On the Power of (even a little) Centralization in Distributed Processing

John N. Tsitsiklis *
MIT, LIDS
Cambridge, MA 02139
jnt@mit.edu

Kuang Xu *
MIT, LIDS
Cambridge, MA 02139
kuangxu@mit.edu

ABSTRACT

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 *steady-state delay*, 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 centralization.

We prove convergence to a fluid limit, and characterize both the transient and steady-state behavior of the finite system, in the limit as the number of stations N goes to infinity. We show that the queue-length process 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.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Queuing theory, Markov processes; C.2.1 [**Network Architecture and Design**]: Centralized networks, Distributed networks

* Research supported in parts by an MIT Jacobs Presidential Fellowship, an MIT-Xerox Fellowship, a Siebel Scholarship, and NSF grant CCF-0728554.

SIGMETRICS'11, June 7–11, 2011, San Jose, California, USA.

General Terms

Performance, Theory

Keywords

Phase transition, Dynamic resource allocation, Partial centralization

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 utilization of computing resources at the cost of implementation complexity and communication overhead. A natural question is how performance varies with the *degree of centralization*. 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 clusters, which involve a trade-off between performance (e.g., delay) and cost (e.g., communication infrastructure, energy consumption, etc.). In this paper, we address this problem by formulating and analyzing a multi-server model with an *adjustable* 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 by an independent stream of tasks, arriving at a rate of λ tasks per second, with $0 < \lambda < 1$.[1] Each station is equipped with a *local server* with identical performance; the server is local in the sense that it only serves its own station. All stations are also connected to a single *centralized server* which will serve a station with the longest queue whenever possible.

We consider an N-station system. The system designer is granted a total amount of N 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 is faced with the problem of allocating 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

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

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 $p \in (0, 1)$.

When the arrival process 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 to optimize system performance? Alternatively, we can ask an even more fundamental question: is there any 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 the system depicted in Figure 2. The arrival assumptions are the same as in the 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. Otherwise, it processes a task from a station selected uniformly at random; if the randomly chosen station is empty, 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. Hence, while it is intuitively obvious that longest-queue-first (LQF) scheduling is more desirable, it may not always be possible to obtain up-to-date information on the system state (i.e., the queue lengths at all stations). Thus, the central server may be forced to allocate service blindly. In this setting, a system designer is interested in setting the optimal *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 the current paper.

1.3 Overview of Main Contributions

We provide here an overview of the main contributions. Exact statements of our 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, 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).

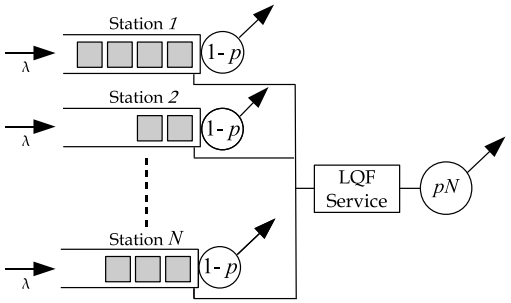

Figure 1: Server Farm with Local and Central Servers

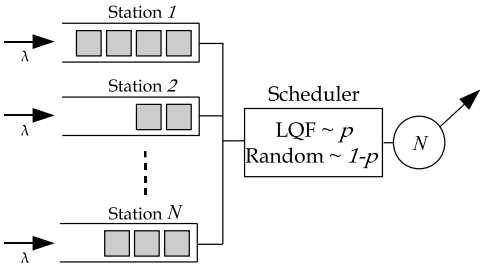

Figure 2: Centralized Scheduling with Communication Constraints

Fluid approximations typically lead to 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): 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] In terms of the question raised at the end of Section 1.1, this suggests that for large systems, even a small degree of centralization in-

[2]The $\sim$ notation used in this paper is to be understood as *asymptotic closeness* in the following sense: $f(x) \sim g(x)$, 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.

deed provides significant improvements in the system's delay performance, in the heavy traffic regime.

2. On the **technical end**, we show:

   (a) Given any finite initial queue sizes, and with high probability, the evolution of the queue length process can be approximated by the unique solution to a fluid model, over any finite time interval, as $N \to \infty$.

   (b) All solutions to the fluid model converge to a unique invariant state, as time $t \to \infty$, for any finite initial condition (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 distributions in (c) also becomes non-trivial due to the presence of discontinuous drifts. To derive this result, we will first establish the uniqueness of solutions to the fluid model and a uniform speed of convergence of stochastic sample paths to the solution of the fluid model over a compact set of initial conditions.

1.4 Related Work

To the best of our knowledge, the proposed model for the splitting of computing resources between distributed 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. A variation of this approach in a scheduling setting with channel uncertainties is examined in [5], but $\mathbf{s}_i$ 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 tail probabilities $\mathbf{s}_i$ become zero for sufficiently large finite i, which is significantly 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 generalized the supermarket model to arrival and processing times with general distributions. Since the queue length process is no longer Markov, the authors reply on an asymptotic independence property of the limiting system and use tools from statistical physics to establish convergence.

Our system remains 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 needed 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, 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 results 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 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.

Finally, there has been some work on the impact of service flexibilities in routing problems motivated by applications such as multilingual call centers. These date back to the seminal work of [9], with a more recent numerical study in [10]. 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 proofs, and the reader is referred to Section 4 for an overview of the proof structure. Due to space limitations, some of the proofs are sketched, omitted, or relegated to [11].

2. MODEL AND NOTATION

2.1 Model

We present our model using language 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 contains a queue which stores the tasks to be processed. The queue length (i.e., number of tasks) at station i at time t is denoted by $Q_i(t)$, $i \in \{1, 2, \ldots, N\}$. For now, we do not make any assumptions on the queue lengths at time $t = 0$, other than that they are finite.

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 i 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 i ticks at time t, we say that a **local service token** is generated at station i. If $Q_i(t) \neq 0$, exactly one task from station i "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.[4]

   (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 t (i.e., $\sum_{i=1}^{N} Q_i(t) > 0$), exactly one task from some station i, 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.

Equivalence between two the interpretations. We comment 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 process a task at a random station (corresponding to a *local service token*), or 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 described above.

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 token 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

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

Each coordinate $\mathbf{S}_i^N(t)$ represents the fraction of queues with at least i tasks. 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 $\mathbf{S}_i^N(t) = \mathbf{V}_i^N(t) - \mathbf{V}_{i+1}^N(t)$, and that $\mathbf{V}_1^N(t) = \sum_{j=1}^{\infty} \mathbf{S}_j^N(t)$ is equal to the *average queue length* in the system at time t. 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. 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

The following sets will be used throughout the paper (where M is a positive integer):

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

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

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

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

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

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

$$\|\mathbf{x} - \mathbf{y}\|_w^2 = \sum_{i=0}^{\infty} \frac{|\mathbf{x}_i - \mathbf{y}_i|^2}{2^i}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathbb{Z}^+}. \tag{4}$$

We will be using bold letters to denote vectors, and either bold or ordinary letters for scalars. Upper-case letters are in general reserved for random variables (e.g., $\mathbf{V}^{(0,N)}$) or scholastic processes (e.g., $\mathbf{V}^N(t)$), and lower-case letters are used for constants (e.g., $\mathbf{v}^0$) and deterministic functions

[4]The generation of a token can also be thought of as a 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. The same interpretation holds for the central service token.

(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

3.1 Definition of Fluid Model

Before introducing the main results, we first define the fluid model, with some intuitive justification.

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:*

(1) $\mathbf{v}(0) = \mathbf{v}^0$

(2) For all $t \geq 0$, $\mathbf{v}_0(t) - \mathbf{v}_1(t) = 1$ *and* $1 \geq \mathbf{v}_i(t) - \mathbf{v}_{i+1}(t) \geq \mathbf{v}_{i+1}(t) - \mathbf{v}_{i+2}(t) \geq 0$ *for all* $i \geq 0$.

(3) For almost all $t \in [0,\infty)$, *every* $\mathbf{v}_i(t)$ *is differentiable and satisfies*

$$\dot{\mathbf{v}}_i(t) = \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i) - (1-p)(\mathbf{v}_i - \mathbf{v}_{i+1}) - g_i(\mathbf{v}), \quad (5)$$

where

$$g_i(\mathbf{v}) = \begin{cases} p, & \mathbf{v}_i > 0, \\ \min\{\lambda \mathbf{v}_{i-1}, p\}, & \mathbf{v}_i = 0, \mathbf{v}_{i-1} > 0, \\ 0, & \mathbf{v}_i = 0, \mathbf{v}_{i-1} = 0, \end{cases} \quad (6)$$

We can write Eq. (5) more compactly as

$$\dot{\mathbf{v}}(t) = \mathbf{F}(\mathbf{v}), \quad (7)$$

where $\mathbf{F}(\mathbf{v})$ is called the drift at point $\mathbf{v}$.

Interpretation of the fluid model. The 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 (1) and (2) correspond to initial and boundary conditions, respectively. Before rigorously establishing the validity of approximation, we provide some intuition for each of the drift terms in Eq. (5):

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 one that is 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 $N \to \infty$, and multiply by the total arrival rate, $N\lambda$, times 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 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 to keep $\mathbf{v}_i$ at zero. Thus, the total rate of central service tokens dedicated to $\mathbf{v}_i$ matches exactly the rate of increase of $\mathbf{v}_i$ due to arrivals.[5]

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 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

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. $\mathbf{F}(\mathbf{v}^I) = 0$). Letting $\mathbf{s}_i^I \triangleq \mathbf{v}_i^I - \mathbf{v}_{i+1}^I$ for all $i \geq 0$, the exact expression for the invariant state is given as follows:*

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

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

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

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

where $\tilde{i}^*(p,\lambda) \triangleq \left\lfloor \frac{1-p}{p} \right\rfloor$.[6]

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

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

where

$$i^*(p,\lambda) \triangleq \left\lfloor \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda} \right\rfloor, \quad (8)$$

Proof. The proof consists of simple algebra to compute the solution to $\mathbf{F}(\mathbf{v}^I) = 0$. See Appendix A.1 in [11] for a proof.

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

[5]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.

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

$\sum_{i=1}^{\infty} \mathbf{s}_i^I$ 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*

$$\mathbf{v}_1^I \triangleq \sum_{i=1}^{\infty} \mathbf{s}_i^I = \frac{(1-p)(1-\lambda)}{(1-p-\lambda)^2}\left[1-\left(\frac{\lambda}{1-p}\right)^{i^*(p,\lambda)}\right] - \frac{p}{1-p-\lambda} i^*(p,\lambda), \quad (9)$$

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

$$\mathbf{v}_1^I \sim i^*(p,\lambda) \sim \log_{\frac{1}{1-p}} \frac{1}{1-\lambda}, \text{ as } \lambda \to 1. \quad (10)$$

The scaling of the average queue length in Eq. (10) with respect to arrival rate λ is contrasted with (and is *dramatically better* than) the familiar $\frac{1}{1-\lambda}$ scaling when no centralized resource is available ($p = 0$).

Intuition for Exponential Phase Transition. The exponential improvement in the scaling of $\mathbf{v}_1^I$ is surprising, because the expressions for $\mathbf{s}_i^I$ look ordinary and do not contain any super-geometric terms in i. However, a closer look reveals that for any $p > 0$, the tail probabilities $\mathbf{s}^I$ have **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 a tail probability with "finite-support" implies 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. $\diamond$

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 more quickly, and reaches zero at around $i = 40$. Figure 4 demonstrates the exponential phase transition in the average queue length as the traffic intensity reaches 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 theoretical model offers good predictions for even a moderate number of servers ($N = 100$) [7]. 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. $\diamond$

[7]The detailed simulation setup can be found in Appendix C in [11].

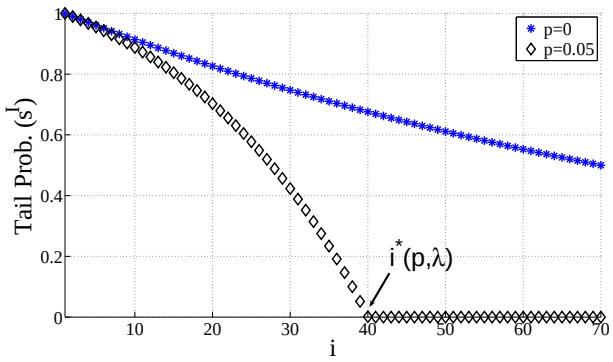

Figure 3: Values of $\mathbf{s}_i^I$, as a function of i, for $p = 0$ and $p = 0.05$, with traffic intensity $\lambda = 0.99$.

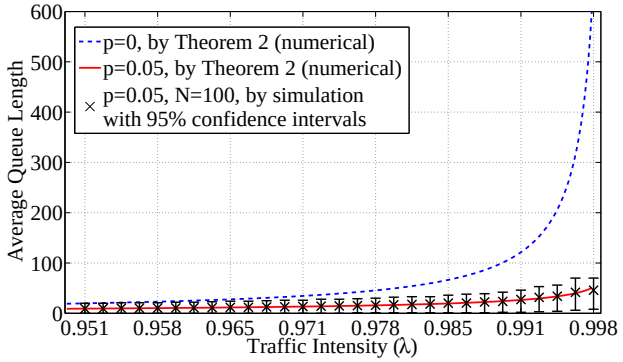

Figure 4: Illustration of the exponential improvement in average queue length from $O(\frac{1}{1-\lambda})$ to $O(\log \frac{1}{1-\lambda})$ as $\lambda \to 1$, when we compare $p = 0$ to $p = 0.05$.

Theorem 2 characterizes the invariant state of the fluid model, without saying if 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* $\mathbf{v}(t)$, $t \in [0,\infty)$.

Proof. See Section 7.1.

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*

$$\lim_{t\to\infty} \left\| \mathbf{v}\left(\mathbf{v}^0,t\right) - \mathbf{v}^I \right\|_w = 0, \quad (11)$$

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

Proof. See Section 7.3.

3.3 Convergence to a Fluid Solution - Finite-time and Steady-state

The two theorems in this section justify the use of the fluid model as an approximation for the finite stochastic system. The first theorem states that with high probability, the

$p = \backslash \lambda =$	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

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

evolution of the aggregated 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 as $N \to \infty$.

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

$$\lim_{N\to\infty} \mathbb{P}\left(\|\mathbf{V}^N(0) - \mathbf{v}^0\|_w > \gamma\right) = 0, \quad \forall \gamma > 0, \tag{12}$$

then

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \|\mathbf{V}^N(t) - \mathbf{v}\left(\mathbf{v}^0,t\right)\|_w > \gamma\right) = 0, \quad \forall \gamma > 0. \tag{13}$$

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 7.2.

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

$$\lim_{t\to\infty}\lim_{N\to\infty} \mathbf{V}^N(t) = \mathbf{v}^I, \text{ in distribution.} \tag{14}$$

If we switch the order in which the limits over t and N are taken in Eq. (14), the question becomes to describe 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. (14), we may expect that

$$\lim_{N\to\infty}\lim_{t\to\infty} \mathbf{V}^N(t) = \mathbf{v}^I, \text{ in distribution.} \tag{15}$$

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 system is positive recurrent, and $\mathbf{V}^N(t)$ admits a unique steady-state distribution π^N. Moreover,*

$$\lim_{N\to\infty} \pi^N = \mathbf{v}^I, \text{ weakly.} \tag{16}$$

PROOF. The proof is based on the tightness of the sequence of steady-state distributions π^N, and a uniform rate of convergence of $\mathbf{V}^N(\cdot)$ to $\mathbf{v}(\cdot)$ over any compact set of initial conditions. See Appendix B in [11] for a proof.

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

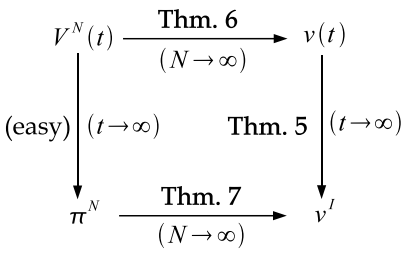

Figure 5: Relationships between convergence results.

4. PROOF OVERVIEW

The remainder of the paper will be devoted to proving the results summarized in Section 3. We begin by coupling the sample paths of processes of interest (e.g., $\mathbf{V}^N$) with those of two fundamental processes that drive the system dynamics (Section 5). This approach allows us to link deterministically the convergence properties of the sample paths of interest to the convergence 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 6 that almost all sample paths of $\mathbf{V}^N$ are "tight" in the sense that they are uniformly approximated by a set of Lipschitz-continuous trajectories, which we refer to as the fluid limits, as $N \to \infty$, and that all such fluid limits are valid solutions to the fluid model. This makes the connection between the finite stochastic system and the deterministic fluid solutions. Section 7 studies the properties of the fluid model, and provides proofs for Theorem 4 and 5. Note that Theorem 6 (convergence of $\mathbf{V}^N$ to the unique fluid solution, over a finite time horizon) now follows from the tightness results in Section 6 and the uniqueness of fluid solutions (Theorem 4). The proof of Theorem 2 stands alone, and due to space constraints, is included in Appendix A.1 in [11]. Finally, the proof of Theorem 7 (convergence of steady state distributions to $\mathbf{v}^I$), which is more technical, is given in Appendix B of [11].

5. PROBABILITY SPACE AND COUPLING

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

5.1 Definition of Probability Space

Definition 8. (Fundamental Processes and Initial Conditions)

(1) **The Total Event Process** $W(t)$*, 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** $U(n)$*, 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** $\{\mathbf{V}^{(0,N)}\}$ *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 in* $\overline{\mathcal{V}}^{\infty} \cap \mathcal{Q}^N$*.* [8] $\mathbf{V}^{(0,N)}$ *represents the initial queue length distribution.*

For the rest of the paper, we will be working with the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ defined as the **product space** of $(\Omega_W, \mathcal{F}_W, \mathbb{P}_W)$, $(\Omega_U, \mathcal{F}_U, \mathbb{P}_U)$ and $(\Omega_0, \mathcal{F}_0, \mathbb{P}_0)$. 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 $\mathbf{V}^{(0,N)}$.

5.2 A Coupled Construction of Sample Paths

Recall the interpretation of the fluid model drift terms in Section 3.1. Mimicking the expression of $\dot{\mathbf{v}}_i(t)$ in Eq. (5), we would like to decompose $\mathbf{V}_i^N(t)$ into three non-decreasing right-continuous processes,

$$\mathbf{V}_i^N(t) = \mathbf{V}_i^N(0) + \mathbf{A}_i^N(t) - \mathbf{L}_i^N(t) - \mathbf{C}_i^N(t), \quad i \geq 1, \quad (17)$$

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 event process, by $W^N(\omega, t) \triangleq W(\omega, Nt)$, $\forall t \geq 0$, $\omega \in \Omega$.

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:

(1) If $U(\omega, k) \in \frac{\lambda}{1+\lambda}\left[0, \mathbf{V}_{i-1}^N(\omega, t_k-) - \mathbf{V}_i^N(\omega, t_k-)\right)$ for some $i \geq 1$,[9], the event corresponds to an **arrival** to a station with at least $i-1$ tasks. Hence we increase $\mathbf{A}_i^N(\omega, t)$ by $\frac{1}{N}$ at all such i.

(2) If $U(\omega, k) \in \frac{\lambda}{1+\lambda} + \frac{1-p}{1+\lambda}\left[0, \mathbf{V}_i^N(\omega, t_k-) - \mathbf{V}_{i+1}^N(\omega, t_k-)\right)$ 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 $\mathbf{L}_i^N(\omega, t)$ 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) For all other values of $U(\omega, k)$, 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 in 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., $i^*(t) = \sup\{i : \mathbf{V}_i^N(\omega, t-) > 0\}$. We increase $\mathbf{C}_j^N(\omega, t)$ 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. (17), and keep the values of all processes unchanged between t_k and t_{k+1}. We set $\mathbf{V}_0^N \triangleq \mathbf{V}_1^N + 1$, just to stay consistent with the definition of $\mathbf{V}_0^N$.

6. FLUID LIMITS OF STOCHASTIC SAMPLE PATHS

In the sample-path-wise construction in Section 5.2, 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 $\mathbf{V}^N$ 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, over a finite time interval $[0, T]$, can be summarized as follows:

(1) Find 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) Show that for all ω in this nice set, the derived sample paths $\mathbf{V}^N$ are also "nice", and contain a subsequence converging to a Lipschitz-continuous trajectory $\mathbf{v}(\cdot)$, as $N \to \infty$.

(3) Characterize the derivative at any regular point[10] 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.

(4) Finally, show that given any finite initial condition, $\mathbf{v}(t)$ converges to a unique invariant state $\mathbf{v}^I$ as $t \to \infty$.

The proof will be presented according to the above order.

6.1 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* $\omega \in \mathcal{C}$*,*

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \left|W^N(\omega, t) - (1+\lambda)t\right| = 0, \quad (18)$$

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{[a,b)}(U(\omega, i)) = b - a, \ \forall [a, b) \subset [0, 1]. \quad (19)$$

Proof. Eq. (18) is based on the Functional Law of Large Numbers and Eq. (19) is a consequence of the Glivenko-Cantelli theorem. See Appendix A.2 in [11] for a proof.

[8] 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.

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

[10] Regular points are points where the derivative exists. Since the trajectories are Lipschitz-continuous, almost all points are regular.

Definition 10. *We call the 4-tuple,* $\mathbf{X}^N \triangleq \left(\mathbf{V}^N, \mathbf{A}^N, \mathbf{L}^N, \mathbf{C}^N\right)$, *the* N**th system**. *Note that all four components are infinite-dimensional processes.* [11]

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, $d(\cdot,\cdot)$:

$$d(x,y) \triangleq \sup_{t\in[0,T]} |x(t) - y(t)|, \; x,y \in D[0,T]. \tag{20}$$

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 $D^{\infty}[0,T]$:

$$d^{\mathbb{Z}^+}(\mathbf{x},\mathbf{y}) \triangleq \sup_{t\in[0,t]} \|\mathbf{x}(t) - \mathbf{y}(t)\|_w, \mathbf{x},\mathbf{y} \in D^{\mathbb{Z}^+}[0,T], \tag{21}$$

with $\|\cdot\|_w$ defined in Eq. (4).

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

$$\lim_{N\to\infty} \|\mathbf{V}^N(\omega,0) - \mathbf{v}^0\|_w = 0, \tag{22}$$

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*

$$|\mathbf{x}_i(a) - \mathbf{x}_i(b)| \le L|a-b|, \quad \forall a,b \in [0,T],\, i \in \mathbb{Z}^+, \tag{23}$$

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: 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 further subsequence with a Lipschitz-continuous limit. We then use this coordinate-wise convergence result 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 of multiclass queuing 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 a particular sample. 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 L-Lipschitz-continuous (closedness). This completes the proof for coordinate-wise convergence.

Using this coordinate-wise convergence, we now construct the limit points of $\mathbf{X}^N$ in the space $D^{\mathbb{Z}^+}[0,T]$. Let $\mathbf{v}_1(\cdot)$ be any L-Lipschitz-continuous limit point of $\mathbf{V}_1^N$, so that a subsequence $\mathbf{V}_1^{N_j^1}(\omega,\cdot) \to \mathbf{v}_1(\cdot)$, as $j \to \infty$, with respect to $d(\cdot,\cdot)$. Then, we proceed recursively by letting $\mathbf{v}_{i+1}(\cdot)$ 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. We claim that $\mathbf{v}$ is indeed a limit point of $\mathbf{V}^N$ under the norm $d^{\mathbb{Z}^+}(\cdot,\cdot)$. Note that since $\mathbf{v}_1(0) = \mathbf{v}_1^0$, $0 \le \mathbf{V}_i^N(t) \le \mathbf{V}_1^N(t)$, and $\mathbf{v}_1(\cdot)$ is L-Lipschitz-continuous, we have that

$$\sup_{t\in[0,T]} |\mathbf{v}_i(t)| \le \sup_{t\in[0,T]} |\mathbf{v}_1(t)| \le \left|\mathbf{v}_1^0\right| + LT, \quad \forall i \in \mathbb{Z}^+. \tag{24}$$

Set $N_1 = 1$, and let, for $k \ge 2$,

$$N_k = \min\left\{N \ge N_{k-1} : \sup_{1\le i\le k} d(\mathbf{V}_i^N(\omega,\cdot), \mathbf{v}_i) \le \frac{1}{k}\right\}. \tag{25}$$

Note that the construction of $\mathbf{v}$ implies that N_k is well defined and finite for all k. From Eqs. (24) and (25), we have, for all $k \ge 2$,

$$\begin{aligned} d^{\mathbb{Z}^+}\left(\mathbf{V}^{N_k}(\omega,\cdot),\mathbf{v}\right) &= \sup_{t\in[0,T]} \sqrt{\sum_{i=0}^{\infty} \frac{\left|\mathbf{V}_i^{N_k}(\omega,t) - \mathbf{v}_i(t)\right|^2}{2^i}} \\ &\le \frac{1}{k} + \sqrt{\left(|\mathbf{v}_1^0| + LT\right)^2 \sum_{i=k+1}^{\infty} \frac{1}{2^i}} \\ &= \frac{1}{k} + \frac{1}{2^{k/2}}\left(\left|\mathbf{v}_1^0\right| + LT\right). \end{aligned} \tag{26}$$

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. □

6.2 Derivatives of the Fluid Limits

The previous section established that any sequence of "good" sample paths ($\left\{\mathbf{X}^N(\omega,\cdot)\right\}$ 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 might expect, that they are the same as the drift terms in the fluid model. 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*

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

differentiability of all coordinates of $\mathbf{x}$. *Then, for all* $i \in \mathbb{N}$,

$$\dot{\mathbf{a}}_i(t) = \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i), \quad (27)$$

$$\dot{\mathbf{l}}_i(t) = (1-p)(\mathbf{v}_i - \mathbf{v}_{i+1}), \quad (28)$$

$$\dot{\mathbf{c}}_i(t) = g_i(\mathbf{v}), \quad (29)$$

where g *was defined in Eq.* (6), *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 $\left\{\mathbf{X}^{N_k}(\cdot)\right\}_{k=1}^{\infty}$ 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. (27)-(29)) separately, and 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 $\left\{A^{N_k}(\cdot)\right\}_{k=1}^{\infty}$. 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:

$$\mathbf{A}_i^{N_k}(t+\epsilon) - \mathbf{A}_i^{N_k}(t) = \frac{1}{N_k} \sum_{j=N_k W^{N_k}(t)}^{N_k W^{N_k}(t+\epsilon)} \mathbb{I}_{I_j}\big(U(j)\big), \quad (30)$$

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 jth jump in $W^N(\cdot)$, i.e.,

$$t_j^N \triangleq \inf\left\{t \geq 0 : W^N(t) \geq \frac{j}{N}\right\}. \quad (31)$$

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

$$\lim_{k\to\infty}\left(\mathbf{A}_i^{N_k}(t+\epsilon) - \mathbf{A}_i^{N_k}(t)\right) = \mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t). \quad (32)$$

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* $\epsilon > 0$

$$|\mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t) - \epsilon\lambda(\mathbf{v}_{i-1}(t) - \mathbf{v}_i(t))| \leq 2\epsilon^2 L \quad (33)$$

Proof outline: The proof is based on the fact that $\omega \in \mathcal{C}$. Using Lemma 9, Eq. (33) follows from Eq. (30) by applying the convergence properties of $W^N(t)$ (Eq. (18)) and $U(n)$ (Eq. (19)). See Appendix A.3 in [11] for a proof. □

Since by assumption $\mathbf{a}(\cdot)$ is differentiable at t, Claim 1 follows from Lemma 13 by noting $\dot{\mathbf{a}}_i(t) \triangleq \lim_{\epsilon\downarrow 0} \frac{\mathbf{a}_i(t+\epsilon) - \mathbf{a}_i(t)}{\epsilon}$.

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: $\dot{\mathbf{c}}_i(t) = g_i(\mathbf{v})$. 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

$$\dot{\mathbf{c}}_i(t) = \dot{\mathbf{a}}_i(t) - \dot{\mathbf{l}}_i(t) - \dot{\mathbf{v}}_i(t). \quad (34)$$

We calculate each of the three terms on the right-hand side of the above equation. By Claim 1, $\dot{\mathbf{a}}_i(t) = \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i) = 0$, and by Claim 2, $\dot{\mathbf{l}}_i(t) = \lambda(\mathbf{v}_i - \mathbf{v}_{i+1}) = 0$. 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**: $\dot{\mathbf{c}}_i(t) = \min\{\lambda\mathbf{v}_{i-1}, p\}$, if $\mathbf{v}_i = 0, \mathbf{v}_{i-1} > 0$.

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 $\mathbf{v}_{i-1}(t) \leq \frac{p}{\lambda}$. Here the line of arguments is similar to the one in Case 1. By Claim 1, $\dot{\mathbf{a}}_i(t) = \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i) = \lambda\mathbf{v}_{i-1}$, and by Claim 2, $\dot{\mathbf{l}}_i(t) = \lambda(\mathbf{v}_i - \mathbf{v}_{i+1}) = 0$. Using again the same trick as in Case 1, the non-negativity of $\mathbf{v}_i$ 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

$$\dot{\mathbf{c}}_i(t) = -\dot{\mathbf{v}}_i(t) + \dot{\mathbf{a}}_i(t) - \dot{\mathbf{l}}_i(t) = \lambda\mathbf{v}_{i-1}. \quad (35)$$

Intuitively, here the drift due to random arrivals to queues with $i-1$ tasks, $\lambda\mathbf{v}_{i-1}$, 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 $\dot{\mathbf{c}}_i(t) \leq p$. This simply means $\mathbf{v}_i(t)$ cannot be differentiable at time t, if $\mathbf{v}_i(t) = 0, \mathbf{v}_{i-1}(t) > \frac{p}{\lambda}$. 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 $k \geq K$, $s \in [t, t+\epsilon]$

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

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, $\dot{\mathbf{v}}_i(t) = \lim_{\epsilon\downarrow 0}\lim_{k\to\infty} \frac{\mathbf{V}_i^{N_k}(t+\epsilon) - \mathbf{V}_i^{N_k}(t)}{\epsilon} = p$.

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

7. PROPERTIES OF THE FLUID MODEL

7.1 Uniqueness of Fluid Limit & 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 $\mathbf{v}(t)$ 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).

PROOF. **(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 $i^p(\mathbf{v}) \triangleq \sup\{i : \mathbf{v}_i > 0\}$.[12] 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 holding if both are infinite. Denoting by $\mathbf{a}^{\mathbf{v}}$ the arrival process $\mathbf{a}$ corresponding to the fluid limit $\mathbf{v}$ (and similarly for $\mathbf{l}$ and $\mathbf{c}$), we have:

$$
\begin{aligned}
&\frac{d}{dt}\left\|\mathbf{v}-\mathbf{w}\right\|_w^2 \triangleq \frac{d}{dt}\sum_{i=0}^{\infty}\frac{\left|\mathbf{v}_i-\mathbf{w}_i\right|^2}{2^i} \overset{(a)}{=} \sum_{i=0}^{\infty}\frac{\left(\mathbf{v}_i-\mathbf{w}_i\right)\left(\dot{\mathbf{v}}_i-\dot{\mathbf{w}}_i\right)}{2^{i-1}} \\
= &\sum_{i=0}^{\infty}\frac{\left(\mathbf{v}_i-\mathbf{w}_i\right)\left[\left(\dot{\mathbf{a}}_i^{\mathbf{v}}-\dot{\mathbf{l}}_i^{\mathbf{v}}\right)-\left(\dot{\mathbf{a}}_i^{\mathbf{w}}-\dot{\mathbf{l}}_i^{\mathbf{w}}\right)\right]}{2^i} \\
&-\sum_{i=0}^{\infty}\frac{\left(\mathbf{v}_i-\mathbf{w}_i\right)\left(\dot{\mathbf{c}}_i^{\mathbf{v}}-\dot{\mathbf{c}}_i^{\mathbf{w}}\right)}{2^{i-1}} \\
\overset{(b)}{\leq} &C\left\|\mathbf{v}-\mathbf{w}\right\|_w^2-\sum_{i=0}^{\infty}\frac{\left(\mathbf{v}_i-\mathbf{w}_i\right)\left(\dot{\mathbf{c}}_i^{\mathbf{v}}-\dot{\mathbf{c}}_i^{\mathbf{w}}\right)}{2^{i-1}} \\
= &C\left\|\mathbf{v}-\mathbf{w}\right\|_w^2-\sum_{i=0}^{i^p(\mathbf{v})}\frac{1}{2^{i-1}}\left(\mathbf{v}_i-\mathbf{w}_i\right)(p-p) \\
&-\frac{1}{2^{i^p(\mathbf{v})}}(0-\mathbf{w}_{i^p(\mathbf{v})+1})(\min\{\lambda\mathbf{v}_{i^p(\mathbf{v})},p\}-p) \\
&-\sum_{i=i^p(\mathbf{v})+2}^{i^p(\mathbf{w})}\frac{1}{2^{i-1}}(0-\mathbf{w}_i)(0-p) \\
&-\sum_{j=i^p(\mathbf{w})+1}^{\infty}\frac{1}{2^{i-1}}(0-0)\left(\dot{\mathbf{c}}_i^{\mathbf{v}}-\dot{\mathbf{c}}_i^{\mathbf{w}}\right) \\
\leq &C\left\|\mathbf{v}-\mathbf{w}\right\|_w^2, \qquad (36)
\end{aligned}
$$

where $C = 6(\lambda + 1 - p)$. The existence of the derivative $\frac{d}{dt}\left\|\mathbf{v}-\mathbf{w}\right\|_w^2$ and the exchange of limits in (a) require some care, but they are based on the fact that $\mathbf{v}_i(t)$ and $\mathbf{w}_i(t)$ are L-Lipschitz for all i[13]. Step (b) follows from the fact that $\dot{\mathbf{a}}$ and $\dot{\mathbf{l}}$ are both continuous and linear in $\mathbf{v}$ (see Eqs. (27) – (29)). The specific value of C can be derived after some simple algebra. Now suppose that $\mathbf{v}^0 = \mathbf{w}^0$. By Gronwall's inequality and Eq. (36), we have

$$\left\|\mathbf{v}(t)-\mathbf{w}(t)\right\|_w^2 \leq \left\|\mathbf{v}(0)-\mathbf{w}(0)\right\|_w^2 e^{Ct} = 0,\ \forall t \in [0,\infty), \quad (37)$$

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

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

Corollary 14. (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 $\mathbf{w}^n \in \overline{\mathcal{V}}^\infty, \forall n$, and $\left\|\mathbf{w}^n-\mathbf{v}^0\right\|_w \to 0$ as $n \to \infty$, then for all $t \geq 0$,*

$$\lim_{n\to\infty}\left\|\mathbf{v}(\mathbf{w}^n,t)-\mathbf{v}(\mathbf{v}^0,t)\right\|_w = 0. \quad (38)$$

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

[13] In particular, this implies that there exists $L' > 0$ such that $h_i(t) \triangleq |\mathbf{v}_i(t)-\mathbf{w}_i(t)|^2$ are L'-Lipschitz in a small neighborhood around t for all i, i.e. $\left|\frac{h_i(t+\epsilon)-h_i(t)}{\epsilon}\right| \leq L'$ for all i and all sufficiently small ϵ. The existence of $\frac{d}{dt}\left\|\mathbf{v}-\mathbf{w}\right\|_w^2$ and the exchange of limits then follow from the dominated convergence theorem. See a more elaborate version of this proof in [11] for details.

PROOF. The continuity with respect to the initial condition is a direct consequence of Eq. (37): if $\mathbf{v}(\mathbf{w}^n,\cdot)$ is a sequence of fluid limits with initial conditions $\mathbf{w}^n \in \overline{\mathcal{V}}^\infty$ and if $\left\|\mathbf{w}^n-\mathbf{v}^0\right\|_w^2 \to 0$ as $N \to \infty$, then for all $t \in [0,\infty)$,

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

This completes the proof. □

$\mathbf{V}^N(t)$ versus $\mathbf{S}^N(t)$: The above uniqueness proof (Theorem 4) demonstrates the power of using $\mathbf{V}^N(t)$ and $\mathbf{v}(t)$ as a state representation. The proof technique exploits a property of the drifts, also known as the one-sided-Lipschitz condition in the dynamical systems literature. In fact, if we instead use $\mathbf{s}(t)$ to construct the fluid mode, the resulting drift terms, given by the relation $\mathbf{s}_i(t) = \mathbf{v}_i(t) - \mathbf{v}_{i+1}(t)$, fail to be one-sided-Lipschitz-continuous. The uniqueness result should still hold, but the proof would be much more difficult, requiring an examination of all points of discontinuity in the space. The intuitive reason is that the total 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 create one-sided-Lipschitz continuity and greatly simplify the analysis seems interesting in itself.

7.2 Proof of Theorem 6

PROOF. **(Theorem 6)** The proof follows from the sample-path tightness in Proposition 11 and the uniqueness of the fluid limit 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 $T < \infty$,

$$\lim_{N\to\infty}\sup_{t\in[0,T]}\left\|\mathbf{V}^N(\omega,t)-\mathbf{v}(\mathbf{v}^0,t)\right\|_w = 0,\ \mathbb{P}\text{–almost surely}, \quad (39)$$

which implies convergence in probability, and Eq. (13) holds.□

7.3 Convergence to the Invariant State $\mathbf{v}^I$ (Proof of Theorem 5)

In this section, we will switch to the alternative state representation, $\mathbf{s}(t)$, where

$$\mathbf{s}_i(t) \triangleq \mathbf{v}_{i+1}(t) - \mathbf{v}_i(t),\ \forall i \geq 0 \quad (40)$$

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 an invariant state. We note that $\mathbf{s}_0(t) = 1$ for all t, and 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. (40).

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 cannot use $\mathbf{v}$ but must rely on $\mathbf{s}$ to establish the result.

Definition 15. (Coordinate-wise Dominance) *For any $\mathbf{s}, \mathbf{s}' \in \overline{\mathcal{S}}^{\infty}$, we write $\mathbf{s} \succeq \mathbf{s}'$ if $\mathbf{s}_i \geq \mathbf{s}'_i$, for all $i \geq 0$.*

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

Lemma 16. *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 $\mathbf{s}^1(t) \succeq \mathbf{s}^2(t), \forall t \geq 0$.*

The proof of Lemma 16 consists of checking the drift terms of the fluid model. It is straightforward and is omitted. We are now ready to prove Theorem 5.

PROOF. **(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 $\mathbf{s}^u(0) \succeq \mathbf{s}(0) \succeq \mathbf{s}^l(0)$ and $\mathbf{s}^u(0) \succeq \mathbf{s}^I \succeq \mathbf{s}^l(0)$. By Lemma 16, we must have $\mathbf{s}^u(t) \succeq \mathbf{s}^I \succeq \mathbf{s}^l(t)$ 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, for any regular $t > 0$,

$$\begin{aligned}
\dot{\mathbf{v}}_i(t) &= \lambda(\mathbf{v}_{i-1} - \mathbf{v}_i) - (1-p)(\mathbf{v}_i - \mathbf{v}_{i+1}) - g_i(\mathbf{v}) \\
&= \lambda \mathbf{s}_{i-1} - (1-p)\mathbf{s}_i - g_i(\mathbf{v}) \\
&= (1-p)\left(\frac{\lambda \mathbf{s}_{i-1} - g_i(\mathbf{v})}{1-p} - \mathbf{s}_i\right). \qquad (41)
\end{aligned}$$

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. (41) and the fact that $\mathbf{s}_0^u = \mathbf{s}_0^I = 1$, we have

$$\dot{\mathbf{v}}_1^u(t) = (1-p)\left(\frac{\lambda - g_1(\mathbf{v}^u)}{1-p} - \mathbf{s}_1^u\right) \leq (1-p)\left(\mathbf{s}_1^I - \mathbf{s}_1^u\right), \quad (42)$$

for all regular $t \geq 0$. To see why the above inequality holds, note that $\frac{\lambda - g_1(\mathbf{v}^u)}{1-p} = \frac{\lambda - p}{1-p} \leq \mathbf{s}_1^I$ whenever $\mathbf{s}_1^u(t) > 0$, and $\frac{\lambda - g_1(\mathbf{v}^u)}{1-p} = \mathbf{s}_1^u(t) = 0$ whenever $\mathbf{s}_1^u(t) = \mathbf{s}_1^I = 0$.

Since $\mathbf{v}_1^u(0) < \infty$ and $\mathbf{v}_1^u(t) \geq 0$ for all $t \geq 0$, it is not hard to show that Eq. (42) implies that $\lim_{t\to\infty} \left|\mathbf{s}_1^u(t) - \mathbf{s}_1^I\right| = 0$.

We then proceed by induction. Suppose $\lim_{t\to\infty} \left|\mathbf{s}_i^u(t) - \mathbf{s}_i^I\right| = 0$ for some $i \geq 1$. By Eq. (41), we have

$$\begin{aligned}
\dot{\mathbf{v}}_{i+1}^u &= (1-p)\left(\frac{\lambda \mathbf{s}_i^u - g_i(\mathbf{v}^u)}{1-p} - \mathbf{s}_{i+1}^u\right) \\
&= (1-p)\left(\frac{\lambda \mathbf{s}_i^I - g_i(\mathbf{v}^u)}{1-p} - \mathbf{s}_{i+1}^u + \epsilon_i^u\right) \\
&\leq (1-p)\left(\mathbf{s}_{i+1}^I - \mathbf{s}_{i+1}^u + \epsilon_i^u\right), \qquad (43)
\end{aligned}$$

where $\epsilon_i^u \triangleq \frac{\lambda}{1-p}\left(\mathbf{s}_i^u(t) - \mathbf{s}_i^I\right) \to 0$ as $t \to \infty$ by the induction hypothesis. With the same argument as the one for $\mathbf{s}_1$, we obtain $\lim_{t\to\infty} |\mathbf{s}_{i+1}^u(t) - \mathbf{s}_{i+1}^I| = 0$. This establishes the convergence of $\mathbf{s}^u(t)$ to $\mathbf{s}^I$ along all coordinates, which implies $\lim_{t\to\infty} \left\|\mathbf{s}^u(t) - \mathbf{s}^I\right\|_w = 0$. 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. □

8. CONCLUSIONS

The overall theme 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 moderate-sized finite systems ($N = 100$).

For future work, some current modeling 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 arrival and processing times are assumed to be Poisson, while in reality more general traffic distributions (e.g., heavy-tailed traffic) are observed. Finally, the speed of the central server may not be able to scale linearly in N for large N. Further work to extend the current model by incorporating these realistic constraints could be of great interest, although obtaining theoretical characterizations seems quite challenging. Lastly, 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.

9. 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.

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

[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): pp. 20–34, 1996.

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

[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*, pp. 255–312, 2001.

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

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

[11] J.N. Tsitsiklis and K. Xu, On the power of (even a little) centralization in distributed processing (Technical Report). `http://web.mit.edu/jnt/www/Papers/TXSIG11.pdf`.