V}^N(\cdot)$ are solutions to the fluid model.

Proof. We fix some $\omega \in \mathcal{C}$ and for the rest of this proof we will suppress the dependence on ω in our notation. The existence of Lipschitz-continuous limit points for the given $\omega \in \mathcal{C}$ is guaranteed by Proposition 11. Let $\{\mathbf{X}^{N_k}(\cdot)\}_{k=1}^{\infty}$ be a convergent subsequence such that $\lim_{k\to\infty} d^{\mathbb{Z}_+}(\mathbf{X}^{N_k}(\cdot), \mathbf{x}) = 0$. We now prove each of the three claims (Eqs. (42)–(44)) separately. The index i is always fixed unless otherwise stated.

Claim 1. $\dot{\mathbf{a}}_i(t) = \lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t))$. *Consider the sequence of trajectories $\{\mathbf{A}^{N_k}(\cdot)\}_{k=1}^{\infty}$. By construction, $\mathbf{A}_i^N(t)$ receives a jump of magnitude $\frac{1}{N}$ at time t if and only if an event happens at time t and the corresponding selection random variable, $U(\cdot)$, falls in the interval $\frac{\lambda}{1+\lambda}\left[0, \mathbf{V}_{i-1}^N(t-) - \mathbf{V}_i^N(t-)\right)$. Therefore, we can write, for any given $\epsilon > 0$,*

$$\mathbf{A}_i^{N_k}(t+\epsilon) - \mathbf{A}_i^{N_k}(t) = \frac{1}{N_k} \sum_{j=N_k W^{N_k}(t)}^{N_k W^{N_k}(t+\epsilon)} \mathbb{I}_{I_j}\left(U(j)\right), \tag{45}$$

where $I_j \triangleq \frac{\lambda}{1+\lambda}\left[0, \mathbf{V}_{i-1}^{N_k}(t_j^{N_k}-) - \mathbf{V}_i^{N_k}(t_j^{N_k}-)\right)$ *and* t_j^N *is defined to be the time of the* j*th jump in* $W^N(\cdot)$*, i.e.,*

$$t_j^N \triangleq \inf\left\{s \geq 0 : W^N(s) \geq \frac{j}{N}\right\}. \tag{46}$$

Note that by the definition of a fluid limit, we have that

$$\lim_{k\to\infty}\left(\mathbf{A}_i^{N_k}(t+\epsilon) - \mathbf{A}_i^{N_k}(t)\right) = \mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t). \tag{47}$$

The following lemma bounds the change in $\mathbf{a}_i(t)$ *on a small time interval.*

LEMMA 13. *Fix* i *and* t*. For all sufficiently small* $\epsilon > 0$

$$|\mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t) - \epsilon\lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t))| \leq 2\epsilon^2 L. \tag{48}$$

PROOF. While the proof involves heavy notation, it is based on the fact that $\omega \in \mathcal{C}$: using Lemma 9, Eq. (48) follows from Eq. (45) by applying the convergence properties of $W^N(t)$ (Eq. (35)) and $U(n)$ (Eq. (36)).

For the formal proof, fix some $\omega \in \mathcal{C}$. Also, fix $i \geq 1$, $t > 0$, and $\epsilon > 0$. Since the limiting function $\mathbf{x}$ is L-Lipschitz-continuous on all coordinates by Proposition 11, there exists a non-increasing sequence $\gamma_n \downarrow 0$ such that for all $s \in [t, t+\epsilon]$ and all sufficiently large k,

$$\mathbf{V}_j^{N_k}(s) \in \left[\mathbf{v}_j(t) - (\epsilon L + \gamma_{N_k}), \mathbf{v}_j(t) + (\epsilon L + \gamma_{N_k})\right), \quad j \in \{i-1, i, i+1\}. \tag{49}$$

The above leads to:[14]

$$\begin{aligned} &\left[0, \mathbf{V}_{i-1}^{N_k}(s) - \mathbf{V}_i^{N_k}(s)\right) \supset \left[0, [\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t) - 2(\epsilon L + \gamma_{N_k})]^+\right), \\ &\text{and } \left[0, \mathbf{V}_{i-1}^{N_k}(s) - \mathbf{V}_i^{N_k}(s)\right) \subset \left[0, \mathbf{v}_{i-1}(t) - \mathbf{v}_i(t) + 2(\epsilon L + \gamma_{N_k})\right), \end{aligned} \tag{50}$$

for all sufficiently large k.

Define the sequence of set-valued functions $\{\eta^n(t)\}$ as

$$\eta^n(t) \triangleq \frac{\lambda}{1+\lambda}\left[0, \mathbf{v}_{i-1}(t) - \mathbf{v}_i(t) + 2(\epsilon L + \gamma_n)\right). \tag{51}$$

Note that since $\gamma_n \downarrow 0$,

$$\eta^n(t) \supset \eta^{n+1}(t) \text{ and } \bigcap_{n=1}^{\infty} \eta^n(t) = \frac{\lambda}{1+\lambda}\left[0, \mathbf{v}_{i-1}(t) - \mathbf{v}_i(t) + 2\epsilon L\right]. \tag{52}$$

[14] Here $[x]^+ \triangleq \max\{0, x\}$.

We have for all sufficiently large k, and any l such that $1 \le l \le N_k$,

$$\begin{aligned}(53)\qquad & \mathbf{A}_i^{N_k}(t+\epsilon) - \mathbf{A}_i^{N_k}(t) \\ \le\ & \frac{1}{N_k} \sum_{j=N_k W^{N_k}(t)+1}^{N_k W^{N_k}(t+\epsilon)} \mathbb{I}_{\eta^{N_k}(t)}\left(U(j)\right) \\ \le\ & \frac{1}{N_k} \sum_{j=N_k W^{N_k}(t)+1}^{N_k W^{N_k}(t+\epsilon)} \mathbb{I}_{\eta^{l}(t)}\left(U(j)\right) \\ =\ & \frac{1}{N_k}\left(\sum_{j=1}^{N_k W^{N_k}(t+\epsilon)} \mathbb{I}_{\eta^{l}(t)}\left(U(j)\right) - \sum_{j=1}^{N_k W^{N_k}(t)} \mathbb{I}_{\eta^{l}(t)}\left(U(j)\right)\right),\end{aligned}$$

where the first inequality follows from the second containment in Eq. (50), and the second inequality follows from the monotonicity of $\{\eta^n(t)\}$ in Eq. (52).

We would like to show that for all sufficiently small $\epsilon > 0$,

$$(54)\qquad \mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t) - \epsilon\lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)) \le 2\epsilon^2 L.$$

To prove the above inequality, we first claim that for any interval $[a,b) \subset [0,1]$,

$$(55)\qquad \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{NW^N(t)} \mathbb{I}_{[a,b)}\left(U(i)\right) = (\lambda+1)t(b-a).$$

To see this, rewrite the left-hand side of the equation above as

$$\begin{aligned}(56)\qquad & \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{NW^N(t)} \mathbb{I}_{[a,b)}\left(U(i)\right) \\ =\ & \lim_{N\to\infty} (\lambda+1)t \frac{1}{(\lambda+1)Nt} \sum_{i=1}^{(\lambda+1)Nt} \mathbb{I}_{[a,b)}\left(U(i)\right) \\ & + \lim_{N\to\infty} (\lambda+1)t \frac{1}{(\lambda+1)Nt}\left(\sum_{i=1}^{NW^N(t)} \mathbb{I}_{[a,b)}\left(U(i)\right) - \sum_{i=1}^{(\lambda+1)Nt} \mathbb{I}_{[a,b)}\left(U(i)\right)\right).\end{aligned}$$

Because the magnitude of the indicator function $\mathbb{I}\{\cdot\}$ is bounded by 1, we have that

$$(57)\qquad \left| \sum_{i=1}^{NW^N(t)} \mathbb{I}_{[a,b)}\left(U(i)\right) - \sum_{i=1}^{(\lambda+1)Nt} \mathbb{I}_{[a,b)}\left(U(i)\right)\right| \le N\left|(\lambda+1)t - W^N(t)\right|.$$

Since $\omega \in \mathcal{C}$, by Lemma 9 we have that

$$\lim_{N\to\infty} \left|(\lambda+1)t - W^N(t)\right| = 0, \tag{58}$$

$$\lim_{N\to\infty} \frac{1}{(\lambda+1)Nt} \sum_{i=1}^{(\lambda+1)Nt} \mathbb{I}_{[a,b)}\left(U(i)\right) = b-a, \tag{59}$$

for any $t < \infty$. Combining Eqs. (56)–(59), we have that

$$\begin{aligned} &\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{W^N(t)} \mathbb{I}_{[a,b)}\left(U(i)\right) \\ &= (\lambda+1)t \lim_{N\to\infty} \frac{1}{(\lambda+1)Nt} \sum_{i=1}^{(\lambda+1)Nt} \mathbb{I}_{[a,b)}\left(U(i)\right) \\ &\quad + \lim_{N\to\infty} \frac{1}{(\lambda+1)t} \left|(\lambda+1)t - W^N(t)\right| \\ &= (\lambda+1)t(b-a), \end{aligned} \tag{60}$$

which establishes Eq. (55). By the same argument, Eq. (60) also holds when t is replaced by $t+\epsilon$. Applying this result to Eq. (53), we have that

$$\begin{aligned} &\mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t) \\ &= \lim_{k\to\infty} \left(\mathbf{A}_i^{N_k}(t+\epsilon) - \mathbf{A}_i^{N_k}(t)\right) \\ &\le (t+\epsilon-t)(\lambda+1)\frac{\lambda}{\lambda+1}\left[\mathbf{v}_i(t) - \mathbf{v}_{i-1}(t) + 2(\epsilon L + \gamma_l)\right] \\ &= \epsilon\lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)) + \lambda(2\epsilon^2 L + 2\epsilon\gamma_l) \\ &< \epsilon\lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)) + 2\epsilon^2 L + 2\epsilon\gamma_l, \end{aligned} \tag{61}$$

for all $l \ge 1$, where the last inequality is due to the fact that $\lambda < 1$. Taking $l \to \infty$ and using the fact that $\gamma_l \downarrow 0$, we have established Eq. (54).

Similarly, changing the definition of $\eta^n(t)$ to

$$\eta^n(t) \triangleq \frac{\lambda}{1+\lambda}\Big[0, \left[\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t) - 2(\epsilon L + \gamma_n)\right]^+\Big), \tag{62}$$

we can obtain a similar lower bound

$$\mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t) - \epsilon\lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)) \ge -2\epsilon^2 L, \tag{63}$$

which together with Eq. (54) proves the claim. Note that if $\mathbf{v}_i(t) = \mathbf{v}_{i-1}(t)$, the lower bound trivially holds because $\mathbf{A}_i^{N_k}(t)$ is a cumulative arrival process and is hence non-decreasing in t by definition. □

Since by assumption $\mathbf{a}(\cdot)$ is differentiable at t, Lemma 13 implies that $\dot{\mathbf{a}}_i(t) \stackrel{\Delta}{=} \lim_{\epsilon \downarrow 0} \frac{\mathbf{a}_i(t+\epsilon)-\mathbf{a}_i(t)}{\epsilon} = \lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t))$, which establishes Claim 1.

CLAIM 2. $\dot{\mathbf{l}}_i(t) = (1-p)(\mathbf{v}_i(t) - \mathbf{v}_{i+1}(t))$. *Claim 2 can be proved using an identical approach to the one used to prove Claim 1. The proof is hence omitted.*

CLAIM 3. $\dot{\mathbf{c}}_i(t) = g_i(\mathbf{v}(t))$. *We prove Claim 3 by considering separately the three cases in the definition of* $\mathbf{v}$.

(1) **Case 1**: $\dot{\mathbf{c}}_i(t) = 0$, if $\mathbf{v}_{i-1} = 0, \mathbf{v}_i = 0$. Write

$$\dot{\mathbf{c}}_i(t) = \dot{\mathbf{a}}_i(t) - \dot{\mathbf{l}}_i(t) - \dot{\mathbf{v}}_i(t). \tag{64}$$

We calculate each of the three terms on the right-hand side of the above equation. By Claim 1, $\dot{\mathbf{a}}_i(t) = \lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)) = 0$, and by Claim 2, $\dot{\mathbf{l}}_i(t) = \lambda(\mathbf{v}_i(t) - \mathbf{v}_{i+1}(t)) = 0$. To obtain the value for $\dot{\mathbf{v}}_i(t)$, we use the following trick: since $\mathbf{v}_i(t) = 0$ *and* $\mathbf{v}_i$ is *non-negative*, the *only* possibility for $\mathbf{v}_i(t)$ to be differentiable at t is that $\dot{\mathbf{v}}_i(t) = 0$. Since $\dot{\mathbf{a}}_i(t)$, $\dot{\mathbf{l}}_i(t)$, and $\dot{\mathbf{v}}_i(t)$ are all zero, we have that $\dot{\mathbf{c}}_i(t) = 0$.

(2) **Case 2**: $\dot{\mathbf{c}}_i(t) = \min\{\lambda \mathbf{v}_{i-1}(t), p\}$, if $\mathbf{v}_i(t) = 0, \mathbf{v}_{i-1}(t) > 0$.
In this case, the fraction of queues with at least i tasks is zero, hence $\mathbf{v}_i$ receives no drift from the local portion of the service capacity by Claim 2. First consider the case $\mathbf{v}_{i-1}(t) \le \frac{p}{\lambda}$. Here the line of arguments is similar to the one in Case 1. By Claim 1, $\dot{\mathbf{a}}_i(t) = \lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)) = \lambda \mathbf{v}_{i-1}(t)$, and by Claim 2, $\dot{\mathbf{l}}_i(t) = \lambda(\mathbf{v}_i(t) - \mathbf{v}_{i+1}(t)) = 0$. Using again the same trick as in Case 1, the non-negativity of $\mathbf{v}_i(t)$ and the fact that $\mathbf{v}_i(t) = 0$ together imply that we must have $\dot{\mathbf{v}}_i(t) = 0$. Combining the expressions for $\dot{\mathbf{a}}_i(t)$, $\dot{\mathbf{l}}_i(t)$, and $\dot{\mathbf{v}}_i(t)$, we have

$$\dot{\mathbf{c}}_i(t) = -\dot{\mathbf{v}}_i(t) + \dot{\mathbf{a}}_i(t) - \dot{\mathbf{l}}_i(t) = \lambda \mathbf{v}_{i-1}(t). \tag{65}$$

Intuitively, here the drift due to random arrivals to queues with $i-1$ tasks, $\lambda \mathbf{v}_{i-1}(t)$, is "absorbed" by the central portion of the service capacity.
If $\mathbf{v}_{i-1}(t) > \frac{p}{\lambda}$, then the above equation would imply that $\dot{\mathbf{c}}_i(t) = \lambda \mathbf{v}_{i-1}(t) > p$, if $\dot{\mathbf{c}}_i(t)$ exists. But clearly $\dot{\mathbf{c}}_i(t) \le p$. This simply means that if $\mathbf{v}_i(t) = 0, \mathbf{v}_{i-1}(t) > \frac{p}{\lambda}$, then $\mathbf{v}_i(t)$ cannot be differentiable at time t. Hence we have the claimed expression.

(3) **Case 3**: $\dot{\mathbf{c}}_i(t) = p$, if $\mathbf{v}_i > 0, \mathbf{v}_{i+1} > 0$.
Since there is a positive fraction of queues with more than i tasks, it follows that $\mathbf{V}_i^N$ is decreased by $\frac{1}{N}$ *whenever* a central token becomes

available. Formally, for some small enough ϵ, there exists K such that $\mathbf{V}_i^{N_k}(s) > 0$ for all $k \geq K$, $s \in [t, t+\epsilon]$. Given the coupling construction, this implies for all $k \geq K$, $s \in [t, t+\epsilon]$,

$$\mathbf{C}_i^{N_k}(s) - \mathbf{C}_i^{N_k}(t) = \frac{1}{N_k} \sum_{j=N_k W^{N_k}(t)}^{N_k W^{N_k}(s)} \mathbb{I}_{[1-\frac{p}{1+\lambda},1)}\left(U(j)\right).$$

Using the same arguments as in the proof of Lemma 13, we see that the right-hand side of the above equation converges to $(s-t)p + o(\epsilon)$ as $k \to \infty$. Hence, $\dot{\mathbf{c}}_i(t) = \lim_{\epsilon \downarrow 0} \lim_{k\to\infty} \frac{\mathbf{C}_i^{N_k}(t+\epsilon) - \mathbf{C}_i^{N_k}(t)}{\epsilon} = p$.

Finally, note that the boundary condition $\mathbf{v}_0(t) - \mathbf{v}_1(t) = 1$ is a consequence of the fact that $\mathbf{V}_0^N(t) - \mathbf{V}_1^N(t) \stackrel{\triangle}{=} \mathbf{S}_1^N(t) = 1$ for all t. Similarly, the boundary condition (14) is automatically satisfied. This concludes the proof of Proposition 12. □

6. Properties of the fluid model. In this section, we establish several key properties of the fluid model. We begin by proving Theorem 2 in Section 6.1, which states that the fluid model admits a unique invariant state for each pair of p and λ. Section 6.2 is devoted to proving that the fluid model admits a *unique* solution $\mathbf{v}(\cdot)$ for any initial condition $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$. As a corollary, we show that the fluid solution, $\mathbf{v}(\cdot)$, depends continuously on the initial condition $\mathbf{v}^0$, which will be used for proving the steady-state convergence theorem in the next section. Using the uniqueness of fluid solutions and the results from the last section, Theorem 6 is proved in Section 6.3, which establishes the convergence of stochastic sample paths to the unique solution of the fluid model over any finite time horizon, with high probability. Finally, in Section 6.4 we prove that the solutions to the fluid model are *globally stable* (Theorem 5): any fluid solution $\mathbf{v}(t)$ converges to the unique invariant state $\mathbf{v}^I$ as $t \to \infty$. This suggests that the qualitative properties derived from the invariant state $\mathbf{v}^I$ serve as a good approximation for the transient behavior of the system, as $t \to \infty$. We note that by the end of this section, we will have established all *transient approximation results*, which correspond to the path

$$\mathbf{V}^N(t) \stackrel{N\to\infty}{\longrightarrow} \mathbf{v}(t) \stackrel{t\to\infty}{\longrightarrow} \mathbf{v}^I, \tag{66}$$

as was illustrated in Figure 5 of Section 3. The other path in Figure 5, namely, the approximation of the steady-state distributions of $\mathbf{V}^N(\cdot)$ by $\mathbf{v}^I$, will be dealt with in the next section.

6.1. *Invariant state of the fluid model.* In this section we prove Theorem 2, which gives explicit expressions for the (unique) invariant state of the fluid model.

PROOF OF THEOREM 2. In this proof we will be working with both $\mathbf{v}^I$ and $\mathbf{s}^I$, with the understanding that $\mathbf{s}_i^I \triangleq \mathbf{v}_i^I - \mathbf{v}_{i+1}^I, \forall i \geq 0$. It can be verified that the expressions given in all three cases are valid invariant states, by checking that $\mathbf{F}(\mathbf{v}^I) = 0$. We show they are indeed unique.

First, note that if $p \geq \lambda$, then $\mathbf{F}_1(\mathbf{v}) < 0$ whenever $\mathbf{v}_1 > 0$. Since $\mathbf{v}_1^I \geq 0$, we must have $\mathbf{v}_1^I = 0$, which by the boundary conditions implies that all other $\mathbf{v}_i^I$ must also be zero. This proves case (2) of the theorem.

Now suppose that $0 < p < \lambda$. We will first prove case (4). We observe that $\mathbf{F}_1(\mathbf{v}) > 0$ whenever $\mathbf{v}_1 = 0$. Hence $\mathbf{v}_1^I$ must be positive. By Eq. (16) this implies that $g_1(\mathbf{v}^I) = p$. Substituting $g_1(\mathbf{v}^I)$ in Eq. (15), along with the boundary condition $\mathbf{v}_0^I - \mathbf{v}_1^I = \mathbf{s}_0^I = 1$, we have

$$0 = \lambda \cdot 1 - (1-p)\mathbf{s}_1^I - p, \tag{67}$$

which yields that $\mathbf{s}_1^I = \frac{\lambda - p}{1-p}$. Repeating the same argument, we obtain the recursion that $\mathbf{s}_i^I = \frac{\lambda \mathbf{s}_{i-1}^I - p}{1-p}$, for as long as $\mathbf{s}_i^I$ (and therefore, $\mathbf{v}_i^I$) remains positive. Combining this with the expression for $\mathbf{s}_1^I$, we have

$$\mathbf{s}_i^I = \frac{1-\lambda}{1-(p+\lambda)}\left(\frac{\lambda}{1-p}\right)^i - \frac{p}{1-(p+\lambda)}, \quad 1 \leq i \leq i^*(p,\lambda), \tag{68}$$

where $i^*(p,\lambda) \triangleq \left\lfloor \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda} \right\rfloor$ marks the last coordinate where $\mathbf{s}_i^I$ remains non-negative. This proves uniqueness of $\mathbf{s}_i^I$ up to $i \leq i^*(p,\lambda)$. We can then use the same argument as in case (2), to show that $\mathbf{s}_i^I$ must be equal to zero for all $i > i^*(p,\lambda)$. Cases (1) and (3) can be established using similar arguments as those used in proving case (4). This completes the proof. □

REMARK (a finite-support property of $\mathbf{v}(\cdot)$). As was discussed in Section 3.2, Theorem 2 shows that for all $p > 0$, the unique invariant state $\mathbf{v}^I$ admits a finite support (i.e., $\mathbf{v}_i^I = 0$ for all $i \geq i^*$ for some $i^* < \infty$). It turns out that, when $p > 0$, this finite-support property also holds for the fluid solution $\mathbf{v}(t)$ at all $t > 0$. For a more elaborate discussion on this topic, see Appendix C.

6.2. *Uniqueness of fluid limits & continuous dependence on initial conditions.* We now prove Theorem 4, which states that given an initial condition

$\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$, a solution to the fluid model exists and is unique. As a direct consequence of the proof, we obtain an important corollary, that the unique solution $\mathbf{v}(\cdot)$ depends *continuously* on the initial condition $\mathbf{v}^0$.

The uniqueness result justifies the use of the fluid approximation, in the sense that the evolution of the stochastic system is close to a *single* trajectory. The uniqueness along with the continuous dependence on the initial condition will be used to prove convergence of steady-state distributions to $\mathbf{v}^I$ (Theorem 7).

We note that, in general, the uniqueness of solutions is not guaranteed for a differential equation with a discontinuous drift (see, e.g., [22]). In our case, $\mathbf{F}(\cdot)$ is discontinuous on the domain $\overline{\mathcal{V}}^\infty$ due to the drift associated with central service tokens (Eq. (18)).

PROOF OF THEOREM 4. The existence of a solution to the fluid model follows from the fact that $\mathbf{V}^N$ has a limit point (Proposition 11) and that all limit points of $\mathbf{V}^N$ are solutions to the fluid model (Proposition 12). We now show uniqueness. Define $i^p(\mathbf{v}) \triangleq \sup\{i : \mathbf{v}_i > 0\}$.[15] Let $\mathbf{v}(t), \mathbf{w}(t)$ be two solutions to the fluid model such that $\mathbf{v}(0) = \mathbf{v}^0$ and $\mathbf{w}(0) = \mathbf{w}^0$, with $\mathbf{v}^0, \mathbf{w}^0 \in \overline{\mathcal{V}}^\infty$. At any regular point $t \geq 0$, where all coordinates of $\mathbf{v}(t), \mathbf{w}(t)$ are differentiable, without loss of generality, assume $i^p(\mathbf{v}(t)) \leq i^p(\mathbf{w}(t))$, with equality if both are infinite. Let $\mathbf{a}^{\mathbf{v}}(\cdot)$ and $\mathbf{a}^{\mathbf{w}}(\cdot)$ be the arrival trajectories corresponding to $\mathbf{v}(\cdot)$ and $\mathbf{w}(\cdot)$, respectively, and similarly for $\mathbf{l}$ and $\mathbf{c}$. Since $\mathbf{v}_0(t) = \mathbf{v}_1(t) + 1$ for all $t \geq 0$ by the boundary condition (Eq. (13)), and $\dot{\mathbf{v}}_1 = \dot{\mathbf{a}}_1^{\mathbf{v}} - \dot{\mathbf{l}}_1^{\mathbf{v}} - \dot{\mathbf{c}}_1^{\mathbf{v}}$, for notational convenience we will write

$$\dot{\mathbf{v}}_0 = \dot{\mathbf{a}}_0^{\mathbf{v}} - \dot{\mathbf{l}}_0^{\mathbf{v}} - \dot{\mathbf{c}}_0^{\mathbf{v}}, \tag{69}$$

where

$$\dot{\mathbf{a}}_0^{\mathbf{v}} \triangleq \dot{\mathbf{a}}_1^{\mathbf{v}}, \quad \dot{\mathbf{l}}_0^{\mathbf{v}} \triangleq \dot{\mathbf{l}}_1^{\mathbf{v}}, \quad \text{and } \dot{\mathbf{c}}_0^{\mathbf{v}} \triangleq \dot{\mathbf{c}}_1^{\mathbf{v}}. \tag{70}$$

The same notation will be used for $\dot{\mathbf{w}}(t)$.

We have,[16]

$$\begin{aligned} \frac{d}{dt}\|\mathbf{v}-\mathbf{w}\|_w^2 &\triangleq \frac{d}{dt}\sum_{i=0}^{\infty}\frac{|\mathbf{v}_i-\mathbf{w}_i|^2}{2^i} \overset{(a)}{=} \sum_{i=0}^{\infty}\frac{(\mathbf{v}_i-\mathbf{w}_i)(\dot{\mathbf{v}}_i-\dot{\mathbf{w}}_i)}{2^{i-1}} \\ &= \sum_{i=0}^{\infty}\frac{(\mathbf{v}_i-\mathbf{w}_i)\left[\left(\dot{\mathbf{a}}_i^{\mathbf{v}}-\dot{\mathbf{l}}_i^{\mathbf{v}}\right)-\left(\dot{\mathbf{a}}_i^{\mathbf{w}}-\dot{\mathbf{l}}_i^{\mathbf{w}}\right)\right]}{2^{i-1}} - \sum_{i=0}^{\infty}\frac{(\mathbf{v}_i-\mathbf{w}_i)(\dot{\mathbf{c}}_i^{\mathbf{v}}-\dot{\mathbf{c}}_i^{\mathbf{w}})}{2^{i-1}} \end{aligned} \tag{71}$$

[15] $i^p(\mathbf{v})$ can be infinite; this happens if all coordinates of $\mathbf{v}$ are positive.

[16] For notational convenience, we remove the dependence on t throughout this derivation, with the convention that all derivatives are taken with respect to t.

$$
\begin{aligned}
&\overset{(b)}{\leq} && C\left\|\mathbf{v}-\mathbf{w}\right\|_w^2 - \sum_{i=0}^{\infty} \frac{\left(\mathbf{v}_i-\mathbf{w}_i\right)\left(\dot{\mathbf{c}}_i^{\mathbf{v}}-\dot{\mathbf{c}}_i^{\mathbf{w}}\right)}{2^{i-1}} \\
&= && C\left\|\mathbf{v}-\mathbf{w}\right\|_w^2 - \sum_{i=0}^{i^p(\mathbf{v})} \frac{1}{2^{i-1}}\left(\mathbf{v}_i-\mathbf{w}_i\right)(p-p) \\
&&& -\frac{1}{2^{i^p(\mathbf{v})}}(0-\mathbf{w}_{i^p(\mathbf{v})+1})(\min\{\lambda \mathbf{v}_{i^p(\mathbf{v})}, p\}-p) \\
&&& -\sum_{i=i^p(\mathbf{v})+2}^{i^p(\mathbf{w})} \frac{1}{2^{i-1}}(0-\mathbf{w}_i)(0-p) \\
&&& -\sum_{j=i^p(\mathbf{w})+1}^{\infty} \frac{1}{2^{i-1}}(0-0)\left(\dot{\mathbf{c}}_i^{\mathbf{v}}-\dot{\mathbf{c}}_i^{\mathbf{w}}\right) \\
&\leq && C\left\|\mathbf{v}-\mathbf{w}\right\|_w^2,
\end{aligned}
$$

where $C = 6(\lambda + 1 - p)$. We first justify the existence of the derivative $\frac{d}{dt}\left\|\mathbf{v}-\mathbf{w}\right\|_w^2$ and the exchange of limits in (a). Because $\mathbf{v}_i(t)$ and $\mathbf{w}_i(t)$ are L-Lipschitz-continuous for all i, it follows that there exists $L' > 0$ such that for all i, $h(i,s) \overset{\triangle}{=} |\mathbf{v}_i(s) - \mathbf{w}_i(s)|^2$ is L'-Lipschitz-continuous in the second argument, within a small neighborhood around $s = t$. In other words,

$$
\left|\frac{h(i,t+\epsilon)-h(i,t)}{\epsilon}\right| \leq L' \tag{72}
$$

for all i and all sufficiently small ϵ. Then,

$$
\begin{aligned}
\frac{d}{dt}\left\|\mathbf{v}-\mathbf{w}\right\|_w^2 &= \lim_{\epsilon \downarrow 0} \sum_{i=0}^{\infty} 2^{-i} \frac{h(i,t+\epsilon)-h(i,t)}{\epsilon} \\
&= \lim_{\epsilon \downarrow 0} \int_{i \in \mathbb{Z}_+} \frac{h(i,t+\epsilon)-h(i,t)}{\epsilon}\, d\mu_{\mathbb{N}},
\end{aligned} \tag{73}
$$

where $\mu_{\mathbb{N}}$ is a measure on $\mathbb{Z}_+$ defined by $\mu_{\mathbb{N}}(i) = 2^{-i}, i \in \mathbb{Z}_+$. By Eq. (72) and the dominated convergence theorem, we can exchange the limit and integration in Eq. (73) and obtain

$$
\begin{aligned}
\frac{d}{dt}\left\|\mathbf{v}-\mathbf{w}\right\|_w^2 &= \lim_{\epsilon \downarrow 0} \int_{i \in \mathbb{Z}_+} \frac{h(i,t+\epsilon)-h(i,t)}{\epsilon}\, d\mu_{\mathbb{N}} \\
&= \int_{i \in \mathbb{Z}_+} \lim_{\epsilon \downarrow 0} \frac{h(i,t+\epsilon)-h(i,t)}{\epsilon}\, d\mu_{\mathbb{N}} \\
&= \sum_{i=0}^{\infty} \frac{\left(\mathbf{v}_i-\mathbf{w}_i\right)\left(\dot{\mathbf{v}}_i-\dot{\mathbf{w}}_i\right)}{2^{i-1}},
\end{aligned} \tag{74}
$$

which justifies step (a) in Eq. (71). Step (b) follows from the fact that $\dot{\mathbf{a}}$ and $\dot{\mathbf{l}}$ are both continuous and linear in $\mathbf{v}$ (see Eqs. (42)–(44)).

The specific value of C can be derived after some straightforward algebra, which we isolate in the following claim:

CLAIM 14.

$$\sum_{i=0}^{\infty} \frac{(\mathbf{v}_i - \mathbf{w}_i)\left[\left(\dot{\mathbf{a}}_i^{\mathbf{v}} - \dot{\mathbf{l}}_i^{\mathbf{v}}\right) - \left(\dot{\mathbf{a}}_i^{\mathbf{w}} - \dot{\mathbf{l}}_i^{\mathbf{w}}\right)\right]}{2^{i-1}} \leq 6(\lambda + 1 - p)\,\|\mathbf{v} - \mathbf{w}\|_w^2\,, \tag{75}$$

PROOF. See Appendix A.2. □

Now suppose that $\mathbf{v}(0) = \mathbf{w}(0)$. By Gronwall's inequality[17] and Eq. (71), we have

$$\|\mathbf{v}(t) - \mathbf{w}(t)\|_w^2 \leq \|\mathbf{v}(0) - \mathbf{w}(0)\|_w^2\, e^{Ct} = 0, \quad \forall t \in [0, \infty), \tag{76}$$

which establishes the uniqueness of fluid limits on $[0, \infty)$. □

The following corollary is an easy, but important, consequence of the uniqueness proof.

COROLLARY 15 (Continuous Dependence on Initial Conditions). *Denote by* $\mathbf{v}(\mathbf{v}^0, \cdot)$ *the unique solution to the fluid model given initial condition* $\mathbf{v}^0 \in \overline{\mathcal{V}}^{\infty}$. *If* $\mathbf{w}^n \in \overline{\mathcal{V}}^{\infty}$ *for all* n, *and* $\|\mathbf{w}^n - \mathbf{v}^0\|_w \to 0$ *as* $n \to \infty$, *then for all* $t \geq 0$,

$$\lim_{n\to\infty} \|\mathbf{v}(\mathbf{w}^n, t) - \mathbf{v}(\mathbf{v}^0, t)\|_w = 0. \tag{77}$$

PROOF. The continuity with respect to the initial condition is a direct consequence of the inequality in Eq. (76): if $\mathbf{v}(\mathbf{w}^n, \cdot)$ is a sequence of fluid solutions with initial conditions $\mathbf{w}^n \in \overline{\mathcal{V}}^{\infty}$ and if $\|\mathbf{w}^n - \mathbf{v}^0\|_w^2 \to 0$ as $N \to \infty$, then for any $t \in [0, \infty)$,

$$\left\|\mathbf{v}(\mathbf{v}^0, t) - \mathbf{v}(\mathbf{w}^n, t)\right\|_w^2 \leq \left\|\mathbf{v}^0 - \mathbf{w}^n\right\|_w^2 e^{Ct} \to 0, \quad \text{as } n \to \infty.$$

This completes the proof. □

REMARK ($\mathbf{v}(\cdot)$ versus $\mathbf{s}(\cdot)$, and the Uniqueness of Fluid Limits). As was mentioned in Section 2.2, we have chosen to work primarily with the aggregate queue length process, $\mathbf{V}^N(\cdot)$ (Eq. (2)), instead of the *normalized queue length process*, $\mathbf{S}^N(\cdot)$ (Eq. (3)). Recall that for any finite N, the

[17] We use a special case of Gronwall's inequality, which states that if $\dot{u}(t) \leq Cu(t)$ on $t \in [a, b]$, then $u(t) \leq u(a)e^{Ct}$ for all $t \in [a, b]$. For an introduction to Gronwall's inequality, see [28] and references therein.

two processes are related by simple transformations, namely, for all $i \geq 0$, $\mathbf{V}_i^N(t) \stackrel{\triangle}{=} \sum_{j=i}^{\infty} \mathbf{S}_j^N(t)$, and $\mathbf{S}_i^N(t) \stackrel{\triangle}{=} \mathbf{V}_i^N(t) - \mathbf{V}_{i+1}^N(t)$. Therefore, there seems to be no obvious reason to favor one representation over the other when N is finite. However, in the limit of $N \to \infty$, it turns out that the *fluid model* associated with $\mathbf{V}^N(\cdot)$ is much easier to work with in establishing uniqueness of fluid solutions (Theorem 4).

A key to the the proof of Theorem 4 is a contraction of the drifts (Eq. (71)), which, surprisingly, would have failed if we had used the the alternative state representation $\mathbf{s}_i(t) = \mathbf{v}_i(t) - \mathbf{v}_{i+1}(t)$. The uniqueness result should still hold, but the proof would be much more difficult. The intuitive reason is that the sum of the drifts of the $\mathbf{s}_i$'s provided by the centralized service remains *constant* as long as the system is non-empty; hence, by adding up all the coordinates of $\mathbf{s}_i$, we eliminate many of the drift discontinuities. The fact that such a simple linear transformation can greatly simplify the analysis of an otherwise much more complex dynamical system may be of independent interest.

A more elaborate discussion on this topic, along with a counterexample, is provided in Appendix B.

6.3. *Convergence to a fluid solution over a finite horizon.* We now prove Theorem 6.

Proof of Theorem 6. The proof follows from the sample-path tightness in Proposition 11 and the uniqueness of fluid limits from Theorem 4. By assumption, the sequence of initial conditions $\mathbf{V}^{(0,N)}$ converges to some $\mathbf{v}^0 \in \overline{\mathcal{V}}^{\infty}$, in probability. Since the space $\overline{\mathcal{V}}^{\infty}$ is separable and complete under the $\|\cdot\|_w$ metric, by Skorohod's representation theorem, we can find a probability space $(\Omega_0, \mathcal{F}_0, \mathbb{P}_0)$ on which $\mathbf{V}^{(0,N)} \to \mathbf{v}^0$ almost surely. By Proposition 11 and Theorem 4, for almost every $\omega \in \Omega$, any subsequence of $\mathbf{V}^N(\omega, t)$ contains a further subsequence that converges to the unique fluid limit $\mathbf{v}(\mathbf{v}^0, t)$ uniformly on any compact interval $[0, T]$. Therefore for all $T < \infty$,

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \left\| \mathbf{V}^N(\omega, t) - \mathbf{v}(\mathbf{v}^0, t) \right\|_w = 0, \quad \mathbb{P}\text{–almost surely}, \tag{78}$$

which implies convergence in probability, and Eq. (28) holds. □

6.4. *Convergence to the invariant state* $\mathbf{v}^I$. We will prove Theorem 5 in this section. We switch to the alternative state representation, $\mathbf{s}(t)$, where

$$\mathbf{s}_i(t) \stackrel{\triangle}{=} \mathbf{v}_{i+1}(t) - \mathbf{v}_i(t), \quad \forall i \geq 0, \tag{79}$$

to study the evolution of a fluid solution as $t \to \infty$. It turns out that a nice monotonicity property of the evolution of $\mathbf{s}(t)$ induced by the drift structure will help establish the convergence to the invariant state. We recall that $\mathbf{s}_0(t) = 1$ for all t, and note that for all points where $\mathbf{v}$ is differentiable,

$$\dot{\mathbf{s}}_i(t) = \dot{\mathbf{v}}_i(t) - \dot{\mathbf{v}}_{i+1}(t) = \lambda(\mathbf{s}_{i-1} - \mathbf{s}_i) - (1-p)(\mathbf{s}_i - \mathbf{s}_{i+1}) - g_i^s(\mathbf{s}),$$

for all $i \geq 1$, where $g_i^s(\mathbf{s}) \triangleq g_i(\mathbf{v}) - g_{i+1}(\mathbf{v})$. Throughout this section, we will use both representations $\mathbf{v}(t)$ and $\mathbf{s}(t)$ to refer to the *same* fluid solution, with their relationship specified in Eq. (79).

The approach we will be using is essentially a variant of the convergence proof given in [3]. The idea is to partition the space $\overline{\mathcal{S}}^\infty$ into dominating classes, and show that (i) dominance in initial conditions is preserved by the fluid model, and (ii) any solution $\mathbf{s}(t)$ to the fluid model with an initial condition that dominates, or is dominated by, the invariant state $\mathbf{s}^I$ converges to $\mathbf{s}^I$ as $t \to \infty$. Properties (i) and (ii) together imply the convergence of the fluid solution $\mathbf{s}(t)$ to $\mathbf{s}^I$, as $t \to \infty$, for any finite initial condition. It turns out that such dominance in $\mathbf{s}$ is much stronger than a similarly defined relation for $\mathbf{v}$. For this reason we do not use $\mathbf{v}$ but instead rely on $\mathbf{s}$ to establish the result.

Definition 16 (Coordinate-wise Dominance). For any $\mathbf{s}, \mathbf{s}' \in \overline{\mathcal{S}}^\infty$, we write

1. $\mathbf{s} \succeq \mathbf{s}'$ if $\mathbf{s}_i \geq \mathbf{s}'_i$, for all $i \geq 0$;
2. $\mathbf{s} \succ \mathbf{s}'$ if $\mathbf{s} \neq \mathbf{s}'$, $\mathbf{s} \succeq \mathbf{s}'$ and $\mathbf{s}_i > \mathbf{s}'_i$, for all $i \geq 1$ for which $\mathbf{s}'_i > 0$.[18]

The following lemma states that $\succeq$-dominance in initial conditions is preserved by the fluid model.

Lemma 17. *Let* $\mathbf{s}^1(\cdot)$ *and* $\mathbf{s}^2(\cdot)$ *be two solutions to the fluid model such that* $\mathbf{s}^1(0) \succeq \mathbf{s}^2(0)$. *Then* $\mathbf{s}^1(t) \succeq \mathbf{s}^2(t), \forall t \geq 0$.

Proof. By the continuous dependence of a fluid limit on its initial condition (Corollary 15), it suffices to verify that $\mathbf{s}^1(t) \succeq \mathbf{s}^2(t), \forall t \geq 0$, whenever $\mathbf{s}^1(0) \succ \mathbf{s}^2(0)$ (strictly dominated initial conditions).

Let t_1 be the first time that there exists a coordinate for which $\mathbf{s}^1(t)$ and $\mathbf{s}^2(t)$ are equal *and* positive:

$$t_1 \triangleq \inf\left\{t \geq 0 : \mathbf{s}^1(t_1) \neq \mathbf{s}^2(t_1), \mathbf{s}_i^1(t) = \mathbf{s}_i^2(t) > 0, \text{ for some } i \geq 1\right\}, \tag{80}$$

If $t_1 = \infty$, one of the following must be true:

[18] We need the condition $\mathbf{s} \neq \mathbf{s}'$ in order to rule out the case where $\mathbf{s} = \mathbf{s}' = 0$.

(1) $\mathbf{s}^1(t) \succ \mathbf{s}^2(t)$, for all $t \geq 0$, in which case the claim holds.
(2) $\mathbf{s}^1(t') = \mathbf{s}^2(t')$ at some $t' < \infty$. By the uniqueness of solutions, $\mathbf{s}^1(t) = \mathbf{s}^2(t)$ for all $t \geq t'$, in which case the claim also holds.

Hence, we assume $t_1 < \infty$. Let k be the smallest coordinate index such that $\mathbf{s}^1(t_1)$ and $\mathbf{s}^2(t_1)$ are equal at k, but differ on *at least one* of the two adjacent coordinates, $k-1$ and $k+1$:

$$k \triangleq \min\left\{ i \geq 0 : \mathbf{s}_i^1(t_1) = \mathbf{s}_i^2(t_1) > 0, \max_{j\in\{1,-1\}} \{\mathbf{s}_{i+j}^1(t_1) - \mathbf{s}_{i+j}^2(t_1)\} > 0 \right\} \tag{81}$$

Since $\mathbf{s}^1(t) \succ \mathbf{s}^2(t)$, at all regular points $t < t_1$ that are close enough to t_1,

$$\dot{\mathbf{s}}_k^1(t) - \dot{\mathbf{s}}_k^2(t) = \lambda(\mathbf{s}_{k-1}^1 - \mathbf{s}_{k-1}^2) - (1-p)(\mathbf{s}_{k+1}^2 - \mathbf{s}_{k+1}^1) - (g_k^s(\mathbf{s}^1) - g_k^s(\mathbf{s}^2)), \tag{82}$$

where

$$\begin{aligned} & g_k^s(\mathbf{s}^1) - g_k^s(\mathbf{s}^2) \\ \leq\ & 0 \cdot \mathbb{I}\left\{\mathbf{s}_{k+1}^2 > 0\right\} + \left[\left(p - \min\{p, \lambda \mathbf{s}_k^1\}\right) - \left(p - \min\{p, \lambda \mathbf{s}_k^2\}\right)\right] \cdot \mathbb{I}\left\{\mathbf{s}_{k+1}^2 = 0\right\} \\ =\ & 0, \end{aligned} \tag{83}$$

and where the last equality comes from the fact that $\mathbf{s}_k^1(t) = \mathbf{s}_k^2(t)$ by the definition of k. Because $\mathbf{s}^1(t)$ and $\mathbf{s}^2(t)$ are continuous functions of t in every coordinate, we can find a time $t_0 < t_1$ such that $\mathbf{s}_k^1(t_0) > \mathbf{s}_k^2(t_0)$ and

$$\dot{\mathbf{s}}_k^1(t) - \dot{\mathbf{s}}_k^2(t) > 0, \tag{84}$$

for all regular $t \in (t_0, t_1)$. Since $\mathbf{s}_k^1(t_1) - \mathbf{s}_k^2(t_1) = \mathbf{s}_k^1(t_0) - \mathbf{s}_k^2(t_0) + \int_{t_0}^{t_1} (\dot{\mathbf{s}}_k^1(t) - \dot{\mathbf{s}}_k^2(t))\, dt$, this contradicts the fact that $\mathbf{s}_k^1(t_1) = \mathbf{s}_k^2(t_1)$, and hence proves the claim. □

We are now ready to prove Theorem 5.

Proof of Theorem 5. Let $\mathbf{s}(\cdot)$, $\mathbf{s}^u(\cdot)$, and $\mathbf{s}^l(\cdot)$ be three fluid limits with initial conditions in $\overline{\mathcal{S}}^\infty$ such that $\mathbf{s}^u(0) \geq \mathbf{s}(0) \geq \mathbf{s}^l(0)$ and $\mathbf{s}^u(0) \geq \mathbf{s}^I \geq \mathbf{s}^l(0)$. By Lemma 17, we must have $\mathbf{s}^u(t) \geq \mathbf{s}^I \geq \mathbf{s}^l(t)$ for all $t \geq 0$. Hence it suffices to show that $\lim_{t\to\infty} \left\|\mathbf{s}^u(t) - \mathbf{s}^I\right\|_w = \lim_{t\to\infty} \left\|\mathbf{s}^l(t) - \mathbf{s}^I\right\|_w = 0$. Recall that for any regular $t > 0$,

$$\begin{aligned} \dot{\mathbf{v}}_i(t) &= \lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t)) - (1-p)(\mathbf{v}_i(t) - \mathbf{v}_{i+1}(t)) - g_i(\mathbf{v}(t)) \\ &= \lambda \mathbf{s}_{i-1}(t) - (1-p)\mathbf{s}_i(t) - g_i(\mathbf{v}(t)) \\ &= (1-p)\left(\frac{\lambda \mathbf{s}_{i-1}(t) - g_i(\mathbf{v}(t))}{1-p} - \mathbf{s}_i\right). \end{aligned} \tag{85}$$

Recall, from the expressions for $\mathbf{s}_i^I$ in Theorem 2, that $\mathbf{s}_{i+1}^I \geq \frac{\lambda \mathbf{s}_i^I - p}{1-p}$, $\forall i \geq 0$. From Eq. (85) and the fact that $\mathbf{s}_0^u = \mathbf{s}_0^I = 1$, we have

$$\dot{\mathbf{v}}_1^u(t) = (1-p)\left(\frac{\lambda - g_1(\mathbf{v}^u(t))}{1-p} - \mathbf{s}_1^u(t)\right) \leq (1-p)\left(\mathbf{s}_1^I - \mathbf{s}_1^u(t)\right), \tag{86}$$

for all regular $t \geq 0$. To see why the above inequality holds, note that

$$\frac{\lambda - g_1(\mathbf{v}^u(t))}{1-p} = \frac{\lambda - p}{1-p} \leq \mathbf{s}_1^I, \tag{87}$$

whenever $\mathbf{s}_1^u(t) > 0$, and

$$\frac{\lambda - g_1(\mathbf{v}^u(t))}{1-p} = \mathbf{s}_1^u(t) = 0, \tag{88}$$

whenever $\mathbf{s}_1^u(t) = \mathbf{s}_1^I = 0$.

We argue that Eq. (86) implies that

$$\lim_{t\to\infty} \left|\mathbf{s}_1^I - \mathbf{s}_1^u(t)\right| = 0. \tag{89}$$

To show this, we use the following technical lemma. The proof is elementary and is omitted.

LEMMA 18. *Let $f(t) : [0,\infty) \to \mathbb{R}$ be non-negative and L-Lipschitz continuous. If $f(t)$ is integrable, then $\lim_{t\to\infty} f(t) = 0$.*

Let $f_1(t) \triangleq \mathbf{s}_1^u(t) - \mathbf{s}_1^I$. By Eq.(86), we have that $\dot{\mathbf{v}}_1^u(t) = -(1-p)f_1(t)$. Since $\mathbf{v}_1^u(t)$ is non-negative for all t, we have that $f_1(t)$ is integrable. We then invoke Lemma 18, which yields that

$$\lim_{t\to\infty} \left|\mathbf{s}_1^I - \mathbf{s}_1^u(t)\right| = \lim_{t\to\infty} f_1(t) = 0. \tag{90}$$

We now continue by induction. Let $f_i(t) \triangleq \mathbf{s}_i^u(t) - \mathbf{s}_i^I$, and suppose that $f_i(t)$ is integrable. By Eq. (85), we have that

$$\begin{aligned} \dot{\mathbf{v}}_{i+1}^u(t) &= (1-p)\left(\frac{\lambda \mathbf{s}_i^u(t) - g_i(\mathbf{v}^u(t))}{1-p} - \mathbf{s}_{i+1}^u(t)\right) \\ &= (1-p)\left[\frac{\lambda \mathbf{s}_i^I - g_i(\mathbf{v}^u(t))}{1-p} - \mathbf{s}_{i+1}^u(t) + \frac{\lambda}{1-p}\left(\mathbf{s}_i^u(t) - \mathbf{s}_i^I\right)\right] \\ &\leq -(1-p)f_{i+1}(t) + \lambda f_i(t), \end{aligned} \tag{91}$$

where the last inequality follows by an argument parallel to the one used in Eqs. (86)–(88). Since $\mathbf{v}^u_{i+1}(t)$ is non-negative for all t, by Eq. (91), we have that $[-(1-p)f_{i+1}(t) + \lambda f_i(t)]$ is integrable. Since $f_i(t)$ is integrable by the induction hypothesis, this implies that $f_{i+1}(t)$ is also integrable. Now, we again invoke Lemma 18, and we have that

$$\lim_{t\to\infty} |\mathbf{s}^u_{i+1}(t) - \mathbf{s}^I_{i+1}| = \lim_{t\to\infty} f_{i+1}(t) = 0. \tag{92}$$

This establishes the convergence of $\mathbf{s}^u(t)$ to $\mathbf{s}^I$ along all coordinates. Using also the fact that $0 \le \mathbf{s}^u_i(t) \le 1$ for all i and t, it is not hard to show that this coordinate-wise convergence also implies that

$$\lim_{t\to\infty} \left\|\mathbf{s}^u(t) - \mathbf{s}^I\right\|_w = 0. \tag{93}$$

Using the same set of arguments, we can show that $\lim_{t\to\infty} \left\|\mathbf{s}^l(t) - \mathbf{s}^I\right\|_w = 0$. This completes the proof. □

7. Convergence of steady-state distributions. We will prove Theorem 7 in this section, which states that, for all N, the Markov process $\mathbf{V}^N(t)$ converges to a unique steady-state distribution, π^N, as $t \to \infty$, and that the sequence $\{\pi^N\}_{N\ge1}$ concentrates on the unique invariant state of the fluid model, $\mathbf{v}^I$, as $N \to \infty$. This result is of practical importance, as it guarantees that key quantities, such as the average queue length, as derived from the expressions for $\mathbf{v}^I$, also serve as accurate approximations for the steady state of the actual (finite) stochastic system.

Note that by the end of this section, we will have established our *steady-state approximation* results, i.e.,

$$\mathbf{V}^N(t) \overset{t\to\infty}{\longrightarrow} \pi^N \overset{N\to\infty}{\longrightarrow} \mathbf{v}^I, \tag{94}$$

as illustrated in Figure 5 of Section 3. Together with the transient approximation results established in the previous sections, these conclude the proofs of all approximation theorems in this paper.

Before proving Theorem 7, we first give an important proposition which strengthens the finite-horizon convergence result stated in Theorem 6: we establish a uniform speed of convergence over any compact set of initial conditions. This proposition will be critical to the proof of Theorem 7 which will appear later in the section.

7.1. *Uniform rate of convergence to the fluid limit.* Let the probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ be the product space of $(\Omega_W, \mathcal{F}_W, \mathbb{P}_W)$ and $(\Omega_U, \mathcal{F}_U, \mathbb{P}_U)$.

Intuitively, $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ captures all exogenous arrival and service information. Fixing $\omega_1 \in \Omega_1$ and $\mathbf{v}^0 \in \overline{\mathcal{V}}^M \cap \mathcal{Q}^N$, denote by $\mathbf{V}^N(\mathbf{v}^0, \omega_1, t)$ the resulting sample path of $\mathbf{V}^N$ given the initial condition $\mathbf{V}^N(0) = \mathbf{v}^0$. Also, denote by $\mathbf{v}\left(\mathbf{v}^0, t\right)$ the solution to the fluid model for a given initial condition $\mathbf{v}^0$. We have the following proposition.

PROPOSITION 19 (Uniform Rate of Convergence to the Fluid Limit). *Fix $T > 0$ and $M \in \mathbb{N}$. Let $K^N \triangleq \overline{\mathcal{V}}^M \cap \mathcal{Q}^N$. We have*

$$\lim_{N\to\infty} \sup_{\mathbf{v}^0 \in K^N} d^{\mathbb{Z}_+}\left(\mathbf{V}^N(\mathbf{v}^0, \omega_1, \cdot), \mathbf{v}(\mathbf{v}^0, \cdot)\right) = 0, \quad \mathbb{P}_1\text{-almost surely}, \tag{95}$$

where the metric $d^{\mathbb{Z}_+}(\cdot,\cdot)$ was defined in Eq. (39).

PROOF. The proof highlights the convenience of the sample-path based approach. By the same argument as in Lemma 9, we can find sets $\mathcal{C}_W \subset \Omega_W$ and $\mathcal{C}_U \subset \Omega_U$ such that the convergence in Eqs. (35) and (36) holds over $\mathcal{C}_W$ and $\mathcal{C}_U$, respectively, and such that $\mathbb{P}_W(\mathcal{C}_W) = \mathbb{P}_U(\mathcal{C}_U) = 1$. Let $\mathcal{C}_1 \triangleq \mathcal{C}_W \times \mathcal{C}_U$. Note that $\mathbb{P}_1(\mathcal{C}_1) = 1$.

To prove the claim, it suffices to show that

$$\lim_{N\to\infty} \sup_{\mathbf{v}^0 \in K^N} d^{\mathbb{Z}_+}\left(\mathbf{V}^N(\mathbf{v}^0, \omega_1, \cdot), \mathbf{v}(\mathbf{v}^0, \cdot)\right) = 0, \quad \forall \omega_1 \in \mathcal{C}_1. \tag{96}$$

We start by assuming that the above convergence fails for some $\tilde{\omega}_1 \in \mathcal{C}_1$, which amounts to having a sequence of "bad" sample paths of $\mathbf{V}^N$ that are always a positive distance away from the corresponding fluid solution with the same initial condition, as $N \to \infty$. We then find nested subsequences within this sequence of bad sample paths, and construct two solutions to the fluid model with the *same* initial condition, contradicting the uniqueness of fluid model solutions.

Assume that there exists $\tilde{\omega}_1 \in \mathcal{C}_1$ such that

$$\limsup_{N\to\infty} \sup_{\mathbf{v}^0 \in K^N} d^{\mathbb{Z}_+}\left(\mathbf{V}^N(\mathbf{v}^0, \tilde{\omega}_1, \cdot), \mathbf{v}(\mathbf{v}^0, \cdot)\right) > 0. \tag{97}$$

This implies that there exist $\epsilon > 0$, $\{N_i\}_{i=1}^{\infty} \subset \mathbb{N}$, and $\left\{\mathbf{v}^{(0,N_i)}\right\}_{i=1}^{\infty}$ with $\mathbf{v}^{(0,N_i)} \in K^{N_i}$, such that

$$d^{\mathbb{Z}_+}\left(\mathbf{V}^N(\mathbf{v}^{(0,N_i)}, \tilde{\omega}_1, \cdot), \mathbf{v}(\mathbf{v}^{(0,N_i)}, \cdot)\right) > \epsilon, \tag{98}$$

for all $i \in \mathbb{N}$. We make the following two observations:

1. The set $\overline{\mathcal{V}}^M$ is closed and bounded, and the fluid solutions $\mathbf{v}(\mathbf{v}^{(0,N_i)},\cdot)$ are L-Lipschitz-continuous for all i. Hence the sequence of functions $\{\mathbf{v}(\mathbf{v}^{(0,N_i)},\cdot)\}_{i=1}^{\infty}$ are equicontinuous and uniformly bounded on $[0,T]$. We have by the Arzela-Ascoli theorem that there exists a subsequence $\left\{N_i^2\right\}_{i=1}^{\infty}$ of $\left\{N_i^1\right\}_{i=1}^{\infty}$ such that

$$d^{\mathbb{Z}_+}\left(\mathbf{v}\left(\mathbf{v}^{(0,N_i^2)},\cdot\right),\tilde{\mathbf{v}}^a(\cdot)\right)\to 0, \tag{99}$$

as $i\to\infty$, for some Lipschitz-continuous function $\tilde{\mathbf{v}}^a(\cdot)$ with $\tilde{\mathbf{v}}^a(0)\in\overline{\mathcal{V}}^M$. By the *continuous dependence of fluid solutions on initial conditions* (Corollary 15), $\tilde{\mathbf{v}}^a(\cdot)$ must be the unique solution to the fluid model with initial condition $\tilde{\mathbf{v}}^a(0)$, i.e.,

$$\tilde{\mathbf{v}}^a(t)=\mathbf{v}\left(\tilde{\mathbf{v}}^a(0),t\right),\quad \forall t\in[0,T]. \tag{100}$$

2. Since $\omega_1\in\mathcal{C}_1$, by Propositions 11 and 12, there exists a further subsequence $\left\{N_i^3\right\}_{i=1}^{\infty}$ of $\left\{N_i^2\right\}_{i=1}^{\infty}$ such that $\mathbf{V}^{N_i^3}\left(\mathbf{v}^{(0,N_i^3)},\tilde{\omega}_1,\cdot\right)\to\tilde{\mathbf{v}}^b(\cdot)$ uniformly over $[0,T]$ as $i\to\infty$, where $\tilde{\mathbf{v}}^b(\cdot)$ is a solution to the fluid model. Note that since $\left\{N_i^3\right\}_{i=1}^{\infty}\subset\left\{N_i^2\right\}_{i=1}^{\infty}$, we have $\tilde{\mathbf{v}}^b(0)=\tilde{\mathbf{v}}^a(0)$. Hence,

$$\tilde{\mathbf{v}}^b(t)=\mathbf{v}\left(\tilde{\mathbf{v}}^a(0),t\right),\quad \forall t\in[0,T]. \tag{101}$$

By the definition of $\tilde{\omega}_1$ (Eq. (97)) and the fact that $\tilde{\omega}_1\in\mathcal{C}_1$, we must have $\sup_{t\in[0,T]}\left\|\tilde{\mathbf{v}}^a(t)-\tilde{\mathbf{v}}^b(t)\right\|_w>\epsilon$, which, in light of Eqs. (100) and (101), contradicts the uniqueness of the fluid limit (Theorem 4). This completes the proof. □

The following corollary, stated in terms of convergence in probability, follows directly from Proposition 19. The proof is straightforward and is omitted.

COROLLARY 20. *Fix $T>0$ and $M\in\mathbb{N}$. Let $K^N\triangleq\overline{\mathcal{V}}^M\cap\mathcal{Q}^N$. Then, for all $\delta>0$,*

$$\lim_{N\to\infty}\mathbb{P}_1\left(\omega_1\in\Omega_1:\sup_{\mathbf{v}^0\in K^N}d^{\mathbb{Z}_+}\left(\mathbf{V}^N\left(\mathbf{v}^0,\omega_1,\cdot\right),\mathbf{v}(\mathbf{v}^0,\cdot)\right)>\delta\right)=0. \tag{102}$$

7.2. *Proof of Theorem 7.* We first state a tightness result that will be needed in the proof of Theorem 7.

PROPOSITION 21. *For every $N < \infty$ and $p \in (0,1]$, $\mathbf{V}^N(t)$ is positive-recurrent and $\mathbf{V}^N(t)$ converges in distribution to a unique steady-state distribution $\pi^{N,p}$ as $t \to \infty$. Furthermore, the sequence $\{\pi^{N,p}\}_{N=1}^{\infty}$ is tight, in the sense that for any $\epsilon > 0$, there exists $M > 0$ such that*

$$\pi^{N,p}\left(\overline{\mathcal{V}}^M\right) \triangleq \pi^{N,p}\left(\mathbf{V}_1^N \le M\right) \ge 1-\epsilon, \quad \forall N \ge 1. \tag{103}$$

PROOF SKETCH. The proposition is proved using a stochastic dominance argument, by coupling with the case $p = 0$. While the notation may seem heavy, the intuition is simple: when $p = 0$, the system degenerates into a collection of $M/M/1$ queues with independent arrivals and departures (but possibly correlated initial queue lengths), and it is easy to show that the system is positive recurrent and that the resulting sequence of steady-state distributions is tight as $N \to \infty$. The bulk of the proof is to formally argue that when $p > 0$, the system behaves "no worse" than when $p = 0$ in terms of positive recurrence and tightness of steady-state distributions. See Appendix A.3 for a complete proof using this stochastic dominance approach. □

REMARK. It is worth mentioning that the tightness of $\pi^{N,p}$ could alternatively be established by defining a Lyapunov function on $\overline{\mathcal{V}}^N$ and checking its drift with respect to the underlying embedded discrete-time Markov chain. By applying the Foster-Lyapunov stability criterion, one should be able to prove positive recurrence of $\mathbf{V}^N$ and give an explicit upper-bound on the expected value of $\mathbf{V}_1^N$ in steady state.[19] If this expectation is bounded as $N \to \infty$, we will have obtained the desirable result by the Markov inequality. We do not pursue this direction in this paper, because we believe that the stochastic dominance approach adopted here provides more insight by exploiting the monotonicity in p of the steady-state queue length distribution.

PROOF OF THEOREM 7. For the rest of the proof, since p is fixed, we will drop p in the super-script of $\pi^{N,p}$. By Proposition 21, the sequence of distributions π^N is tight, in the sense that for any $\epsilon > 0$, there exists $M(\epsilon) \in \mathbb{N}$ such that for all $M \ge M(\epsilon)$, $\pi^N\left(\overline{\mathcal{V}}^M\right) \ge 1-\epsilon$, for all N. (This is the same as the usual notion of tightness, because the set $\overline{\mathcal{V}}^M$ is compact.)

The rest of the proof is based on a classical technique based on continuous test functions (see Section 4 of [2]). The continuous dependence on initial conditions and the uniform rate of convergence established previously will

[19] For an overview of the use of the Foster-Lyapunov criterion in proving stability of queueing networks, see, e.g., [25].

be used here. Let $\overline{C}$ be the space of bounded continuous functions from $\overline{\mathcal{V}}^{\infty}$ to $\mathbb{R}$. Define the mappings $T^N(t)$ and $T(t)$ on $\overline{C}$ by:

$$\begin{aligned}\left(T^N(t)f\right)(\mathbf{v}^0) &\triangleq \mathbb{E}\left[f\left(\mathbf{V}^N(t)\right) \mid \mathbf{V}^N(0)=\mathbf{v}^0\right],\\ \left(T(t)f\right)(\mathbf{v}^0) &\triangleq \mathbb{E}\left[f\left(\mathbf{v}(t)\right) \mid \mathbf{v}(0)=\mathbf{v}^0\right] = f(\mathbf{v}(\mathbf{v}^0,t)), \text{ for } f\in\overline{C}.\end{aligned}$$

With this notation, π^N being a steady-state distribution for the Markov process $\mathbf{V}^N(t)$ is equivalent to having for all $t\geq 0$, $f\in\overline{C}$,

$$\int_{\mathbf{v}^0\in\overline{\mathcal{V}}^{\infty}\cap\mathcal{Q}^N} (T^N(t)f)(\mathbf{v}^0)\,d\pi^N = \int_{\mathbf{v}^0\in\overline{\mathcal{V}}^{\infty}\cap\mathcal{Q}^N} f(\mathbf{v}^0)\,d\pi^N. \tag{104}$$

Since $\{\pi^N\}$ is tight, it is sequentially compact under the topology of weak convergence, by Prokhorov's theorem. Let π be the weak limit of some subsequence of $\{\pi^N\}$. We will show that for all $t\geq 0$, $f\in\overline{C}$,

$$\left|\int_{\mathbf{v}^0\in\overline{\mathcal{V}}^{\infty}} (T(t)f)(\mathbf{v}^0)\,d\pi(\mathbf{v}^0) - \int_{\mathbf{v}^0\in\overline{\mathcal{V}}^{\infty}} f(\mathbf{v}^0)\,d\pi(\mathbf{v}^0)\right| = 0. \tag{105}$$

In other words, π is also a steady-state distribution for the deterministic fluid limit. Since by Theorem 2, the invariant state of the fluid limit is unique, Eq. (105) will imply that $\pi\left(\mathbf{v}^I\right)=1$, and this will prove the theorem.

To show Eq. (105), we write

$$\begin{aligned}\left|\int T(t)f\,d\pi - \int f\,d\pi\right| \leq{}& \limsup_{N\to\infty}\left|\int T(t)f\,d\pi - \int T(t)f\,d\pi^N\right|\\ &+\limsup_{N\to\infty}\left|\int T(t)f\,d\pi^N - \int T^N(t)f\,d\pi^N\right|\\ &+\limsup_{N\to\infty}\left|\int T^N(t)f\,d\pi^N - \int f\,d\pi\right|\end{aligned} \tag{106}$$

We will show that all three terms on the right-hand side of Eq. (106) are zero. Since $\mathbf{v}(\mathbf{v}^0,t)$ depends continuously on the initial condition $\mathbf{v}^0$ (Corollary 15), we have $T(t)f\in\overline{C}, \forall t\geq 0$, which along with $\pi^N\Rightarrow\pi$ implies that the first term is zero. For the third term, since π^N is the steady-state distribution of $\mathbf{V}^N$, we have that $\int T^N(t)f d\pi^N = \int f d\pi^N$, $\forall t\geq 0$, $f\in\overline{C}$. Since $\pi^N\Rightarrow\pi$, this implies that the last term is zero.

To bound the second term, fix some $M\in\mathbb{N}$ and let $K=\overline{\mathcal{V}}^M$. We have

$$\begin{aligned}&\limsup_{N\to\infty}\left|\int T(t)f d\pi^N - \int T^N(t)f d\pi^N\right|\\ \leq{}& \limsup_{N\to\infty}\left|\int_K T(t)f d\pi^N - \int_K T^N(t)f d\pi^N\right|\end{aligned} \tag{107}$$

$$
\begin{aligned}
&+\limsup_{N\to\infty}\left|\int_{K^c} T(t)f\,d\pi^N - \int_{K^c} T^N(t)f\,d\pi^N\right| \\
&\overset{(a)}{\leq} \limsup_{N\to\infty}\int_K \left|T^N(t)f - T(t)f\right| d\pi^N + \limsup_{N\to\infty} 2\,\|f\|\,\pi^N(K^c) \\
&\overset{(b)}{=} \limsup_{N\to\infty} 2\,\|f\|\,\pi^N(K^c)\,,
\end{aligned}
$$

where $K^c \triangleq \overline{\mathcal{V}}^\infty - K$ and $\|f\| \triangleq \sup_{\mathbf{v}\in\overline{\mathcal{V}}^\infty} |f(\mathbf{v})|$. The inequality (a) holds because $T(t)$ and $T^N(t)$ are both conditional expectations and are hence contraction mappings with respect to the sup-norm $\|f\|$. Equality (b) ($\limsup_{N\to\infty} \int_K \left|T^N(t)f - T(t)f\right| d\pi^N = 0$) can be shown using an argument involving interchanges of the order of integration, which essentially follows from the uniform rate of convergence to the fluid limit over the compact set K of initial conditions (Corollary 20). We isolate equality (b) in the following claim:

CLAIM 22. *Let K be a compact subset of $\overline{\mathcal{V}}^\infty$, we have*

$$
\limsup_{N\to\infty}\int_K \left|T^N(t)f - T(t)f\right| d\pi^N = 0 \tag{108}
$$

PROOF. Fix some $\delta > 0$. There exists $N(\delta) > 0$ such that for all $N \geq N(\delta)$, we have

$$
\begin{aligned}
&\int_K \left|T(t)f - T^N(t)f\right| d\pi^N \\
&= \int_{\mathbf{v}^0\in K} \left|f(\mathbf{v}(\mathbf{v}^0,t)) - \mathbb{E}\left[f\left(\mathbf{V}^N(t)\right) \,\middle|\, \mathbf{V}^N(0) = \mathbf{v}^0\right]\right| d\pi^N(\mathbf{v}^0) \\
&\leq \int_{\mathbf{v}^0\in K}\left(\int_{\mathbf{v}^t\in\overline{V}^\infty\cap\mathcal{Q}^N} \left|f\left(\mathbf{v}(\mathbf{v}^0,t)\right) - f\left(\mathbf{v}^t\right)\right| d\mathbb{P}_{\mathbf{V}^N(t)|\mathbf{V}^N(0)}\left(\mathbf{v}^t \,\middle|\, \mathbf{v}^0\right)\right) d\pi^N(\mathbf{v}^0) \\
&\overset{(a)}{\leq} \int_{\mathbf{v}^0\in K} \sup_{\mathbf{v}^t\in\overline{\mathcal{V}}^\infty, \|\mathbf{v}^t - \mathbf{v}(\mathbf{v}^0,t)\|_w\leq\delta} \left|f(\mathbf{v}(\mathbf{v}^0,t)) - f(\mathbf{v}^t)\right| d\pi^N(\mathbf{v}^0) \\
&\leq \omega_f(K^\delta,\delta),
\end{aligned}
$$

where K^δ is the δ-extension of K,

$$
K^\delta \triangleq \left\{\mathbf{x}\in\overline{\mathcal{V}}^\infty : \|\mathbf{x}-\mathbf{y}\|_w \leq \delta \text{ for some } y\in K\right\}, \tag{109}
$$

and $\omega_f(X,\delta)$ is the modulus of continuity of f restricted to set X:

$$
\omega_f(K,\delta) \triangleq \sup_{\mathbf{x},\mathbf{y}\in X, \|\mathbf{x}-\mathbf{y}\|_w\leq\delta} |f(\mathbf{x}) - f(\mathbf{y})|\,. \tag{110}
$$

To see why inequality (a) holds, recall that by Corollary 20, starting from a compact set of initial conditions, the sample paths of a finite system stay uniformly close to a fluid limit on a compact time interval, with high probability. Inequality (a) then follows from Eq. (102) and the fact that f is bounded. Because K is a compact set, it is not difficult show that K^{δ^0} is also compact for any fixed $\delta^0 > 0$. Hence f is uniformly continuous on K^{δ^0}, and we have

$$\limsup_{N\to\infty} \left| \int_K T(t) f \, d\pi^N - \int_K T^N(t) f \, d\pi^N \right| \le \limsup_{\delta\to 0} \omega_f(K^{\delta^0}, \delta) = 0, \tag{111}$$

which establishes the claim. □

Going back, since Eq. (107) holds for *any* $K = \overline{\mathcal{V}}^M$, $M \in \mathbb{N}$, we have, by the tightness of π^N, that the middle term in Eq. (106) is also zero. This shows that any limit point π of $\{\pi^N\}$ is indeed the unique invariant state of the fluid model $(\mathbf{v}^I)$. This completes the proof of Theorem 7. □

8. Conclusions and future work. The overall objective of this paper is to study how the degree of centralization in allocating computing or processing resources impacts performance. This investigation was motivated by applications in server farms, cloud centers, as well as more general scheduling problems with communication constraints. Using a fluid model and associated convergence theorems, we showed that any small degree of centralization induces an exponential performance improvement in the steady-state scaling of system delay, for sufficiently large systems. Simulations show good accuracy of the model even for moderately-sized finite systems ($N = 100$).

There are several interesting and important questions which we did not address in this paper. We have left out the question of what happens when the central server adopts a scheduling policy different from the Longest-Queue-First (LQF) policy considered in this paper. Since scheduling a task from a longest queue may require significant global communication overhead, other scheduling policies that require less global information may be of great practical interest. Some alternatives include

1. (*Random k-Longest-Queues*) The central server always serves a task from a queue chosen uniformly at random among the k most loaded queues, where $k \ge 2$ is a fixed integer. Note that the LQF policy is a sub-case, corresponding to $k = 1$.
2. (*Random Work-Conserving*) The central server always serves a task from a queue chosen uniformly at random among all non-empty queues.

It will be interesting to see whether a similar exponential improvement in the delay scaling is still present under these other policies. Based on the

analysis done in this paper, as well as some heuristic calculations using the fluid model, we conjecture that in order for the phase transition phenomenon to occur, a *strictly positive* fraction of the central service tokens must be used to serve a longest queue. Hence, between the two policies listed above, the former is more likely to exhibit a similar delay scaling improvement than the latter.

Assuming that the LQF policy is used, another interesting question is whether a non-trivial delay scaling can be observed if p, instead of being fixed, is a function of N and decreases to zero as $N \to \infty$. This is again of practical relevance, because having a central server whose processing speed scales linearly with N may be expensive or infeasible for certain applications.

On the modeling end, some of our current assumptions could be restrictive for practical applications. For example, the transmission delays between the local and central stations are assumed to be negligible compared to processing times; this may not be true for data centers that are separated by significant geographic distances. Also, the arrivals and services are modeled by Poisson processes, while in reality more general traffic distributions (e.g., heavy-tailed traffic) are observed. Further work to extend the current model by incorporating these realistic constraints could be of great interest, although obtaining theoretical characterizations seems quite challenging.

Last, the surprisingly simple expressions in our results make it tempting to ask whether similar performance characterizations can be obtained for other stochastic systems with partially centralized control laws; insights obtained here may find applications beyond the realm of queueing theory.

APPENDIX A: ADDITIONAL PROOFS

A.1. Proof of Proposition 11. Here we will follow a line of argument in [1] to establish the existence of fluid limits. We begin with some definitions. Recall the uniform metric, $d(\cdot,\cdot)$, defined on $D[0,T]$:

$$d(x,y) \stackrel{\triangle}{=} \sup_{t\in[0,T]} |x(t)-y(t)|, \quad x,y \in D[0,T]. \tag{112}$$

Definition 23. Let E_c be a non-empty compact subset of $D[0,T]$. A sequence of subsets of $D[0,T]$, $\mathcal{E} = \{E_N\}_{N\geq 1}$, is said to be *asymptotically close* to the set E_c if the distance to E_c of any element in $\mathcal{E}$ decreases to zero uniformly, i.e., if

$$\lim_{N\to\infty} \sup_{x\in E_N} d(x,E_c) = 0, \tag{113}$$

where the distance from a point to a set is defined as

$$d(x, E_c) \stackrel{\triangle}{=} \inf_{y \in E_c} d(x, y). \tag{114}$$

DEFINITION 24. A point $y \in D[0,T]$ is said to be a cluster point of a sequence $\{x_N\}_{N\geq 1}$ if

$$\liminf_{N\to\infty} d(x_N, y) = 0. \tag{115}$$

A point $y \in D[0,T]$ is a cluster point of a sequence of subsets $\mathcal{E} = \{E_N\}_{N\geq 1}$, if it is a cluster point of some $\{x_N\}_{N\geq 1}$ such that:

$$x_N \in E_N, \quad \forall N \geq 1. \tag{116}$$

LEMMA 25. *Let $C(\mathcal{E})$ be the set of cluster points of $\mathcal{E} = \{E_N\}_{N\geq 1}$. If $\mathcal{E}$ is asymptotically close to a compact set E_c, then*

1. *$\mathcal{E}$ is asymptotically close to $C(\mathcal{E})$.*
2. *$C(\mathcal{E}) \subset E_c$.*

PROOF. Suppose that the first claim is false. Then, there exists a subsequence $\{x_{N_i}\}_{i\geq 1}$, where $x_{N_i} \in E_{N_i}, \forall i$, such that

$$d(x_{N_i}, C(\mathcal{E})) = \gamma > 0, \quad \forall i \geq 1. \tag{117}$$

However, since $\mathcal{E}$ is asymptotically close to E_c by assumption, there exists a sequence $\{y_i\} \subset E_c$ such that

$$d(x_{N_i}, y_i) \to 0, \quad \text{as } i \to \infty. \tag{118}$$

Since E_c is compact, $\{y_i\}$ has a convergent subsequence with limit $\tilde{y}$. By Eq. (118), $\tilde{y}$ is a cluster point of $\{x_{N_i}\}$, and hence a cluster point of $\mathcal{E}$, contradicting Eq. (117). This proves the first claim.

The second claim is an easy consequence of the closedness of E_c. Let $\tilde{x}$ be any point in $C(\mathcal{E})$. There exists a subsequence $\{x_{N_i}\}$, where $x_{N_i} \in E_{N_i}, \forall i$, such that $\lim_{i\to\infty} d(x_{N_i}, \tilde{x}) = 0$, by the definition of a cluster point. By the same reasoning as in the first part of the proof (Eq. (118)), there exists a sequence $\{y_i\} \subset E_c$ which also converges to $\tilde{x}$. Since E_c is closed, $\tilde{x} \in E_c$. □

We now put the above definition into our context. Define $\mathcal{E} = \{E_N\}_{N\geq 1}$ to be a sequence of subsets of $D[0,T]$ such that

$$\begin{aligned} E_N \;=\; \big\{ & x \in D[0,T] : \; |x(0) - x^0| \leq M_N, \;\; \text{and} \\ & |x(a) - x(b)| \leq L|a-b| + \gamma_N, \quad \forall a, b \in [0,T] \big\}, \end{aligned} \tag{119}$$

where x^0 is a constant, $M_N \downarrow 0$ and $\gamma_N \downarrow 0$ are two sequences of diminishing non-negative numbers. We first characterize the set of cluster points of the sequence $\mathcal{E}$. Loosely speaking, $\mathcal{E}$ represents a sequence of sample paths that get increasingly "close" to the set of L-Lipschitz continuous functions, in that all elements of $\mathcal{E}$ are "γ_N-approximately" Lipschitz-continuous. The definition below and the lemma that follows formalize this notion.

Define E_c as the set of *Lipschitz-continuous* functions on $[0,T]$ with Lipschitz constant L and initial values bounded by a positive constant M:

(120)
$$E_c \stackrel{\triangle}{=} \{x \in D[0,T] : |x(0)| \le M, \text{ and } |x(a) - x(b)| \le L|a-b|, \forall a,b \in [0,T]\}.$$

We have the following property of E_c.

LEMMA 26. *E_c is compact.*

PROOF. E_c is a set of L-Lipschitz continuous functions $x(\cdot)$ on $[0,T]$ with initial values contained in a closed and bounded interval. By the Arzela-Ascoli theorem, every sequence of elements of E_c contains a further subsequence which converges to some $x^*(\cdot)$ uniformly on $[0,T]$. Since all elements in E_c are L-Lipschitz continuous, $x^*(\cdot)$ is also Lipschitz-continuous on $[0,T]$. It is clear that $x^*(\cdot)$ also satisfies $x^*(0) \le M$. Hence, $x^*(\cdot) \in E_c$. □

LEMMA 27. *$\mathcal{E}$ is asymptotically close to E_c.*

PROOF. The proof involves an elementary but tedious interpolation argument; see Appendix A.1 of [29] for the details of this argument. □

Finally, the following lemma states that all sample paths $\mathbf{X}^N(\omega,\cdot)$ with $\omega \in \mathcal{C}$ belong to E_N, with appropriately chosen $\{M_N\}_{N\ge1}$ and $\{\gamma_N\}_{N\ge1}$.

LEMMA 28. *Suppose that there exists $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$ such that for all $\omega \in \mathcal{C}$*

$$\left\|\mathbf{V}^N(\omega,0) - \mathbf{v}^0\right\|_w \le \tilde{M}_N, \tag{121}$$

for some $\tilde{M}_N \downarrow 0$. Then for all $\omega \in \mathcal{C}$ and $i \in \mathbb{Z}_+$, there exist $L > 0$ and sequences $M_N \downarrow 0$ and $\gamma_N \downarrow 0$ such that

$$\mathbf{X}_i^N(\omega,\cdot) \in E_N, \tag{122}$$

where E_N is defined as in Eq. (119).

PROOF. Intuitively, the lemma follows from the uniform convergence of scaled sample paths of the event process $W^N(\omega,t)$ to $(1+\lambda)t$ (Lemma 9),

because jumps along any coordinate of the sample path have a magnitude of $\frac{1}{N}$, and because all coordinates of $\mathbf{X}^N$ are dominated in terms of W and the total number of jumps.

Based on our coupling construction (cf. Section 4.3), each coordinate of $\mathbf{A}^N, \mathbf{L}^N$, and $\mathbf{C}^N$ is monotonically non-decreasing, with a positive jump at time t of magnitude $\frac{1}{N}$ only if there is a jump of same size at time t in $W(\omega,\cdot)$. Hence for all $i \geq 1$,

$$\left|\mathbf{A}_i^N(\omega,a) - \mathbf{A}_i^N(\omega,b)\right| \leq \left|W^N(\omega,a) - W^N(\omega,b)\right|, \quad \forall a,b \in [0,T]. \tag{123}$$

The same inequalities hold for $\mathbf{L}^N$ and $\mathbf{C}^N$. Since by construction,

$$\mathbf{V}_i^N(\omega,t) = \mathbf{V}_i^N(\omega,0) + \mathbf{A}_i^N(\omega,t) - \mathbf{L}_i^N(\omega,t) - \mathbf{C}_i^N(\omega,t), \quad \forall i \geq 1, \tag{124}$$

we have, for all $i \geq 1$,

$$\left|\mathbf{X}_i^N(\omega,a) - \mathbf{X}_i^N(\omega,b)\right| \leq 3\left|W^N(\omega,a) - W^N(\omega,b)\right|, \quad \forall a,b \in [0,T]. \tag{125}$$

Since $\omega \in \mathcal{C}$, $\mathbf{W}^N(\omega,\cdot)$ converges uniformly to $(\lambda+1)t$ on [0,T] by Lemma 9. This implies that there exists a sequence $\tilde{\gamma}_N \downarrow 0$ such that for all $N \geq 1$,

$$\left|W^N(\omega,a) - W^N(\omega,b)\right| \leq (\lambda+1)|a-b| + \tilde{\gamma}_N, \quad \forall a,b \in [0,T], \tag{126}$$

which, in light of Eq. (125), implies that

$$\left|\mathbf{X}_i^N(\omega,a) - \mathbf{X}_i^N(\omega,b)\right| \leq 3(\lambda+1)|a-b| + 3\tilde{\gamma}_N, \quad \forall a,b \in [0,T], i \geq 1. \tag{127}$$

Finally, note that all coordinates of $\mathbf{X}^N(\omega,0)$ except for $\mathbf{V}^N(\omega,0)$ are equal to 0 by definition. The proof is completed by setting $M_N = 2^i \tilde{M}_N$, $\gamma_N = 3\tilde{\gamma}_N$, and $L = 3(\lambda+1)$. □

We are now ready to prove Proposition 11.

Proof of Proposition 11. Let us first summarize the key results we have so far:

1. (Lemma 26) E_c is a set of L-Lipschitz continuous functions with bounded values at 0, and it is compact and closed.
2. (Lemma 27) $\mathcal{E} = \{E_N\}_{N\geq 1}$, a sequence of sets of γ_N-approximate L-Lipschitz-continuous functions with convergent initial values, is asymptotically close E_c.
3. (Lemma 28) For all $\omega \in \mathcal{C}$, $\mathbf{X}^N(\omega,\cdot)$ is in E_N for all $N \geq 1$.

The rest is straightforward: Pick any $\omega \in \mathcal{C}$. By the above statements, for any $i \in \mathbb{Z}_+$ one can find a subsequence $\left\{\mathbf{X}^{N_j}(\omega,\cdot)\right\}_{j=1}^{\infty}$ and a sequence $\{y_j\}_{j=1}^{\infty} \subset E_c$ such that

$$d\left(\mathbf{X}_i^{N_j}(\omega,\cdot), y_j\right) \longrightarrow 0, \text{ as } j \to \infty. \tag{128}$$

Since by Lemma 26 (statement 1 above), E_c is compact and closed, $\{y_j\}_{j=1}^{\infty}$ has a limit point y^* in E_c, which implies that a further subsequence of $\left\{\mathbf{X}_i^{N_j}(\omega,\cdot)\right\}_{i=1}^{\infty}$ converges to y^*. Moreover, since $\mathbf{V}^N(\omega,0) \to \mathbf{v}^0$ and $A^N(\omega,0) = L^N(\omega,0) = C^N(\omega,0) = 0$, $y^*(0)$ is unique. This proves the existence of a L-Lipschitz-continuous limit point $y^*(\cdot)$ at any single coordinate i of $\mathbf{X}^N(\cdot)$.

Starting with the coordinate-wise limit points, we then use a diagonal argument to construct the limit points of $\mathbf{X}^N$ in the $D^{\mathbb{Z}_+}[0,T]$ space. Let $v_1(t)$ be any L-Lipschitz-continuous limit point of $\mathbf{V}_1^N$, so that a subsequence $\mathbf{V}_1^{N_j^1}(\omega,\cdot) \to v_1$ as $j \to \infty$ in $d(\cdot,\cdot)$. Then, proceed recursively by letting $v_{i+1}(t)$ be a limit point of a subsequence of $\left\{\mathbf{V}_{i+1}^{N_j^i}(\omega,\cdot)\right\}_{j=1}^{\infty}$, where $\{N_j^i\}_{j=1}^{\infty}$ are the indices for the ith subsequence. Finally, define

$$\mathbf{v}_i = v_i, \quad \forall i \in \mathbb{Z}_+. \tag{129}$$

We claim that $\mathbf{v}$ is indeed a limit point of $\mathbf{V}^N$ in the $d^{\mathbb{Z}_+}(\cdot,\cdot)$ norm. To see this, first note that for all N,

$$\mathbf{V}_1^N(\omega,t) \geq \mathbf{V}_i^N(\omega,t) \geq 0, \quad \forall i \geq 1, t \in [0,T]. \tag{130}$$

Since we constructed the limit point $\mathbf{v}$ by repeatedly selecting nested subsequences, this property extends to $\mathbf{v}$, i.e.,

$$\mathbf{v}_1(t) \geq \mathbf{v}_i(t) \geq 0, \quad \forall i \geq 1, t \in [0,T]. \tag{131}$$

Since $\mathbf{v}_1(0) = \mathbf{v}_1^0$ and $\mathbf{v}_1(t)$ is L-Lipschitz-continuous, we have that

$$\sup_{t\in[0,T]} |\mathbf{v}_i(t)| \leq \sup_{t\in[0,T]} |\mathbf{v}_1(t)| \leq \left|\mathbf{v}_1^0\right| + LT, \quad \forall i \in \mathbb{Z}_+. \tag{132}$$

Set $N_1 = 1$, and let

$$N_k = \min\left\{N \geq N_{k-1} : \sup_{1\leq i\leq k} d(\mathbf{V}_i^N(\omega,\cdot), \mathbf{v}_i) \leq \frac{1}{k}\right\}, \quad \forall k \geq 2. \tag{133}$$

Note that the construction of $\mathbf{v}$ implies N_k is well defined and finite for all k. From Eqs. (132)–(133), we have for all $k \geq 2$,

$$
\begin{aligned}
(134) \qquad d^{\mathbb{Z}_+}\left(\mathbf{V}^{N_k}(\omega,\cdot),\mathbf{v}\right) &= \sup_{t\in[0,T]} \sqrt{\sum_{i=0}^{\infty} \frac{\left|\mathbf{V}_i^{N_k}(\omega,t)-\mathbf{v}_i(t)\right|^2}{2^i}} \\
&\leq \frac{1}{k} + \sqrt{\left(\left|\mathbf{v}_1^0\right|+LT\right)^2 \sum_{i=k+1}^{\infty} \frac{1}{2^i}} \\
&= \frac{1}{k} + \frac{1}{2^{k/2}}\left(\left|\mathbf{v}_1^0\right|+LT\right)
\end{aligned}
$$

Hence $d^{\mathbb{Z}_+}\left(\mathbf{V}^{N_k}(\omega,\cdot),\mathbf{v}\right) \to 0$ as $k \to \infty$. The existence of the limit points $\mathbf{a}(t), \mathbf{l}(t)$, and $\mathbf{c}(t)$ can be established by an identical argument. This completes the proof. □

A.2. Proof of Claim 14.

Proof of Claim 14. Let $\mathbf{m}_i \triangleq \mathbf{v}_i - \mathbf{w}_i$. Note that for all $i \geq 1$

$$
\begin{aligned}
(135) \qquad & (\mathbf{v}_i-\mathbf{w}_i)\left[\left(\dot{\mathbf{a}}_i^{\mathbf{v}}-\dot{\mathbf{l}}_i^{\mathbf{v}}\right)-\left(\dot{\mathbf{a}}_i^{\mathbf{w}}-\dot{\mathbf{l}}_i^{\mathbf{w}}\right)\right] \\
&= (\mathbf{v}_i-\mathbf{w}_i)\left[\lambda(\mathbf{v}_{i-1}-\mathbf{w}_{i-1})-\lambda\left(\mathbf{v}_i-\mathbf{w}_i\right)-(1-p)\left(\mathbf{v}_i-\mathbf{w}_i\right)\right. \\
&\quad \left.+(1-p)\left(\mathbf{v}_{i+1}-\mathbf{w}_{i+1}\right)\right] \\
&= \mathbf{m}_i\left(\lambda\mathbf{m}_{i-1}-\lambda\mathbf{m}_i-(1-p)\mathbf{m}_i+(1-p)\mathbf{m}_{i+1}\right) \\
&\leq \frac{\lambda}{2}\left(\mathbf{m}_{i-1}^2+\mathbf{m}_i^2\right)-(\lambda+1-p)\,\mathbf{m}_i^2+\frac{1-p}{2}\left(\mathbf{m}_i^2+\mathbf{m}_{i+1}^2\right) \\
&\leq \lambda\mathbf{m}_{i-1}^2+(1-p)\,\mathbf{m}_{i+1}^2-\frac{\lambda+1-p}{2}\mathbf{m}_i^2 \\
&\leq \lambda\mathbf{m}_{i-1}^2+(1-p)\,\mathbf{m}_{i+1}^2.
\end{aligned}
$$

For $i = 0$, by Eq. (70), we have

$$
\begin{aligned}
(\mathbf{v}_0-\mathbf{w}_0)\left[\left(\dot{\mathbf{a}}_0^{\mathbf{v}}-\dot{\mathbf{l}}_0^{\mathbf{v}}\right)-\left(\dot{\mathbf{a}}_0^{\mathbf{w}}-\dot{\mathbf{l}}_0^{\mathbf{w}}\right)\right] &= (\mathbf{v}_1-\mathbf{w}_1)\left[\left(\dot{\mathbf{a}}_1^{\mathbf{v}}-\dot{\mathbf{l}}_1^{\mathbf{v}}\right)-\left(\dot{\mathbf{a}}_1^{\mathbf{w}}-\dot{\mathbf{l}}_1^{\mathbf{w}}\right)\right] \\
(136) \qquad &\leq \lambda\mathbf{m}_0^2+(1-p)\mathbf{m}_2^2.
\end{aligned}
$$

Combining Eqs. (135) and (136), we have

$$
\begin{aligned}
(137) \qquad & \sum_{i=0}^{\infty} \frac{(\mathbf{v}_i-\mathbf{w}_i)\left[\left(\dot{\mathbf{a}}_i^{\mathbf{v}}-\dot{\mathbf{l}}_i^{\mathbf{v}}\right)-\left(\dot{\mathbf{a}}_i^{\mathbf{w}}-\dot{\mathbf{l}}_i^{\mathbf{w}}\right)\right]}{2^{i-1}} \\
&\leq 2(\lambda\mathbf{m}_0^2+(1-p)\mathbf{m}_2^2)+\sum_{i=1}^{\infty}\frac{1}{2^{i-1}}\left(\lambda\mathbf{m}_{i-1}^2+(1-p)\mathbf{m}_{i+1}^2\right)
\end{aligned}
$$

$$\begin{aligned} &\leq 6(\lambda+1-p)\left(\sum_{i=0}^{\infty}\frac{1}{2^i}\mathbf{m}_i^2\right)\\ &= 6(\lambda+1-p)\|\mathbf{v}-\mathbf{w}\|_w^2. \end{aligned}$$

This proves the claim. □

A.3. Proof of Proposition 21.

Proof of Proposition 21. Fix $N>0$ and $p\in(0,1]$. Let $\{\mathbf{V}^N[n]\}_{n\geq 0}$ be the discrete-time embedded Markov chain for $\mathbf{V}^N(t)$, defined by

$$\mathbf{V}^N[n]\stackrel{\Delta}{=}\mathbf{V}^N(t_n),\quad n\geq 0, \tag{138}$$

where t_n, $n\geq 1$, as defined previously, is the time of the nth event taking place in the system (i.e., the time of the nth jump in $W^N(\cdot)$), with the convention that $t_0=0$. Let also $\mathbf{U}^N(t)$ be the sample path obtained if we set p to zero, and let $\mathbf{U}^N[n]$ be the corresponding embedded chain.

Definition 29 (Stochastic Dominance). Let $\{X[n]\}_{n\geq 0}$ and $\{Y[n]\}_{n\geq 0}$ be two discrete-time stochastic processes taking values in $\mathbb{R}^{\mathbb{Z}_+}$. We say that $\{X[n]\}_{n\geq 0}$ is stochastically dominated by $\{Y[n]\}_{n\geq 0}$, denoted by $\{X[n]\}_{n\geq 0}\preceq_{st}\{Y[n]\}_{n\geq 0}$, if there exist random processes $\{X'[n]\}_{n\geq 0}$ and $\{Y'[n]\}_{n\geq 0}$ defined on a common probability space $(\Omega,\mathcal{F},\mathbb{P})$, such that

1. X' and Y' have the same distributions as X and Y, respectively.
2. $X'[n]\leq Y'[n]$, $\forall n\geq 0$, $\mathbb{P}$–almost surely.

We have the following lemma:

Lemma 30. *Fix any $p\in(0,1]$. If $\mathbf{V}^N[0]=\mathbf{U}^N[0]$, then $\{\mathbf{V}_1^N[n]\}_{n\geq 0}\preceq_{st}\{\mathbf{U}_1^N[n]\}_{n\geq 0}$.*

Proof. We will first interpret the system with $p>0$ as that of an optimal scheduling policy with a time-varying channel. The result will then follow from the classical result in Theorem 3 of [23], with a slightly modified arrival assumption, but almost identical proof steps. Recall the Secondary Motivation described in Section 1.2. Here we will use a similar but modified interpretation: instead of thinking of the central server as deciding between serving a most-loaded station versus serving a random station, imagine that the central server always serves a most-loaded station among the ones that are *connected* to it. The channel between the central server and local stations, represented by a set of connected stations, evolves according to the following dynamics and is independent across different time slots:

1. With probability p, all N stations are connected to the central server.
2. Otherwise, only one station, chosen uniformly at random from the N stations, is connected to the central server.

It is easy to see that, under the above channel dynamics, a system in which a central server always serves a most-loaded stations among *connected* stations will produce the same distribution for $\mathbf{V}^N[n]$ as our original system. Note that the case $p = 0$ is equivalent to scheduling tasks under the same channel condition just described, but with a server that servers a station chosen *uniformly at random* among all connected stations. The advantage of the above interpretation is that it allows us to treat $\mathbf{V}^N[n]$ and $\mathbf{U}^N[n]$ as the resulting aggregate queue length processes by applying two *different* scheduling policies to the *same* arrival, token generation, and channel processes. In particular, $\mathbf{V}_1^N[n]$ corresponds to the resulting normalized total queue length process ($\mathbf{V}^N \stackrel{\triangle}{=} \frac{1}{N}\sum_{i=1}^N Q_i(t_n)$), when a longest-queue-first policy is applied, and $\mathbf{U}_1^N[n]$ corresponds to the normalized total queue length process, when a fully random scheduling policy is applied. Theorem 3 of [23] states that when the arrival and channel processes are symmetric with respect to the identities of stations, the total queue length process under a longest-queue-first policy is stochastically dominated by all other causal policies (i.e., policies that use only information from the past). Since the arrival and channel processes are symmetric in our case, and a random scheduling policy falls under the category of causal policies, the statement of Theorem 3 of [23] implies the validity of our claim.

There is, however, a minor difference in the assumptions of Theorem 3 of [23] and our setup that we note here. In [23], it is possible that both arrivals and service occur during the same slot, while in our case, each event corresponds either to an arrival to a queue or the generation of a service token, but not both. This technical difference can be handled easily by discussing separately, whether the current slot corresponds to an arrival or a service. The structure of the proof for Theorem 3 in [23] remains unchanged after this modification, and is hence not repeated here. □

Having established a stochastic dominance relation between the two (discrete-time) embedded Markov chains, and using the fact that the continuous-time chains $\mathbf{V}(t)$ and $\mathbf{U}(t)$ have transitions at the jump times of the common underlying Poisson process $W(t)$, it is elementary to check that $\mathbf{V}(t)$ is stochastically dominated by $\mathbf{U}(t)$; see Appendix A.2 in [29] for the details of this argument.

We now look at the behavior of $\mathbf{U}^N(t)$. When $p = 0$, only local service tokens are generated. Hence, it is easy to see that the system degenerates

into N individual $M/M/1$ queues with independent and identical statistics for arrivals and service token generation. In particular, for any station i, the arrival follows a Poisson process of rate λ and the generation of service tokens follows a Poisson process of rate 1. Since $\lambda < 1$, it is not difficult to verify that the process $\mathbf{U}^N(t)$ is positive recurrent, and admits a unique steady-state distribution, denoted by $\pi^{N,0}$, which satisfies:

$$\pi^{N,0}(\mathbf{V}_1 \le x) = \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} E_i \le x\right), \quad \forall x \in \mathbb{R}, \tag{139}$$

where $\{E_i\}_{i=1}^N$ is a set of i.i.d. geometrically distributed random variables, with

$$\mathbb{P}(E_i = k) = \lambda^k(1-\lambda), \quad \forall k \in \mathbb{Z}_+, \tag{140}$$

Since the process $\mathbf{V}^N(t)$ is dominated by $\mathbf{U}^N(t)$, it is easily verified that the former is also positive recurrent. In particular, $\mathbf{V}^N(t)$ converges in distribution to a unique steady-state distribution $\pi^{N,p}$ as $t \to \infty$. Combining this with the dominance relation between the two processes, we have that for any initial distribution of $\hat{V}^N(0)$,

$$\begin{aligned} \pi^{N,p}(\overline{\mathcal{V}}^M) &\triangleq \pi^{N,p}(\mathbf{V}_1^N \le M) \\ &= \lim_{t\to\infty} \mathbb{P}\left(\mathbf{V}^N(t) \le M\right) \\ &\ge \lim_{t\to\infty} \mathbb{P}\left(\mathbf{U}^N(t) \le M\right) \\ &= \pi^{N,0}(\mathbf{V}_1^N \le M) \\ &= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} E_i \le M\right), \end{aligned} \tag{141}$$

where the last equality follows from Eq. (139). Since the E_i's are i.i.d. geometric random variables, by Markov's inequality,

$$\pi^{N,p}(\overline{\mathcal{V}}^M) \ge 1 - \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} E_i \ge M\right) \ge 1 - \frac{\mathbb{E}(E_1)}{M} = 1 - \frac{\lambda}{(1-\lambda)M}, \tag{142}$$

for all $M > \mathbb{E}(E_1) = \frac{\lambda}{1-\lambda}$, which establishes the tightness of $\{\pi^{N,p}\}_{N=1}^{\infty}$. This completes the proof of Proposition 21. □

APPENDIX B: $\mathbf{V}(\cdot)$ VERSUS $\mathbf{S}(\cdot)$, AND THE UNIQUENESS OF FLUID LIMITS

In this section, we offer some justification for having chosen to work primarily with the aggregate queue length process, $\mathbf{V}^N(\cdot)$ (Eq. (2)), instead of

the *normalized queue length process*, $\mathbf{S}^N(\cdot)$ (Eq. (3)). The high-level reason is that the fluid model corresponding to the aggregate queue length process, expressed in terms of $\mathbf{v}(\cdot)$, admits a nice contraction property which does not hold for the fluid model expressed in $\mathbf{s}(\cdot)$.

A key to the proof of Theorem 4 (uniqueness of fluid solutions) is a contraction property of the drift associated with $\mathbf{v}(\cdot)$ (Eq. (71)), also known as the one-sided Lipschitz continuity (OSL) condition in the dynamical systems literature (see, e.g., [22]). We first give a definition of OSL that applies to our setting.

DEFINITION 31. Let the coordinates of $\mathbb{R}^\infty$ be indexed by $\mathbb{Z}_+$ so that $\mathbf{x} = (x_0, x_1, x_2, \ldots)$ for all $\mathbf{x} \in \mathbb{R}^\infty$. A function $\mathbf{H} : \mathbb{R}^\infty \to \mathbb{R}^\infty$ is said to be one-sided Lipschitz-continuous (OSL) over a subset $D \subset \mathbb{R}^\infty$, if there exists a constant C, such that for every $\mathbf{x}, \mathbf{y} \in D$,

$$\langle \mathbf{x} - \mathbf{y}, \mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y}) \rangle_w \le C \, \|\mathbf{x} - \mathbf{y}\|_w^2 , \tag{143}$$

where the inner product $\langle \cdot, \cdot \rangle_w$ on $\mathbb{R}^\infty$ is defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle_w \stackrel{\triangle}{=} \sum_{i=0}^{\infty} \frac{x_i y_i}{2^i}. \tag{144}$$

What is the usefulness of the above definition in the context of proving uniqueness of solutions to a fluid model? Recall that $\mathbf{F}(\cdot)$ is the drift of the fluid model, as in Eq. (18), i.e.,

$$\dot{\mathbf{v}}(t) = \mathbf{F}(\mathbf{v}(t)), \tag{145}$$

whenever $\mathbf{v}(\cdot)$ is differentiable at t. Let $\mathbf{v}(\cdot)$ and $\mathbf{w}(\cdot)$ be two solutions to the fluid model such that both are differentiable at t, as in the proof of Theorem 4. We have

$$\frac{d}{dt} \|\mathbf{v}(t) - \mathbf{w}(t)\|_w^2 = 2 \langle \mathbf{v}(t) - \mathbf{w}(t), \mathbf{F}(\mathbf{v}(t)) - \mathbf{F}(\mathbf{w}(t)) \rangle_w . \tag{146}$$

Therefore, if $\mathbf{F}(\cdot)$ is one-sided Lipschitz-continuous, as defined by Eq. (143), we immediately obtain the key inequality in Eq. (71), from which the uniqueness of fluid solutions follows. The computation carried out in Eq. (71) was essentially verifying the OSL condition of $\mathbf{F}(\cdot)$ on the domain $\overline{\mathcal{V}}^\infty$.

For the state representation based on $\mathbf{s}(\cdot)$, can one use the same proof technique to show the uniqueness of $\mathbf{s}(\cdot)$ by working directly with the drift associated with $\mathbf{s}(\cdot)$? Recall that

$$\mathbf{s}_i(t) \stackrel{\triangle}{=} \mathbf{v}_i(t) - \mathbf{v}_{i+1}(t), \quad \forall i \ge 0, \tag{147}$$

so that at all t where $\mathbf{v}(t)$ is differentiable, the drift $\dot{\mathbf{s}}(t)$ is given by

$$\mathbf{H}_i(\mathbf{s}(t)) = \dot{\mathbf{s}}_i(t) = \dot{\mathbf{v}}_i(t) - \dot{\mathbf{v}}_{i+1}(t) = \lambda(\mathbf{s}_{i-1} - \mathbf{s}_i) - (1-p)(\mathbf{s}_i - \mathbf{s}_{i+1}) - g_i^s(\mathbf{s}), \tag{148}$$

for all $i \geq 1$, where $g_i^s(\mathbf{s}) \stackrel{\triangle}{=} g_i(\mathbf{v}) - g_{i+1}(\mathbf{v})$, i.e.,

$$g_i^s(\mathbf{s}) = \begin{cases} 0, & \mathbf{s}_i > 0, \mathbf{s}_{i+1} > 0, \\ p - \min\{\lambda \mathbf{s}_i, p\}, & \mathbf{s}_i > 0, \mathbf{s}_{i+1} = 0, \\ \min\{\lambda \mathbf{s}_{i-1}, p\}, & \mathbf{s}_i = 0, \mathbf{s}_{i-1} > 0, \\ 0, & \mathbf{s}_i = 0, \mathbf{s}_{i-1} = 0. \end{cases} \tag{149}$$

Interestingly, it turns out that the drift $\mathbf{H}(\cdot)$, defined in Eq. (148), does not satisfy the one-sided Lipschitz continuity condition in general. We show this fact by inspecting a specific example. To keep the example as simple as possible, we consider a degenerate case.

CLAIM 32. *If $\lambda = 0$ and $p = 1$, then $\mathbf{H}(\cdot)$ is not one-sided Lipschitz-continuous on its domain $\overline{\mathcal{S}}^{\infty}$, where $\overline{\mathcal{S}}^{\infty}$ was defined in Eq. (8) as*

$$\overline{\mathcal{S}}^{\infty} \stackrel{\triangle}{=} \left\{ \mathbf{s} \in \mathcal{S} : \sum_{i=1}^{\infty} \mathbf{s}_i < \infty \right\}.$$

PROOF. We will look at a specific instance where the condition (143) cannot be satisfied for any C. For the rest of the proof, a vector $\mathbf{s} \in \overline{\mathcal{S}}^{\infty}$ will be written explicitly as $\mathbf{s} = (\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \ldots)$. Consider two vectors

$$\mathbf{s}^a = (1, \alpha, 0, 0, \ldots) \text{ and } \mathbf{s}^b = (1, \alpha + \epsilon, \beta, 0, 0, \ldots), \tag{150}$$

where $1 \geq \alpha + \epsilon \geq \beta > 0$, $1 \geq \alpha > 0$, and $\mathbf{s}_i^a = \mathbf{s}_i^b = 0$ for all $i \geq 3$. Note that $\mathbf{s}_i^a, \mathbf{s}_i^b \in \overline{\mathcal{S}}^{\infty}$.

To prove the claim, it suffices to show that for any value of C, there exist some values of α, β, and ϵ such that

$$\left\langle \mathbf{s}^b - \mathbf{s}^a, \mathbf{H}(\mathbf{s}^b) - \mathbf{H}(\mathbf{s}^a) \right\rangle_w > C \left\| \mathbf{s}^b - \mathbf{s}^a \right\|_w^2. \tag{151}$$

Since $\lambda = 0$ and $p = 1$, by the definition of $\mathbf{H}(\cdot)$ (Eqs. (148) and (149)), we have

$$\mathbf{H}(\mathbf{s}^a) = (0, -1, 0, 0, \ldots) \text{ and } \mathbf{H}(\mathbf{s}^b) = (0, 0, -1, 0, 0, \ldots). \tag{152}$$

Combining Eqs. (150) and (152), we have

$$\begin{aligned} \mathbf{s}^b - \mathbf{s}^a &= (0, \epsilon, \beta, 0, 0, \ldots), \\ \text{and } \mathbf{H}(\mathbf{s}^b) - \mathbf{H}(\mathbf{s}^a) &= (0, 1, -1, 0, 0, \ldots), \end{aligned}$$

which yields

$$\left\langle \mathbf{s}^b - \mathbf{s}^a, \mathbf{H}(\mathbf{s}^b) - \mathbf{H}(\mathbf{s}^a) \right\rangle_w = \frac{1}{2}\epsilon - \frac{1}{4}\beta. \tag{153}$$

Since

$$C \left\| \mathbf{s}^b - \mathbf{s}^s \right\|_w^2 \triangleq C \sum_{i=0}^{\infty} \frac{1}{2^i} \left(\mathbf{s}_i^b - \mathbf{s}_i^a \right)^2 = C \left(\frac{1}{2}\epsilon^2 + \frac{1}{4}\beta^2 \right), \tag{154}$$

we have that for all C and all $\epsilon < \frac{1}{C}$,

$$\left\langle \mathbf{s}^a - \mathbf{s}^b, \mathbf{H}(\mathbf{s}^b) - \mathbf{H}(\mathbf{s}^a) \right\rangle_w = \frac{1}{2}\epsilon - \frac{1}{4}\beta > C \left(\frac{1}{2}\epsilon^2 + \frac{1}{4}\beta^2 \right) = C \left\| \mathbf{s}^a - \mathbf{s}^b \right\|_w^2, \tag{155}$$

for all sufficiently small β, which proves Eq. (151). This completes the proof of the claim. □

Claim 32 indicates that a direct proof of uniqueness of fluid solutions using the OSL property of the drift will not work for $\mathbf{s}(\cdot)$. The uniqueness of $\mathbf{s}(\cdot)$ should still hold, but the proof can potentially be much more difficult, requiring an examination of all points of discontinuity of $\mathbf{H}(\cdot)$ on the domain $\overline{\mathcal{S}}^{\infty}$.

We now give some intuition as for why the discontinuity in Claim 32 occurs for $\mathbf{H}(\cdot)$, but not for $\mathbf{F}(\cdot)$. The key difference lies in the expressions of the *drifts due to central service tokens* in two fluid models, namely, $g(\cdot)$ (Eq. (16)) for $\mathbf{v}(\cdot)$ and $g^s(\cdot)$ (Eq. (149)) for $\mathbf{s}(\cdot)$. For $g^s(\cdot)$, note that

$$g_i^s(\mathbf{s}) = 0, \text{ if } \mathbf{s}_i > 0 \text{ and } \mathbf{s}_{i+1} > 0, \tag{156}$$

$$\text{and } g_i^s(\mathbf{s}) = p - \min\left\{\lambda \mathbf{s}_i, p\right\}, \text{ if } \mathbf{s}_i > 0 \text{ and } \mathbf{s}_{i+1} = 0. \tag{157}$$

In other words, the ith coordinate of $\mathbf{s}(t)$, $\mathbf{s}_i(t)$ receives *no drift* due to the central service tokens if there is a strictly positive fraction of queues in the system with at least $i+1$ tasks, that is, if $\mathbf{s}_{i+1}(t) > 0$ (Eq. (156)). However, as soon as $\mathbf{s}_{i+1}(t)$ becomes zero, $\mathbf{s}_i(t)$ immediately receives a strictly positive amount of drift due to the central service tokens (Eq. (157)), as long as $\lambda \mathbf{s}_i(t) < p$. Physically, since the central server always targets the longest queues, this means that when $\mathbf{s}_{i+1}(t)$ becomes zero, the set of queues with exactly i tasks becomes the longest in the system, and begins to receive a positive amount of attention from the central server. Such a sudden change in the drift of $\mathbf{s}_i(t)$ as a result of $\mathbf{s}_{i+1}(t)$ hitting zero is a main cause of the failure of the OSL condition, and this can be observed in Eq. (155) as $\beta \to 0$. In general, the type of discontinuity that was exploited in the proof of

Claim 32 can happen at infinitely many points in $\overline{\mathcal{S}}^{\infty}$. The particular choices of $\lambda = 0$ and $p = 1$ were non-essential, and were only chosen to simplify the calculations.

We now turn to the expression for $g(\cdot)$, the drift of $\mathbf{v}(\cdot)$ due to the central service tokens. We have that

$$g_i(\mathbf{v}) = p, \text{ whenever } \mathbf{v}_i > 0. \tag{158}$$

Note that the above-mentioned discontinuity in $g^s(\cdot)$ is not present in $g(\cdot)$. This is not surprising: since $\mathbf{v}_i(t) \stackrel{\Delta}{=} \sum_{j=i}^{\infty} s_j(t)$, $\mathbf{v}_i(t)$ receives a *constant amount* of drift from the central service token as long as $\mathbf{v}_i(t) > 0$, *regardless* of the values of $\mathbf{v}_j(t), j \geq i+1$. By adding up the coordinates $\mathbf{s}_j(\cdot)$, $j \geq i$, to obtain $\mathbf{v}_i(\cdot)$, we have effectively eliminated many of the drift discontinuities in $\mathbf{s}(\cdot)$. This is a key reason for the one-sided Lipschitz continuity condition to hold for $\mathbf{F}(\cdot)$.

To illustrate this "smoothing" effect, consider again the examples of $\mathbf{s}^a$ and $\mathbf{s}^b$ in Eq. (150). In terms of $\mathbf{v}$, we have

$$\mathbf{v}^a = (1+\alpha, \alpha, 0, 0, \ldots) \text{ and } \mathbf{v}^b = (1+\alpha+\epsilon+\beta, \alpha+\epsilon+\beta, \beta, 0, 0, \ldots). \tag{159}$$

We then have

$$\mathbf{F}(\mathbf{v}^a) = (-1, -1, 0, 0, \ldots) \text{ and } \mathbf{F}(\mathbf{v}^b) = (-1, -1, -1, 0, 0, \ldots). \tag{160}$$

Combining Eqs. (159) and (160), we have

$$\begin{aligned} \mathbf{v}^b - \mathbf{v}^a &= (\epsilon+\beta, \epsilon+\beta, \beta, 0, 0, \ldots), \\ \text{and } \mathbf{F}(\mathbf{v}^b) - \mathbf{F}(\mathbf{v}^a) &= (0, 0, -1, 0, 0, \ldots). \end{aligned}$$

This implies that for all $C \geq 0$,

$$\left\langle \mathbf{v}^a - \mathbf{v}^b, \mathbf{F}(\mathbf{v}^b) - \mathbf{F}(\mathbf{v}^a) \right\rangle_w = -\frac{1}{4}\beta \leq C \left\| \mathbf{v}^a - \mathbf{v}^b \right\|_w^2, \tag{161}$$

for all $\beta \geq 0$. Contrasting Eq. (161) with Eq. (155), notice that the $\frac{1}{2}\epsilon$ term is no longer present in the expression for the inner product, as a result of the additional "smoothness" of $\mathbf{F}(\cdot)$. Therefore, unlike in the case of $\mathbf{H}(\cdot)$, the OSL condition for $\mathbf{F}$ does not break down at $\mathbf{v}^a$ and $\mathbf{v}^b$.

The difference in drift patterns described above can also be observed in finite systems. The two graphs in Figure 7 display the *same* sample path of the embedded discrete-time Markov chain, in the representations of $\mathbf{S}^N$ and $\mathbf{V}^N$, respectively. Here $N = 10000$, $p = 1$, and $\lambda = 0.5$, with an initial condition $\mathbf{S}^N[0] = (1, 0.1, 0.1, 0, 0, \ldots)$ (i.e., 100 queues contain 2 tasks and the

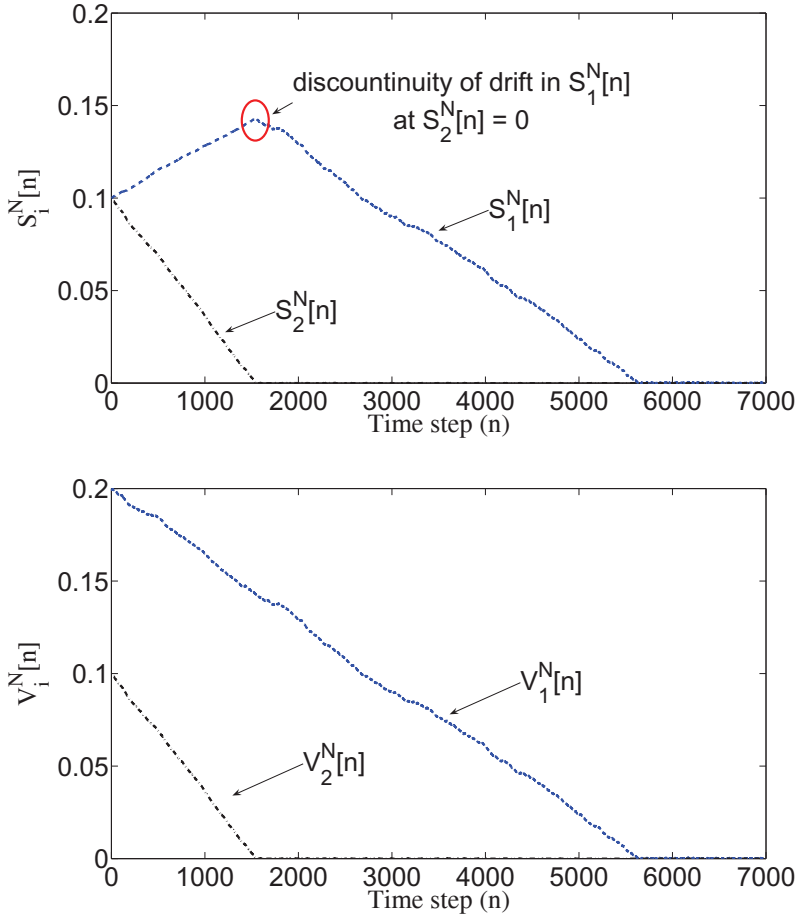

FIG 7. *Comparison between the* $\mathbf{V}^N[\cdot]$ *and* $\mathbf{S}^N[\cdot]$ *representations.*

rest of queues are empty). Notice that when $\mathbf{S}_2^N[n]$ hits zero, $\mathbf{S}_1^N[n]$ immediately receives an extra amount of downward drift. On the other hand, there is no change in drift for $\mathbf{V}_1^N[n]$ when $\mathbf{V}_2^N[n]$ hits zero. This is consistent with the previous analysis on the fluid models.

In summary, the difficulty of proving the uniqueness of fluid solutions is greatly reduced by choosing an appropriate state representation, $\mathbf{v}(\cdot)$. The fact that such a simple (linear) transformation from $\mathbf{s}(\cdot)$ to $\mathbf{v}(\cdot)$ can create one-sided Lipschitz continuity and greatly simplify the analysis may be of independent interest.

APPENDIX C: A FINITE-SUPPORT PROPERTY OF FLUID SOLUTION AND ITS IMPLICATIONS

In this section, we discuss a *finite-support* property of the fluid solution $\mathbf{v}(\cdot)$. Although this property is not directly used in the proofs of other results in our work, we have decided to include it here because it provides important, and somewhat surprising, qualitative insights into the system dynamics.

PROPOSITION 33. *Let* $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$*, and let* $\mathbf{v}(\mathbf{v}^0, \cdot)$ *be the unique solution to the fluid model with initial condition* $\mathbf{v}(\mathbf{v}^0, 0) = \mathbf{v}^0$*. If* $p > 0$*, then* $\mathbf{v}(\mathbf{v}^0, t)$ *has a finite support for all* $t > 0$*, in the sense that*

$$\sup\left\{i : \mathbf{v}_i(\mathbf{v}^0, t) > 0\right\} < \infty, \quad \forall t > 0. \tag{162}$$

Before presenting the proof, we observe that the finite-support property stated in Proposition 33 is *independent* of the size of the support of the initial condition $\mathbf{v}^0$; even if all coordinates of $\mathbf{v}(t)$ are strictly positive at $t = 0$, the support of $\mathbf{v}(t)$ immediately "collapses" to a finite number for any $t > 0$.

A critical assumption in Proposition 33 is that $p > 0$, i.e., that the system has a non-trivial central server. The "collapse" of $\mathbf{v}(\cdot)$ into a finite support is essentially due to the fact that the central server always allocates its service capacity to the longest queues in the system. Proposition 33 illustrates that the worst-case queue-length in the system is well under control *at all times*, thanks to the power of the central server.

Proposition 33 also sheds light on the structure of the invariant state of the fluid model, $\mathbf{v}^I$. Recall from Theorem 2 that $\mathbf{v}^I$ has a finite support *whenever* $p > 0$. Since by the global stability of fluid solutions (Theorem 4), we have that

$$\lim_{t\to\infty} \left\|\mathbf{v}\left(t\right) - \mathbf{v}^I\right\|_w = 0, \tag{163}$$

the fact that $\mathbf{v}(t)$ admits a finite support for any $t > 0$ whenever $p > 0$ provides strong intuition for and partially explains the finite-support property of $\mathbf{v}^I$.

We now prove Proposition 33.

PROOF OF PROPOSITION 33. We fix some $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$, and for the rest of the proof we will write $\mathbf{v}(\cdot)$ in place of $\mathbf{v}(\mathbf{v}^0, \cdot)$. It is not difficult to show, by directly inspecting the drift of the fluid model in Eq. (4), that if we start with an initial condition $\mathbf{v}^0$ with a finite support, then the support remains

finite at all times. Hence, we now assume $\mathbf{v}_i^0 > 0$ for all i. First, the fact that $\mathbf{v}^0 \in \overline{\mathcal{V}}^\infty$ (i.e., $\mathbf{v}_1^0 < \infty$) implies

$$\lim_{i\to\infty} \mathbf{v}_i^0 = 0. \tag{164}$$

This is because all coordinates of the corresponding vector $\mathbf{s}^0$ are non-negative, and

$$\mathbf{v}_i^0 = \mathbf{v}_1^0 - \sum_{j=1}^{i-1} \mathbf{s}_j^0, \tag{165}$$

where the second term on the right-hand side converges to $\mathbf{v}_1^0$.

Assume that $\mathbf{v}_i(t) > 0$ for all i, over some small time interval $t \in [0, s]$. Since the magnitude of the drift on any coordinate $\mathbf{v}_i$ is uniformly bounded from above by $\lambda + 1$, and $\lim_{i\to\infty} \mathbf{v}_i^0 = 0$, for any $\epsilon > 0$ we can find $s', N > 0$ such that for all $i \geq N$ and $t \in [0, s']$,

$$\dot{\mathbf{v}}_i(t) = \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i) - (1-p)(\mathbf{v}_i - \mathbf{v}_{i+1}) - g_i(\mathbf{v}) \leq \epsilon - g_i(\mathbf{v}) = -p + \epsilon. \tag{166}$$

Since $\lim_{i\to\infty} \mathbf{v}_i^0 = 0$, Eq. (166) shows that it is impossible to find any strictly positive time interval $[0, s]$ during which the fluid trajectory $\mathbf{v}(t)$ maintains an infinite support. This proves the claim. □

APPENDIX D: SIMULATION SETUP

The simulation results shown in Figure 4 for a finite system with 100 stations were obtained by simulating the embedded discrete-time Markov chain, $\{Q[n]\}_{n\in\mathbb{N}}$, where the vector $Q[n] \in \mathbb{Z}_+^{100}$ records the queue lengths of all 100 queues at time step n. Specifically, we start with $Q[1] = 0$, and, during each time step, one of the following takes place:

1. With probability $\frac{\lambda}{1+\lambda}$, a queue is chosen uniformly at random from all queues, and one new task is added to this queue. This corresponds to an arrival to the system.
2. With probability $\frac{1-p}{1+\lambda}$, a queue is chosen uniformly at random from all queues, and one task is removed from the queue if the queue is non-empty. If the chosen queue is empty, no change is made to the queue length vector. This corresponds to the generation of a local service token.
3. With probability $\frac{p}{1+\lambda}$, a queue is chosen uniformly at random from the *longest queues*, and one task is removed from the chosen queue if the queue is non-empty. If all queues are empty, no change is made to the queue length vector. This corresponds to the generation of a central service token.

To make the connection between the above discrete-time Markov chain $Q[n]$ and the continuous-time Markov process $Q(t)$ considered in this paper, one can show that $Q(t)$ is uniformized and hence the steady-state distribution of $Q(t)$ coincides with that of the embedded discrete-time chain $Q[n]$.

To measure the steady-state queue length distribution seen by a typical task, we sampled from the chain $Q[n]$ in the following fashion: $Q[n]$ was first run for a burn-in period of 1,000,000 time steps, after which 500,000 samples were collected with 20 time steps between adjacent samples, where each sample recorded the current length of a queue chosen uniformly at random from all queues. Denote by $\mathbf{S}$ the set of all samples. The average queue length, as marked by the symbol "×" in Figure 4, was computed by taking the average over $\mathbf{S}$. The upper (UE) and lower (LE) ends of the 95% confidence intervals were computed by:

$$\begin{aligned} UE &\stackrel{\triangle}{=} \min\{x \in \mathbf{S}: \text{ there are no more than } 2.5\% \\ &\qquad \text{of the elements of } \mathbf{S} \text{ that are strictly greater than } x\}, \\ LE &\stackrel{\triangle}{=} \max\{x \in \mathbf{S}: \text{ there are no more than } 2.5\% \\ &\qquad \text{of the elements of } \mathbf{S} \text{ that are strictly less than } x\}. \end{aligned}$$

Note that this notion of confidence interval is meant to capture the concentration of $\mathbf{S}$ around the mean, and is somewhat different from that used in the statistics literature for parameter estimation.

A separate version of the above experiment was run for each value of λ marked in Figure 4, while the the level of centralization p was fixed at 0.05 across all experiments.

REFERENCES

[1] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems: Theory and Applications*, 30: pp. 89–148, 1998. MR1663763

[2] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence (2nd edition).* Wiley-Interscience, 2005. MR0838085

[3] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich, "Queueing system with selection of the shortest of two queues: An asymptotic approach," *Probl. Inf. Transm*, 32(1): 20–34, 1996. MR1384927

[4] M. Mitzenmacher, "The power of two choices in randomized load balancing," *Ph.D. thesis, U.C. Berkeley,* 1996. MR2695522

[5] M. Alanyali and M. Dashouk, "On power-of-choice in downlink transmission scheduling," *Inform. Theory and Applicat. Workshop*, U.C. San Diego, 2008.

[6] N. Gast and B. Gaujal, "Mean field limit of non-smooth systems and differential inclusions," *INRIA Research Report*, 2010.

[7] M. Bramson, Y. Lu, and B. Prabhakar, "Randomized load balancing with general service time distributions," *ACM Sigmetrics*, New York, 2010.

[8] M. Mitzenmacher, A. Richa, and R. Sitaraman, "The power of two random choices: A survey of techniques and results," *Handbook of Randomized Computing: Volume 1*, 255–312, 2001. MR1966907

[9] W. Jordan and S. C. Graves, "Principles on the benefits of manufacturing process flexibility," *Management Science*, 41(4):577–594, 1995.

[10] D. Simchi-Levi and Y. Wei, "Understanding the performance of the long chain and sparse designs in process flexibility," submitted, 2011.

[11] S. C. Graves and B. T. Tomlin, "Process flexibility in supply chains," *Management Science*, 49:289–328, 2003.

[12] S. Gurumurthi and S. Benjaafar, "Modeling and analysis of flexible queueing systems," *Management Science*, 49:289–328, 2003. MR2071833

[13] S. M. Iravani, M. P. Van Oyen, and K. T. Sims, "Structural flexibility: A new perspective on the design of manufacturing and service operations," *Management Science*, 51(2):151–166, 2005.

[14] R. Wallace and W. Whitt, "A staffing algorithm for call centers with skill-based routing," *Manufacturing and Service Operations Management*, 7:276–294, 2005.

[15] A. Mandelbaum and M. I. Reiman, "On pooling in queueing networks," *Management Science*, 44(7):971-981, 1998.

[16] J. M. Harrison and M. J. Lopez, "Heavy traffic resource pooling in parallel-server systems," *Queueing Systems*, 33:39-368, 1999. MR1742575

[17] S. L. Bell and R. J. Williams, "Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy," *Ann. Appl. Probab.*, 11(3): 608-649, 2001. MR1865018

[18] A. Mandelbaum and A. L. Stolyar, "Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule," *Operations Research*, 52(6):836-855, 2004. MR2104141

[19] A. Bassamboo, R. S. Randhawa, and J. A. Van Mieghem, "A little flexibility is all you need: on the asymptotic value of flexible capacity in parallel queuing systems," submitted, 2011.

[20] G. J. Foschini and J. Salz, "A basic dynamic routing problem and diffusion," *IEEE Trans. on Comm.* 26:320–327, 1978.

[21] Y. T. He and D. G. Down, "On accommodating customer flexibility in service systems," *INFOR*, 47(4): 289–295, 2009. MR2759824

[22] V. Acary and B. Brogliato, *Numerical Methods for Nonsmooth Dynamical Systems: Applications in Mechanics and Electronics*, Springer Verlag, 2008.

[23] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. on Inform. Theory*, 30: 466–478, 1993. MR1224342

[24] J. R. Norris, *Markov Chains*, Cambridge University Press, 1997. MR1600720

[25] S. Foss and T. Konstantopoulos, "An overview of some stochastic stability methods," *Journal of Operations Research Society of Japan*, 47(4), 2004. MR2174067

[26] D. Gamarnik and D. Goldberg, "Steady-state GI/GI/n queue in the Halfin-Whitt regime," Submitted to the *Annals of Applied Probability*, 2011.

[27] Borel-Cantelli Lemma, *Wikipedia*, http://en.wikipedia.org/wiki/Borel-Cantelli_lemma.

[28] Gronwall's Inequality, *Wikipedia*, http://en.wikipedia.org/wiki/Gronwall's_inequality.

[29] K. Xu. On the power of centralization in distributed processing. S.M. thesis, MIT, 2011. http://arxiv.org/pdf/1203.5026v1.pdf.

Laboratory of Information and Decision Systems,
Massachusetts Institute of Technology,
Cambridge, MA, 02139, USA
E-mail: jnt@mit.edu
kuangxu@mit.edu